

THE UNIVERSITY of ADELAIDE

Multi-modal Generation, Synergy and Evaluation

Qi Chen

A thesis submitted for the degree of
DOCTOR OF PHILOSOPHY
The University of Adelaide

May 1, 2024

Contents

Abstract xiii

Declaration of Authorship xv

Acknowledgements xvii

Publications xix

1 Introduction 1
- 1.1 Methodologies for Image Captioning . . . 1
  - 1.1.1 General Domain . . . 1
  - 1.1.2 Specific (Medical) Domain . . . 2
- 1.2 Multi-modal Synergy in Visual Voice Cloning . . . 3
- 1.3 Evaluation for Text-to-Image Synthesis . . . 4

2 Literature Review 7
- 2.1 Generation in Multi-modalities . . . 7
  - 2.1.1 Development of Methodologies . . . 7
  - 2.1.2 Development of Synergy and Applications . . . 8
  - 2.1.3 Development of Evaluation Metrics . . . 10
- 2.2 Major Tasks . . . 10
  - 2.2.1 Methodology: Image Captioning . . . 10
  - 2.2.2 Synergy: Visual Voice Cloning . . . 12
  - 2.2.3 Evaluation: Text-to-Image Evaluation . . . 13

3 Methodologies for Image Captioning 15
- 3.1 Introduction . . . 15
  - 3.1.1 Image Captioning in General Domain . . . 15
  - 3.1.2 Image Captioning in Specific (Medical) Domain . . . 17
- 3.2 Method and Experiments in General Domain . . . 19
  - 3.2.1 Method . . . 19
    - 3.2.1.1 Discrete Mode Learning . . . 20
    - 3.2.1.2 Mode Encoding and Assignment . . . 21
    - 3.2.1.3 Fully Non-autoregressive Decoder . . . 21
    - 3.2.1.4 Learning MIC Models with DML . . . 22
  - 3.2.2 Experiments . . . 22
    - 3.2.2.1 Dataset and Evaluation Metrics . . . 22
    - 3.2.2.2 Implementation Details . . . 23
    - 3.2.2.3 Evaluation on Diversity . . . 23
    - 3.2.2.4 Evaluation on Quality . . . 24
    - 3.2.2.5 Performance Analysis of the CdVAE branch . . . 26
    - 3.2.2.6 Ablation Studies on the CdVAE branch . . . 27
- 3.3 Method and Experiments in Specific (Medical) Domain . . . 28

3.3.1 Method . . . 29
3.3.1.1 Preliminaries . . . 29
3.3.1.2 Radiology-informed Knowledge Aggregation . . . 30
3.3.1.3 Implicit Prior Guidance . . . 31
3.3.2 Experiments . . . 32
3.3.2.1 Datasets . . . 32
3.3.2.2 Metrics and Implementation Details . . . 34
3.3.2.3 Comparison with State-of-the-arts . . . 35
3.3.2.4 Qualitative Evaluation . . . 37
3.3.2.5 More Discussions . . . 38
3.4 Conclusion . . . 42

4 Multi-modal Synergy in Visual Voice Cloning **47**
4.1 Introduction . . . 47
4.2 Task and Dataset . . . 49
4.2.1 Problem Definition for V2C Task . . . 49
4.2.2 Dataset Construction for V2C Task . . . 49
4.2.3 V2C-Animation Dataset vs. Related Datasets . . . 51
4.2.4 More Analysis of V2C-Animation Dataset . . . 51
4.3 Method . . . 54
4.3.1 Multi-modal Encoder for Feature Extraction . . . 56
4.3.2 Synthesizer for Mel-Spectrogram Generation . . . 57
4.3.3 Vocoder for Speech Synthesis . . . 59
4.4 Experiments . . . 59
4.4.1 Evaluation Metrics . . . 60
4.4.2 Implementation Details . . . 61
4.4.3 Quantitative Evaluation . . . 62
4.4.4 Qualitative Evaluation . . . 62
4.4.5 More Discussions . . . 63
4.5 Conclusion . . . 64

5 Likelihood-Based Text-to-Image Evaluation **65**
5.1 Introduction . . . 65
5.2 Method . . . 67
5.2.1 Overview . . . 67
5.2.2 Evaluation via Likelihood Estimation . . . 68
5.2.3 Credit Assignment of Perceptual Significance . . . 69
5.2.4 Credit Assignment of Semantic Significance . . . 69
5.3 Experiments . . . 72
5.3.1 Meta Evaluation . . . 72
5.3.2 Implementation Details . . . 72
5.3.3 Results on Manually Synthesized Data . . . 72
5.3.4 Results on Model Generated Data . . . 75
5.3.5 More Discussions . . . 78
5.4 Conclusion . . . 79

6 Conclusion **83**
6.1 Limitations and Future Work . . . 83
6.2 Future Work in More Various Domains . . . 84

Bibliography **86**

List of Figures

1.1 We decompose the framework of multi-modal content generation into three parts: generation, synergy, and evaluation. In the generation part, we delve into the study of a typical multi-modal generation task (*i.e.*, image captioning) in both general and specific (healthcare/medical) domains. For synergy, instead of focusing on tasks like image captioning that consider partial modalities, we explore interactions among various modalities in a newly proposed task–visual voice cloning (V2C). This task covers all three main modalities: text, vision, and voice. Finally, we review evaluation metrics for text-to-image synthesis and propose an improved metric due to the urgency for human-like assessment. 2

3.1 Captions written by humans vs. captions generated by existing models (Transformer [Vaswani et al., 2017] and AoANet [Huang et al., 2019]), and by our DML. The model-generated captions prefer to use common words or phrases while captions derived from humans are more informative and diverse. The proposed DML generates diverse image captions based on different mode embeddings, and some of the modes tend to yield certain language patterns. *E.g.*, mode-7 is likely to generate captions with the pattern of *"There is ..."*, mode-3 tends to produce complex sentences while mode-58 is prone to brief sentences. 16

3.2 Reports written by radiologists vs. reports generated by the existing model (*e.g.*, R2Gen [Chen et al., 2020g] trained on our merged dataset), and our S4M. We observe that R2Gen remembers some commonly used descriptions across different body regions (highlighted in red) regardless of the semantic alignment with the given radiology images, *e.g.*, the correct diagnosis is *"there is moderate cardiomegaly"* (highlighted in green) while R2Gen keeps *"the heart is normal in size"* (highlighted in red). 18

3.3 Overall of the DML paradigm. The "$\leftrightarrow$" denotes the matching operation by using Hungarian algorithm while the "$\backslash\backslash$" means there is no gradient flow. 19

3.4 Samples of captions that are generated based on different modes. Mode-7, mode-32 and mode-43 tend to generate the captions with a certain pattern, *i.e.*, *"There is ..."*, *"Man ..."*, and *"A close up of ..."*. Mode-3 is apt to use long sentences while mode-58 is prone to brief sentences. . 24

3.5 Comparison between our DML and SoTA baseline COS-CVAE. We show the oracle results of (A) CIDEr, (B) SPICE and (C) METEOR with different numbers of samples. 25

3.6 Effects of the Hungarian mode assignment and the fully non-autoregressive decoding objective. Here, (A) and (B) are DML without NAT objective and Hungarian assign, respectively. (C) is the final DML we design. We visualize the mode embeddings (red "×") and the caption embeddings ("·" with other colors) on the test set by t-SNE. The inactivated modes are not shown for clarity. 27
3.7 Overall of S4M. The framework of our base model is highlighted by the arrow in red. The blue and green arrows refer to Radiology-informed Knowledge Aggregation (RadKA) and Implicit Prior Guidance (IPG) branches, respectively. Notably, IPG would be removed after training, while RadKA would be retained. The "⊗" denotes the indication function. 30
3.8 Examples on our private datasets. Each example contains a frontal image (first column) and another image (second column) with the corresponding radiology report. 33
3.9 Reports generated by S4M (ours) and two baselines, *i.e.*, R2Gen and R2Gen†. Note that R2Gen is trained on IU-Xray only while R2Gen† optimised on our merged training dataset. We highlight the specific areas in the radiology image by coloured bounding boxes (red, blue and orange) as well as the corresponding descriptions in reports with the same colour. 37

4.1 (a) Voice Cloning (VC) vs. (b) Visual Voice Cloning (V2C). Given an input triplet (*i.e.*, subtitle/text, reference audio, and target video), our V2C task seeks to convert the text into a speech, which should be with the voice of reference audio and the emotion derived from reference video. Note that the reference audio only provides an expected voice while its content is irrelevant. 48
4.2 An example of how to cut a movie into a series of video clips according to SRT subtitle files. Note that the SRT files contain both starting and ending time stamps for each video clip. No. 1340 refers to the sequential number of the current utterance. 49
4.3 Examples from the existing Text-to-Speech (TTS) datasets (*i.e.*, LJ speech, LibriTTS and LibriSpeech) and our V2C-Animation dataset. A pitch of 0 Hz refers to an unvoiced segment. 52
4.4 Word cloud of the texts on our V2C-Animation dataset. 53
4.5 Top 30 words on V2C-Animation along with the counts. 53
4.6 Distribution of emotion labels on V2C-Animation. 54
4.7 Distribution of utterance/text length. 56
4.8 Examples of how to cut a movie into a series of video clips according to subtitle files. Note that the subtitle files contain both starting and ending time stamps for each video clip. 56
4.9 Samples of the character's emotion (*e.g.*, happy and sad) involved in the reference video. Here, we take *Elsa* (a character in the movie *Frozen*) as an example. 57
4.10 Overview of V2C-Net. It consists of three main components: a multi-modal encoder, a synthesizer and a vocoder. A triplet (*i.e.*, text, reference audio and reference video) is fed into the multi-modal encoder (Sec. 4.3.1) and it outputs three types of embeddings. Based on these embeddings, the synthesizer (Sec. 4.3.2) generates a mel-spectrogram. Finally, the mel-spectrogram is converted into the waveform (*i.e.*, speech) by the vocoder (Sec. 4.3.3). 57

4.11 Mel-spectrogram of generated and ground-truth audios. Orange curves are F_0 contours, where F_0 is the fundamental frequency of audio. Purple curves refer to the energy (volume) of audio. The horizontal axis is the duration of the audio. We highlight the main difference via a red circle. 62
4.12 More visualized mel-spectrograms of generated and ground-truth audios. The orange curves are F_0 contours, where F_0 denotes the fundamental frequency of audio. The purple curves refer to the energy (volume) of audio. The horizontal axis is the duration of the audio. We highlight the main difference via a red circle. 64

5.1 Generated images for the caption *"A bathroom with a blue shower curtain and walls"* with OPGAN and stable diffusion (the first column) and their visualized likelihood maps (the second column). Although stable diffusion produces a much better result, its average likelihood is 1.401×10^{-4}, less than OPGAN (8.020×10^{-4}). After removing the rare codes (the last column), their likelihoods increase to 3.297×10^{-3} and 1.670×10^{-3}, respectively, where stable diffusion is considered to be better. 68
5.2 (a) are the images generated by the stable diffusion model. (b) and (c) are the likelihood maps before and after the semantic credit assignment, respectively. For these two images, their overall likelihoods are 2.422×10^{-4} and 4.123×10^{-4}, respectively, where the second image which misses the foreground object *"male surfer"* is rated higher than the first image. This can be corrected after the semantic credit assignment, where the overall likelihoods become 1.049×10^{-4} and 4.704×10^{-5} in (c). 70
5.3 The computation graph of the last layer in the original CLIP and our modified CLIP. 71
5.4 (A)-(C): The trend of the proposed LEICA score (Score Value) and the corresponding values of Kendall' Tau (K-Tau) when adding different noises with different disturbance degrees to the images. (D): The trend of the LEICA score and Accuracy (Acc) of its judgments *w.r.t.* semantic alignment when gradually replacing the keywords in texts. 73
5.5 The ranks of each generative model derived by human raters and the proposed evaluation metric. Rank-1 means the best model while rank-11 is the worst one. 75
5.6 Impacts of the number of evaluated samples. We test three commonly used evaluation metrics (*i.e.*, IS, FID and SOA-C) as well as the proposed LEICA for the generated images derived from four different generative models. 76
5.7 The semantic significance maps obtained from the original CLIP and our modified CLIP on generated images. 78
5.8 More visual results of the impact of each component in our LEICA. The generated images are from stable diffusion. 80
5.9 Comparison between the rank of each generative model derived by human raters and that from various automatic evaluation metrics, including IS, FID, KID, SOA-C, SOA-I and CLIP-FID on the MS-COCO dataset. Rank-1 means the best model while rank-11 is the worst one. 81

List of Tables

3.1 Diversity evaluation on the best-5 sentences obtained from consensus re-ranking. 24
3.2 Comparison with baselines *w.r.t.* oracle performance (*i.e.*, best-1 quality) on COCO dataset. "#Sample" refers to the number of generated captions for each image. The best and the second best results are highlighted with **bold** and underline, respectively. 25
3.3 The performance of image captions generated by different modes. For Transformer-DML, we select some representative modes that exhibit clear language patterns or semantic complexities as shown in Figure 3.4. For AoANet-DML, we show the best-performed modes in terms of CIDEr, SPICE and ClipScore, respectively (highlighted with underline). The original results of Transformer and AoANet are obtained using the code and settings in [Luo]. The best results of the two captioning methods on each metric are highlighted with **bold**. 26
3.4 Oracle results of different batch sizes for CdVAE branch and MIC branch. "#effective modes" indicates the total number of modes that have ever been used during the whole training process. "#sampled caps" refers to how many captions per image are sampled to optimize the MIC branch. 28
3.5 Oracle results of different masking strategies for CdVAE branch. "#effective modes" is the total number of modes that have ever been used during the whole training process. "1.0 $\rightarrow$ 0.0" means we gradually reduce the masking probability from 1 to 0 throughout the training process. 28
3.6 Comparison with the recent specialised models on six datasets covering different body parts. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance. We highlight the best results on specialised models with underline while the best results on all models (both specialised and generalist) with **bold**. 36
3.7 Performance analysis of each component. "Base" denotes the model containing only an image encoder and a text decoder. Both RadKA and RadKA* are applied on top of our Base model, where the RadKA* means RadKA without indicator $\mathbb{1}(\cdot)$ in Eq. (3.9). The IPG is applied only on top of the Base+RadKA. 38
3.8 CLIPScore on IU-Xray to assess semantic alignment between images and generated reports. The notation † means we optimise the model on our merged training dataset. 39
3.9 Effect of each part for report-image alignment. 39

3.10 Results of linear probing on CheXpert to evaluate the recognition ability of the image encoder. The notation † means we optimise the model on our merged training dataset. 40
3.11 Impact of hyper-parameter λ in Eq. (3.14) 40
3.12 Effect of different feature extractors. 41
3.13 Effect of different feature extractors. 41
3.14 Results of B4 using partial training data. 42
3.15 **Average** results on the six datasets compared with the recent specialized models. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance. 43
3.16 Comparison with the recent specialized models on IU-Xray. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance. 44
3.17 Comparison with the recent specialized models on Abdomen. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance. 44
3.18 Comparison with the recent specialized models on Knee. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance. 44
3.19 Comparison with the recent specialized models on Hip. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance. 45
3.20 Comparison with the recent specialized models on Wrist. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance. 45
3.21 Comparison with the recent specialized models on Shoulder. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance. 45

4.1 We compare our V2C-Animation dataset with several existing multimodal datasets. "Identity" and "Emotion" indicate whether the datasets contain annotations corresponding to the speaker's identity and emotion. The notations "#Movies", "#Video Clips", "#Audio Clips" and "#Speakers" refer to the number of movies, videos, audios and speakers/characters, respectively. "Avg. S" indicates the average length of subtitle while "Avg. A/V" is the average duration of audio/video. . . 50
4.2 We compare the average and variance of the pitch from our V2C-Animation dataset and the related datasets. "Avg. P" refers to the average values of pitches with the corresponding variances. 52
4.3 Counts of the emotion labels on V2V-Animation dataset. 54
4.4 Movies with the corresponding speakers/characters on the V2C-Animation dataset. 55

4.5 Comparison with the state-of-the-art methods. We provide the results of both objective (*i.e.*, MCD, MCD-DTW and MCD-DTW-SL) and subjective evaluation metrics (*i.e.*, MOS-naturalness and MOS-similarity). "Id. Acc." and "Emo. Acc." are the identity and emotion accuracy of the generated speech, respectively. The method with "*" refers to a variant taking video (emotion) embedding as an additional input. "Ground Truth" denotes the results on ground-truth samples. ↑ (↓) means that the higher (lower) value is better. 59
4.6 Effect of reference audio and video. "ref. A" and "ref. V" denote the reference audio and reference video separately. "Id. Acc." and "Emo. Acc." are the identity and emotion accuracy of the generated speech, respectively. ↑ means a higher value is better. 63
4.7 Comparisons on the difficulties of the conventional Voice Cloning (VC) and Visual Voice Cloning (V2C) tasks. We show the performance of SV2TTS [Jia et al., 2018] trained on different datasets. VCTK, LibriSpeech and Multi-speaker are widely used in VC task. ↑ (↓) means higher (lower) value refers to better performance. 63

5.1 Ablation studies on the components of LEICA. The first row of results indicates the performance of LEICA. The last row is the performance of the averaged log-likelihood. "+" denotes combining the "funny mirror" distortion with the current distortion. dVAE: distortion using the dVAE model of DALL-E. Best results are highlighted with **bold**. . . . 74
5.2 Pearson, Spearman and Kendall's Tau correlations between different evaluation metrics and human judgment. The highest correlation is **bold**. 74
5.3 Pearson, Spearman and Kendall's Tau correlations between different evaluation metrics and human judgment on CUB & Oxford-Flower. The highest correlation is **bold**. 77

The University of Adelaide

Abstract

Multi-modal Generation, Synergy and Evaluation

by Qi Chen

Multi-modal content generation is at the forefront of current research in artificial intelligence, offering novel solutions to complex problems across various modalities and domains such as natural language processing, computer vision, and healthcare. In this thesis, we present an in-depth exploration of these areas, decomposing the AIGC framework into three main components: methods, synergies, and evaluations. We aim to tackle the inherent challenges in multi-modal generation and assess the efficacy of these solutions in real-world scenarios.

First, for the methodologies, we consider a traditional task–image captioning–in both general (natural image) and specific domains (medical image). In the general domain, we propose a Discrete Mode Learning (DML) paradigm, a new learning method in natural image captioning designed to overcome the prevalent mode collapse problem in generative models. DML enhances the diversity and richness of language patterns in generated captions, significantly improving the linguistic variety. In the medical domain, we present a Single-for-Multiple (S4M) framework for automated radiology report generation. S4M efficiently synthesizes accurate and relevant medical reports from diverse imaging data, demonstrating the versatility and precision of multi-modal models in a healthcare context.

Second, we delve into the realm of cross-modal synergy in real-world scenarios. To this end, we propose a Visual Voice Cloning (V2C) task that transcends traditional voice cloning boundaries by incorporating emotional nuances. It incorporates all the modalities, including vision, language, and voice. In this task, we aim to synthesize voice from textual inputs while aligning it with emotions extracted from video data, which offers potential applications in areas ranging from entertainment (*e.g.*, movie dubbing) to assistive technologies.

Finally, we propose a new evaluation metric, named Likelihood-based tExt-to-Image evaluation with Credit Assignment (LEICA), for the text-to-image synthesis task, which is a well-known fundamental generation task. Specifically, we seek to overcome the limitations of existing metrics by directly estimating the likelihood of generated images using a pre-trained model, while incorporating a credit assignment strategy for both perceptual and semantic significance. This metric provides a more accurate and efficient tool for evaluating the alignment and quality of synthesized images from textual descriptions. It not only refines the assessment of multi-modal models but also aligns closely with human perceptual judgments, thus representing a significant advancement in the field.

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Qi Chen

November 28, 2023

Acknowledgements

First and foremost, I would like to thank my PhD supervisor Prof. Qi Wu and co-supervisor Dr. Yuankai Qi, for all their help during my studies at the University of Adelaide. It is fortunate for me to work with such researchers who are always enthusiastic about their work and willing to provide help and guidance whenever needed. They have always been a constant source of support, guidance, and inspiration for me throughout my PhD journey.

I am also grateful to my co-authors, lab mates, and friends, without whom I could never have made it through those tough days. They were always there to lend a listening ear, offer a helping hand, and provide words of encouragement when I needed it the most.

Finally, I am forever grateful to my parents. They helped me to not only survive but also to thrive, and for that, I will always be thankful.

Publications

This thesis contains the following works that have been published or prepared for publication:

- V2C: Visual Voice Cloning

  Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, Qi Wu

  IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022

- Learning Distinct and Representative Modes for Image Captioning

  Qi Chen, Chaorui Deng, Qi Wu

  Advances in Neural Information Processing Systems (NeurIPS), 2022

- S4M: Generating Radiology Reports by A Single Model for Multiple Body Parts

  Qi Chen, Yutong Xie, Biao Wu, Minh-Son To, James Ang, Qi Wu

  ArXiv Preprint, 2023.

- LEICA: Likelihood-Based Text-to-Image Evaluation with Patch-Level Perceptual and Semantic Credit Assignment

  Qi Chen, Chaorui Deng, Zixiong Huang, Bowen Zhang, Mingkui Tan, Qi Wu

  ArXiv Preprint, 2023.

In addition, I have the following papers not included in this thesis:

- Prompt Switch: Efficient CLIP Adaptation for Text-Video Retrieval

  Chaorui Deng, **Qi Chen**, Pengda Qin, Da Chen, Qi Wu

  IEEE International Conference on Computer Vision (ICCV), 2023

- Pareto-aware Neural Architecture Generation for Diverse Computational Budgets

  Yong Guo, Yaofo Chen, Yin Zheng, **Qi Chen**, Peilin Zhao, Junzhou Huang, Jian Chen, Mingkui Tan

  IEEE Conference on Computer Vision and Pattern Recognition (CVPR) workshop, 2023

- Towards Lightweight Super-resolution with Dual Regression Learning

  Yong Guo, Jingdong Wang, **Qi Chen**, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, Jian Chen, Mingkui Tan

  ArXiv Preprint, 2022.

- R-GAN: Exploring Human-like Way for Reasonable Text-to-Image Synthesis via Generative Adversarial Networks

  Yanyuan Qiao, **Qi Chen**, Chaorui Deng, Ning Ding, Yuankai Qi, Mingkui Tan, Xincheng Ren, Qi Wu

  ACM International Conference on Multimedia (ACM MM), 2021

- Towards Accurate and Compact Architectures via Neural Architecture Transformer

  Yong Guo, Yin Zheng, Mingkui Tan, **Qi Chen**, Zhipeng Li, Jian Chen, Peilin Zhao, Junzhou Huang

  IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021

- Contrastive Neural Architecture Search with Neural Architecture Comparators

  Yaofo Chen, Yong Guo, **Qi Chen**, Minli Li, Yaowei Wang, Wei Zeng, Mingkui Tan

  IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021

Chapter 1

Introduction

Multi-modal content generation has attracted much attention from researchers and presents innovative solutions to intricate problems across various domains including natural language processing, computer vision, and healthcare. In this thesis, we delve into an extensive exploration of it by decomposing its framework into three pivotal parts: generation (methodologies), synergy, and evaluation (as shown in Figure 1.1). We aim to address the inherent challenges in multi-modal generation and to assess the efficacy of these methods in practical, real-world scenarios.

1.1 Methodologies for Image Captioning

There are many existing tasks and methods for multi-modal generation. Here, we focus on a typical multi-modal generation task–image captioning. For a thorough study, we consider the image captioning task in both the general domain, using natural images, and the specific domain like healthcare for medical image captioning (a.k.a. medical report generation). While both domains revolve around the same task of image captioning, they present distinct challenges owing to inherent differences among domain-specific images. Thus, in this section, we discuss them separately and briefly for clarity while the details will be introduced in Chapter 3.

1.1.1 General Domain

We begin with image captioning in the general domain. Natural image captioning seeks to produce natural descriptions for given natural images, playing a crucial role in real-world applications such as human-computer interaction and multi-modal recommendation. This has led to a surge in research efforts, with recent state-of-the-art methods [Huang et al., 2019; Li et al., 2020a; Zhou et al., 2020] demonstrating promising results based on evaluation metrics like CIDEr [Vedantam et al., 2015], BLEU [Papineni et al., 2002], and SPICE [Anderson et al., 2016]. However, as highlighted in [Wang et al., 2020], a predominant focus on optimizing these metrics tends to steer image captioning models towards employing commonplace words, phrases, and language patterns from the training corpus. In essence, the model automatically defaults to the most general mode for captioning, resulting in generated captions that lack both semantic and syntactic diversity. It is widely known as the "mode collapse" problem and has been discussed in many prior works on generative modeling [Goodfellow et al., 2014; Zhou et al., 2019]. Due to such a problem, this disparity from human capabilities becomes evident, as humans possess the ability to describe images in a myriad of ways.

In this thesis, we address the aforementioned problem by introducing a novel approach called Discrete Mode Learning (DML) to cultivate distinct and representative modes for image captioning. In contrast to diverse image captioning methods based

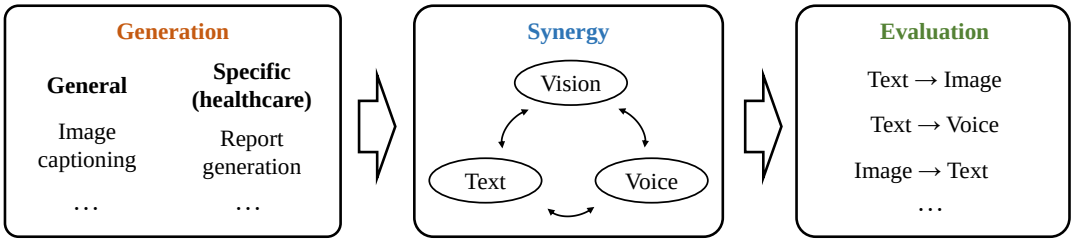


FIGURE 1.1. We decompose the framework of multi-modal content generation into three parts: generation, synergy, and evaluation. In the generation part, we delve into the study of a typical multi-modal generation task (*i.e.*, image captioning) in both general and specific (healthcare/medical) domains. For synergy, instead of focusing on tasks like image captioning that consider partial modalities, we explore interactions among various modalities in a newly proposed task–visual voice cloning (V2C). This task covers all three main modalities: text, vision, and voice. Finally, we review evaluation metrics for text-to-image synthesis and propose an improved metric due to the urgency for human-like assessment.

on Conditional Variational Autoencoders (CVAE), DML adopts a unique strategy of learning a codebook. This codebook serves as an embedding matrix, encompassing a set of mode embeddings that define a discrete latent space. Consequently, the language patterns encoded in different modes become easily discernible and are more interpretable. Different from prior controllable image captioning methods [Aneja et al., 2019; Chen et al., 2018; Cornia et al., 2019; Deshpande et al., 2019; Mahajan and Roth, 2020; Shuster et al., 2019; Wang et al., 2017a], DML stands out for its lack of dependency on additional supervision for mode learning. This characteristic enhances convenience and removes limitations associated with predefined sets of control signals, making DML a versatile and user-friendly approach.

1.1.2 Specific (Medical) Domain

As for the medical domain, interpreting radiology images and generating accurate diagnostic reports pose significant challenges, with inherent complexities and potential for errors. To alleviate this burden, automatic report generation systems offer a valuable solution, providing radiologists with candidate reports in natural language for validation. Additionally, these systems offer a distinct advantage by leveraging machine learning paradigms that can learn directly from free-text reports, a notable departure from other data-intensive medical image analysis applications, such as medical image segmentation [Isensee et al., 2021; Lee et al., 2023; Zhang et al., 2021b], which often rely on extensive quality annotations. Radiologists routinely produce reports covering various body parts, including the chest, abdomen, shoulder, elbow, wrist, hip, knee, ankle, and more. However, current report generation models [Li et al., 2023b; Liu et al., 2021a,b; Ma et al., 2021; Wang et al., 2023; Yan et al., 2023; Zhang et al., 2020] are typically developed following a conventional single-dataset training-and-testing paradigm. This approach leads to notable performance degradation when these models are directly applied to a different dataset involving various body regions.

In this thesis, we design a single model for radiology report generation across multiple body parts, named single-for-multiple (S4M). The S4M framework integrates

explicit and implicit priors to enhance the model's understanding and reasoning capabilities on limited but diverse data from different body parts. Concretely, it introduces two additional branches, namely, a Radiology-informed Knowledge Aggregation (RadKA) branch and an Implicit Prior Guidance (IPG) branch, leveraging pre-defined radiology-informed knowledge and trainable implicit priors from ground-truth reports, respectively. The RadKA branch incorporates domain-specific knowledge to guide the model's behavior and filters out region-irrelevant knowledge, prioritizing region-specific knowledge during report generation. In contrast, the IPG branch enhances the image encoder's recognition capability by learning from ground-truth reports, improving the model's understanding of semantic meaning across modalities.

1.2 Multi-modal Synergy in Visual Voice Cloning

Besides the tasks/applications (*e.g.*, image captioning) that focus on part of the multi-modalities, we seek to further explore the interactions and synergy among all the modalities (*i.e.*, vision, text, and voice). To this end, in this thesis, we define a new task called Visual Voice Cloning (V2C) that covers all these modalities beyond the scope of traditional Voice Cloning (VC) tasks [Arik et al., 2018; Chen et al., 2019b; Jia et al., 2018; Nachmani et al., 2018], where text is converted to speech only based on the voice from reference audio. Our V2C task involves processing a triplet input (text/subtitle, reference audio, reference video). The goal is to generate speech with the same voice as the reference audio but with varying emotions derived from the reference video. The text/subtitle represents the content the generated speech should convey, the reference audio features a pre-recorded voice of the target speaker from a different clip, and our objective is to produce speech with the voice from the reference audio and the character's visual emotion from the reference video, articulating the given text. It can be a demonstration task for exploring the synergy among various modalities.

As there is no readily available dataset for the V2C task, we have curated the inaugural V2C-Animation dataset to advance research in this domain. This dataset comprises 10,217 video clips with corresponding audios and subtitles, featuring content from 26 animated movies and involving 153 characters (speakers) in total. Unlike existing text-to-speech or movie description datasets that typically focus on text and audio, or text and video, our V2C dataset encompasses three modalities—text, audio, and video. Additionally, each audio and video clip in our dataset is annotated with emotions (*e.g.*, happy or sad), similar to previous work [Goodfellow et al., 2013]. To mitigate the impact of background music, we extract the sound channel of the central speaker, ensuring that all audio clips solely contain the sound from speaking characters.

The new task introduces new challenges. First, conventional Voice Cloning (VC) methods [Arik et al., 2018; Blaauw et al., 2019; Chen et al., 2019b; Jia et al., 2018; Nachmani et al., 2018] are inadequate for the V2C task, as they primarily focus on converting input text to speech with the voice/tone of the reference audio, neglecting the emotion and context essential for human-like speech in tasks like movie dubbing. Second, in V2C, the voice emotion must be derived from the reference video, not an unrelated reference audio clip. For instance, in movie dubbing, understanding character emotions requires observing their performances, such as facial expressions or actions, in the corresponding movie clips. While some advanced VC methods [Skerry-Ryan et al., 2018; Wang et al., 2018b] attempt to incorporate voice emotion into generated speech, they capture both emotion and voice from the reference audio,

falling short of V2C requirements. Ideally, an effective V2C method should disentangle voice and emotion from reference audio and video, respectively.

To tackle the challenges in the V2C task, we introduce a novel method named Visual Voice Cloning Network (V2C-Net), built upon the widely used Text-to-Speech (TTS) framework (FastSpeech2 [Ren et al., 2021]). V2C-Net specifically incorporates emotional information from reference video frames. Additionally, we propose an evaluation metric, MCD-DTW weighted by Speech Length (MCD-DTW-SL), based on Mel Cepstral Distortion (MCD [Kubichek, 1993]), to effectively and automatically assess the quality of the generated speech. We will depict more in Chapter 4.

1.3 Evaluation for Text-to-Image Synthesis

Evaluating the performance of multi-modal generative models remains an open question. In this thesis, we diverge from well-established evaluation metrics in the above image captioning and visual voice cloning, which have reached an acceptable level of alignment with human preferences. Instead, our emphasis is on the evaluation metrics in text-to-image synthesis, a field in its early stages and in need of urgent solutions. Although recent advancements [Ramesh et al., 2022; Saharia et al., 2022; Tan et al., 2020; Yang et al., 2021; Yu et al., 2022] have shown promise in the text-to-image generation task, evaluating the performance of text-to-image generation remains a challenging problem, particularly in complex, open-vocabulary scenarios. Most existing models, including state-of-the-art ones, still rely on popular evaluation metrics from general image generation, such as Inception Score (IS) [Salimans et al., 2016] and Fr'echet Inception Distance (FID) [Heusel et al., 2017].

The Inception Score (IS) utilizes a pre-trained image recognition model (Inception-V3 [Szegedy et al., 2016]) to assess the generative performance by providing the recognition confidence of generated images. A higher confidence score may suggest greater similarity to the images used for training Inception-V3 (e.g., ImageNet [Deng et al., 2009]). However, the critical challenge lies in the fact that similarity with pre-training images may not necessarily reflect the perceptual quality or semantic alignment of generated images, given the significant domain shift between the pre-training data (which has limited semantic diversity) and open-vocabulary natural languages. Additionally, pre-trained models can be prone to over/underfitting, making confidence scores potentially unreliable. Similarly, metrics following IS, such as Classification Accuracy Score (CAS) [Ravuri and Vinyals, 2019] and Semantic Object Accuracy (SOA) [Hinz et al., 2020], may also encounter these issues.

In contrast, the Fr'echet Inception Distance (FID) calculates the distribution distance between real and generated images in a latent space, where Inception-V3 serves as the backbone to map images into the latent space. The method imposes a multivariate Gaussian assumption on the latent space to estimate distribution statistics. While FID partially addresses the domain shift issue of IS by incorporating real images during evaluation, it encounters several challenges. Firstly, the assumption of a multivariate Gaussian distribution lacks a solid foundation, making it an unreliable indicator of the perceptual quality of generated images. Secondly, FID is designed to assess the overall similarity of two sets of images, lacking the capability to measure semantic alignment between input text and the corresponding generated image. Lastly, both IS and FID are inefficient, requiring the processing of numerous image samples using the Inception V3 model to stabilize their evaluation results. Many of these limitations also apply to Kernel Inception Distance (KID) [Bińkowski et al., 2018], a kernelized variant of FID.

In this thesis, we introduce a novel metric for text-to-image evaluation based on likelihood estimation. Specifically, given a set of n image descriptions $X = \{x_i\}_{i=1}^n$ and their corresponding generated images $Y = \{y_i\}_{i=1}^n$, we employ a pre-trained likelihood-based text-to-image model $\mathcal{G}$ to estimate the likelihood of each y_i conditioned on x_i, denoted as $P_{\mathcal{G}}(y_i|x_i)$. Ideally, when the model-estimated density closely aligns with the real-world data density, $P_{\mathcal{G}}(y_i|x_i)$ effectively indicates the likelihood of the image y_i existing in the real world given the text x_i [Theis et al., 2016]. To achieve a high $P_{\mathcal{G}}(y_i|x_i)$, y_i should not only be semantically aligned with x_i but also perceptually comparable to real-world images. Otherwise, a low likelihood score for y_i is obtained. Average log-likelihood has been the standard measure, widely considered for detecting out-of-distribution data [Ren et al., 2019a] or quantifying generative modeling performance [Theis et al., 2016; Van Den Oord et al., 2016]. We provide the details in Chapter 5.

Chapter 2

Literature Review

In this chapter, we first review the development of generation in multi-modalities, *w.r.t.* the methodologies, the synergy (applications) among different modalities, and the evaluation methods. Then, we delve into one of the major tasks of each aspect of the multi-modal content generation, *i.e.*, image captioning (methodology), visual voice cloning (synergy), and text-to-image evaluation (evaluation). We depict the details in the following.

2.1 Generation in Multi-modalities

The exploration of multi-modal generation can be categorized into methodologies, applications, and evaluation metrics. Methodologies focus on developing techniques for generating content across modalities like text, image, and audio [Goodfellow et al., 2014; Ho et al., 2020]. Applications/synergies involve practical use cases such as image captioning [Anderson et al., 2018a; Chen et al., 2022a; Huang et al., 2019; Vinyals et al., 2015; Xu et al., 2015] and voice cloning [Gibiansky et al., 2017; Ping et al., 2018; Taigman et al., 2018]. Evaluation metrics [Bińkowski et al., 2018; Heusel et al., 2017; Hinz et al., 2020; Ravuri and Vinyals, 2019; Salimans et al., 2016] provide quantitative measures to assess the quality of generated content. This structured approach aids in understanding advancements and challenges, guiding further research in artificial intelligence-generated content.

2.1.1 Development of Methodologies

Backbone

Initially, multi-modal content generation methodologies focus on simple yet effective integration strategies. Key among these is the use of Convolutional Neural Networks (CNNs) [He et al., 2016; Krizhevsky et al., 2012] for image recognition tasks and Recurrent Neural Networks (RNNs) [Rumelhart et al., 1985] for sequence processing in text. The combination of these two technologies allows for early forms of image captioning and basic text-to-image synthesis. This phase lays the groundwork for understanding and processing multiple modalities within a unified framework.

The advent of deep learning brings a significant leap in multi-modal methodologies. The introduction of more sophisticated neural network architectures, such as Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] networks, enables better handling of sequences in text and audio, leading to improvements in tasks like speech recognition and machine translation. These advancements are crucial in enhancing the fluidity and coherence of generated content across different modalities.

The development of transformer models [Vaswani et al., 2017] marks a turning point in multi-modal content generation. These models, characterized by their attention mechanisms, greatly improve the ability to handle long-range dependencies in

text, leading to more accurate and contextually relevant content generation. Transformers soon find applications beyond text, aiding in tasks such as image captioning [Anderson et al., 2018a; Huang et al., 2019; Vinyals et al., 2015; Xu et al., 2015] and cross-modal retrieval [Chen et al., 2020d; Deng et al., 2023; Dong et al., 2021; Kiros et al., 2014; Yu et al., 2017, 2018b], exemplified by models like BERT [Devlin et al., 2019], UNITER [Chen et al., 2020f], Oscar [Li et al., 2020a], LXMERT [Tan and Bansal, 2019], ViLBERT [Lu et al., 2019], and CLIP [Radford et al., 2021].

Generation Paradigm

Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] introduce a new paradigm in content generation, particularly for images and videos. Initially proposed for natural image generation from input noise, GAN applications have expanded significantly, encompassing diverse areas such as image generation [Arjovsky et al., 2017; Brock et al., 2018; Cao et al., 2018; Gulrajani et al., 2017; Guo et al., 2019; Karras et al., 2018; Radford et al., 2016]. Furthermore, GAN has found utility in various intriguing applications, including image inpainting [Dolhansky and Ferrer, 2018; Yang et al., 2017; Yu et al., 2018a; Zhao et al., 2018], image/video super-resolution [Guo et al., 2018, 2020b; Wang et al., 2018a], image/video deblurring [Kupyn et al., 2018; Zhang et al., 2019], and facial attribute manipulation [Huang et al., 2017; Shen and Liu, 2017; Shu et al., 2017], among others. Specifically, GAN operates by training a generative model through a two-player minimax game, aiming to match the underlying data distribution. Comprising a generator and a discriminator, the generator strives to produce samples whose distribution is indistinguishable from the training samples, while the discriminator acts as a judge to differentiate between generated and real samples. GANs like StyleGAN [Karras et al., 2019] and BigGAN [Brock et al., 2018] demonstrate unprecedented capabilities in generating photo-realistic images, opening up possibilities for creating high-quality synthetic visual content from textual descriptions or other images.

Among the latest advancements in multi-modal content generation methodologies are Diffusion Models (DM) [Ho et al., 2020], which have revolutionized the field. Unlike GANs, which learn to generate content by directly modeling the distribution of training data, diffusion models work by gradually transforming a distribution of random noise into a distribution of the data. This process, akin to reverse denoising, has shown remarkable proficiency in generating high-quality, diverse images. The primary impact of diffusion models has been in the area of image synthesis [Rombach et al., 2022; Zhang et al., 2023]. Their ability to generate intricate and diverse images has positioned them as a formidable alternative to GANs, especially in tasks where detail and variety are paramount. The text-to-image synthesis capabilities of these models [Gu et al., 2022; Nichol et al., 2022; Rombach et al., 2022; Saharia et al., 2022] have shown remarkable proficiency in creating photo-realistic and contextually accurate images from textual descriptions. This breakthrough has implications not just for image generation but also for bridging the gap between textual and visual content, enhancing the scope of multi-modal applications.

2.1.2 Development of Synergy and Applications

The initial stages of synergy in multi-modal content generation focus on basic integrations of different modalities. This includes tasks like image captioning [Anderson et al., 2018a; Chen et al., 2022a; Huang et al., 2019; Vinyals et al., 2015; Xu et al., 2015], where text and images are combined to generate descriptive captions. Early

applications in this domain are primarily experimental, aimed at understanding how different modalities could complement each other.

A significant leap in synergy is observed in audio-visual applications. Technologies like lip-syncing algorithms [Park et al., 2022; Prajwal et al., 2020] and audio-visual speech recognition systems [Afouras et al., 2018; Xu et al., 2020] have begun to emerge. These applications not only synthesize speech from a text but also match it with appropriate lip movements in videos, enhancing the realism of generated content. This development marks a key step in creating more lifelike and interactive AI systems.

The evolution continues with the development of interactive systems that could respond to user inputs in real-time. Interactive storytelling [Missaouib et al., 2021], where narratives could adapt based on user choices, and AI-driven recommendation systems that use user data across multiple modalities (text, visual preferences, and browsing history) show how synergy could lead to more personalized and engaging experiences.

The synergy of AI with creative processes leads to the emergence of AI in art and music generation [Copet et al., 2023; Dong et al., 2018; Goel et al., 2022; Vasquez and Lewis, 2019]. Here, AI algorithms start to not only create artistic content based on text descriptions but also to collaborate with human artists, offering new tools for creative expression. Similarly, in music, AI begins to generate compositions that blend styles and genres, demonstrating the potential of AI in creative industries.

Another one is multi-modal learning, with early efforts [Girdhar et al., 2023; Li et al., 2020a; Lu et al., 2019] exploring ways to extend language models' capabilities to understand visual content. While these models initially showed promise on specific vision-language tasks [Antol et al., 2015], they struggled to generalize to broader scenarios. However, recent advancements have demonstrated that simpler model architectures can achieve impressive results with extensive large-scale pretraining, facilitated by enhanced computational resources and diverse datasets. For instance, PaLI [Chen et al., 2023b] integrates a vision transformer with a language transformer and trains on a vast dataset of 10 billion image-text pairs, showcasing remarkable flexibility in generating and modifying both text and images. Similarly, CM3Leon [Yu et al., 2023] adopts a simple decoder-only transformer architecture, trained on 340 million image-text pairs, yielding strong performance in image-to-text and text-to-image conversions. Additionally, to enable a Language and Vision Model (LLM) capable of generating non-text content, Emu [Sun et al., 2023] combines a stable diffusion model with LLaMA as a decoder, trained on a diverse corpus of 82 million image-text and video-text pairs. This integration represents a significant advancement in the field, highlighting the growing versatility of LLM in multi-modal contexts.

One of the most impactful developments has been in healthcare and assistive technologies [Chen et al., 2023a, 2020g, 2022b; Li et al., 2022b, 2023b; Liu et al., 2021a,b; Wang et al., 2022b, 2023; Zhang et al., 2020]. AI systems that can interpret medical images and generate reports, or convert spoken language to sign language, have shown the potential of multi-modal AI in improving healthcare delivery and accessibility. The integration of AI in augmented reality (AR) and virtual reality (VR) applications represents the latest frontier in multi-modal content generation. These technologies create immersive experiences by combining visual, auditory, and sometimes haptic feedback, driven by AI algorithms that understand and respond to user interactions in a multi-modal environment. Finally, the synergy in multi-modal content generation has found its way into everyday technologies, such as smartphones [Burns et al., 2022; Li et al., 2020b; Sun et al., 2022] and home assistants [Anderson et al., 2018b; Qi et al., 2020]. These devices use AI to understand and process inputs in various

forms–voice commands, visual cues, and touch–and provide responses or actions that integrate multiple modalities.

2.1.3 Development of Evaluation Metrics

Initially, evaluation metrics are primarily designed for single modalities. For text, metrics like BLEU [Papineni et al., 2002] and ROUGE [Lin, 2004] are developed to assess machine translation and summarization tasks. In image generation, metrics such as the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are used to evaluate the quality of synthesized images. These metrics provide a foundation but ae limited in assessing the nuances of multi-modal content. As multi-modal generation tasks become more prevalent, there is a need for metrics that could evaluate the effectiveness of content generation across different modalities. Metrics like CIDEr [Vedantam et al., 2015] and SPICE [Anderson et al., 2016] have emerged for tasks like image captioning, which require simultaneous evaluation of textual and visual quality. These metrics begin to address the alignment and relevance between modalities. With the advancement in specific multi-modal tasks like audio-visual speech synthesis or text-to-image generation, task-specific metrics are developed. For instance, the Lip Reading Sentences (LRS) metric [Son Chung et al., 2017] for evaluating lip-sync accuracy in generated videos, or Fréchet Inception Distance (FID) [Heusel et al., 2017] for assessing the quality of images generated from textual descriptions. These metrics provide more targeted and relevant evaluations for specific multi-modal interactions.

Recognizing the limitations of purely quantitative metrics, the field sees a shift towards human-centric and qualitative assessments. Studies [Chen et al., 2020b; Ramesh et al., 2021; Saharia et al., 2022] begin incorporating user studies and human evaluations to gauge the perceptual quality and effectiveness of multi-modal content. These assessments provide insights into user experience and satisfaction, aspects often missed by automated metrics. The latest developments in evaluation metrics involve assessing semantic coherence and context relevance in multi-modal content. Metrics like CLIPScore [Hessel et al., 2021] use advanced models like CLIP [Radford et al., 2021] to evaluate the alignment between text and images, moving beyond surface-level comparisons to deeper semantic understanding.

2.2 Major Tasks

2.2.1 Methodology: Image Captioning

Natural Image Captioning

Natural image captioning [Anderson et al., 2018a; Huang et al., 2019; Vinyals et al., 2015; Xu et al., 2015] aims to generate descriptions based on given images, garnering significant attention from researchers [LeCun et al., 2015]. The conventional paradigm of image captioning models [Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015] typically comprises two components: a CNN-based image encoder and an RNN-based decoder for caption generation. Numerous works [Anderson et al., 2018a; Huang et al., 2019; Zhou et al., 2020] enhance this paradigm with the introduction of attention mechanisms [Deng et al., 2020b; Li et al., 2018b; Schwartz et al., 2019; Vaswani et al., 2017; Xu et al., 2015], compelling models to focus more on highlighted regions. Additionally, to boost performance, [Yang et al., 2019; Yao et al., 2018] investigate visual relationships through the construction of semantic or scene graphs,

while [Guo et al., 2020a; Rennie et al., 2017] optimize models using Reinforcement Learning (RL) and directly employ CIDEr [Vedantam et al., 2015] for reward computation.

Diverse and Controllable Image Captioning

Diverse image captioning endeavors to train a model capable of generating various captions for the same image. CVAE-based models [Chen et al., 2019a; Mahajan and Roth, 2020; Wang et al., 2017a] achieve this by learning a latent space during training and subsequently generating diverse captions through the sampling of various priors from the latent space. GAN-based models [Dai et al., 2017; Li et al., 2018a; Shetty et al., 2017] achieve caption diversity by utilizing different random noises as inputs in conjunction with given images. While diverse image captioning models can produce varied captions during inference, interpreting the underlying conditional distribution remains a non-trivial challenge, introducing uncertainty about the model's behavior. To impart control over generated captions, some approaches [Chen et al., 2021a, 2020c; Cornia et al., 2019; Deng et al., 2020a; Mathews et al., 2016, 2018] introduce an additional control signal. For instance, Mathews *et al* [Mathews et al., 2016] incorporate a sentiment signal for each caption, aiming to regulate the sentiment of the generated captions. Deng *et al* [Deng et al., 2020a] utilize the length of the caption as a control signal. By conditioning on embeddings representing different length levels, the model can generate descriptions with controllable lengths for the input image. However, many of these methods necessitate reliance on additional tools or annotations to supervise the learning of the control signal, introducing inconvenience in data collection and limiting language patterns within a pre-defined domain.

Medical Image Captioning (Radiology Report Generation)

Radiology report generation is a specialized task targeting medical imaging data, with the goal of producing comprehensive and accurate reports that encapsulate findings, interpretations, and diagnoses related to medical images. Previous works [Jing et al., 2018; Wang et al., 2021; Xue et al., 2018; Yuan et al., 2019] have employed hierarchical LSTMs for generating lengthy paragraphs in medical reports. To further enhance performance, several studies [Chen et al., 2020g, 2022b; Li et al., 2022b; Wang et al., 2022b, 2023] have opted for a transformer as the report decoder instead of LSTM, resulting in notable improvements. Additionally, recent efforts [Li et al., 2023b; Liu et al., 2021a,b; Zhang et al., 2020] have explored the integration of knowledge graphs to capture radiology terminologies and their semantic relationships. These knowledge graphs serve as inputs or optimization constraints, such as classification labels, during the generation process. However, current report generation models adhere to a single-dataset training-and-testing paradigm, and when applied to datasets encompassing different body parts (chest, abdomen, shoulders, etc.), they often suffer from significant performance degradation. Moreover, existing knowledge-based models like [Liu et al., 2021a; Yang et al., 2022] develop a static knowledge graph where the relations (edges) between topics (nodes) cannot be updated during training, limiting their effectiveness in exploiting implicit relationships. Additionally, the knowledge graph in these models focuses solely on chest X-rays, constraining its applicability to other body parts and broader use cases. Therefore, we propose a reorganization of topics in our knowledge set to cover medical terminologies relevant to a diverse range of body parts without predefined and fixed relations, allowing for learnable relationships.

2.2.2 Synergy: Visual Voice Cloning

Text to Speech

Numerous text-to-speech (TTS) synthesis methods [Arık et al., 2017; Chen et al., 2021b; Kalchbrenner et al., 2018; Li et al., 2019; Wang et al., 2017b; Yan et al., 2021] have been proposed for generating natural speech from text. Building on WaveNet, Deep Voice [Arık et al., 2017] divides a TTS model into multiple modules optimized independently. Wang *et al* [Wang et al., 2017b] introduce the Tacotron framework, consolidating all stages essential for text-to-speech synthesis and enabling end-to-end optimization of the speech synthesis model. TransformerTTS [Li et al., 2019] integrates the transformer structure [Vaswani et al., 2017] into the TTS task, while Ren *et al* [Ren et al., 2019b] propose a more efficient transformer (FastSpeech) using a non-autoregressive generation method. Building on FastSpeech, they further develop an enhanced version, FastSpeech2 [Ren et al., 2021], aiming to control generated speech by adjusting pitch and energy.

Voice Cloning

Unlike the text-to-speech (TTS) method, which synthesizes speech with a single voice, the voice cloning (VC) task [Gibiansky et al., 2017; Ping et al., 2018; Taigman et al., 2018] aims to generate multiple speeches with different voices. Building on Deep Voice [Arık et al., 2017] and Tacotron [Wang et al., 2017b], Deep Voice 2 [Gibiansky et al., 2017] maps voices from different speakers into a common space, using the low-dimensional embedding from this common space as a condition to assist the generation process. Jia *et al* [Jia et al., 2018] propose a multi-speaker TTS framework, consisting of three sub-modules (encoder, synthesizer, and vocoder), capable of synthesizing high-quality speech from given text. Recent extensions [Arik et al., 2018; Blaauw et al., 2019; Chen et al., 2019b; Nachmani et al., 2018] focus on synthesizing the voice of an unseen person using only a few samples. Arik *et al* [Arik et al., 2018] explore two approaches for synthesizing a person's voice from limited audio samples: speaker adaptation fine-tunes a trained multi-speaker model for an unseen speaker using a few audio-text pairs, while speaker encoding directly estimates the speaker embedding from audios of an unseen speaker. Chen *et al* [Chen et al., 2019b] propose an adaptive TTS system using a meta-learning approach.

Prosody Transfer

To achieve realistic speech, prosody transfer (PT) [Bian et al., 2019; Hsu et al., 2019; Skerry-Ryan et al., 2018; Stanton et al., 2018; Valle et al., 2020; Wang et al., 2018b; Whitehill et al., 2020] aims to capture prosody from reference audios. Building on Tacotron [Wang et al., 2017b], Skerry-Ryan *et al* [Skerry-Ryan et al., 2018] propose an encoder architecture to learn a representation of prosody from reference spectrogram slices derived from the reference audio. Global Style Tokens (GSTs) [Wang et al., 2018b] models the styles of different speakers using an interpretable embedding, which serves as a condition when transferring various speaking styles. Utilizing the variational autoencoder (VAE) framework [Kingma and Welling, 2014], Hsu *et al* [Hsu et al., 2019] design a neural sequence-to-sequence TTS model that categorizes speaking styles into latent attributes, enabling control over the speaking style by adapting these attributes. To transfer underrepresented speaking styles in the dataset, Whitehill *et al* [Whitehill et al., 2020] propose an adversarial cycle-consistent training procedure for a multi-reference neural TTS system.

2.2.3 Evaluation: Text-to-Image Evaluation

Text-to-Image Synthesis

Generative models can be broadly classified into three types: Generative Adversarial Networks (GAN) [Bi et al., 2022; Chen et al., 2020a; Goodfellow et al., 2014], likelihood-based models, and score-based models. Early works in this domain primarily adopt GAN-based approaches, such as StackGAN [Zhang et al., 2017], AttnGAN [Xu et al., 2018], DM-GAN [Zhu et al., 2019], and OP-GAN [Hinz et al., 2020]. However, GANs focus on modeling the sampling process of a distribution, making it challenging to directly obtain likelihoods of generated samples for text-to-image evaluation. Score-based models, exemplified by diffusion models [Gu et al., 2022; Nichol et al., 2022; Rombach et al., 2022; Saharia et al., 2022], have gained popularity for generating photo-realistic images and artworks. While diffusion models can determine the likelihood of generated images, their computation requires multiple steps and cannot be easily parallelized for efficient evaluation. Likelihood-based models, including DALL-E [Ramesh et al., 2021], OFA [Wang et al., 2022a], CogView [Ding et al., 2021, 2022], and Parti [Yu et al., 2022], predominantly adopt autoregressive structures. These models perform competitively with diffusion models, offering the key advantage of efficiently obtaining the likelihood of a generated image in a single forward step, making them well-suited for evaluating text-to-image generation.

Evaluation Metrics for Text-to-Image Synthesis

Commonly employed metrics for assessing generative image modeling, including IS [Salimans et al., 2016], CAS [Ravuri and Vinyals, 2019], FID [Heusel et al., 2017], and KID [Bińkowski et al., 2018], face challenges in text-to-image evaluation. SOA [Hinz et al., 2020], tailored for text-to-image generation on the MSCOCO [Lin et al., 2014] dataset, filters captions in the validation set for keywords related to MSCOCO object categories. For each filtered caption, images are generated and examined by an object detector (YOLOv3 [Redmon and Farhadi, 2018] trained on MSCOCO) to detect keywords-related objects. SOA-C and SOA-I scores report the average recall of these keywords over all categories or images, respectively, evaluating semantic alignment. While effective, SOA is inconvenient (due to caption pre-processing) and lacks generalizability (bound to MSCOCO). Additionally, it shares challenges with IS. Some retrieval metrics, like R-precision, focus solely on assessing the semantic alignment of generated images [Xu et al., 2018], disregarding perceptual quality. In contrast, average log-likelihood mitigates these issues and is a widely accepted default measure for evaluating generative models [Theis et al., 2016; Van Den Oord et al., 2016].

In this thesis, several contributions have been proposed compared to existing methods:

- **Discrete Mode Learning (DML) framework for Image Captioning**: Introduces a new paradigm to overcome mode collapse in generative models, enhancing linguistic diversity in captions generated based on given images.
- **Single-for-Multiple (S4M) framework for Radiology Report Generation**: Presents a versatile framework capable of generating accurate reports across various body parts, addressing limitations of existing specialized and generalist models.
- **Visual Voice Cloning (V2C) Task**: Extends traditional voice cloning by incorporating emotional nuances from videos, integrating vision, language, and voice modalities.

- **Likelihood-based tExt-to-Image evaluation with Credit Assignment (LEICA)**: Proposes a novel evaluation metric for text-to-image synthesis, overcoming limitations of existing metrics by incorporating perceptual and semantic significance.

Statement of Authorship

Title of Paper	Learning Distinct and Representative Modes for Image Captioning
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Chen, Q., Deng, C. and Wu, Q., 2022. Learning distinct and representative modes for image captioning. Advances in Neural Information Processing Systems, 35, pp.9472-9485.

Principal Author

Name of Principal Author (Candidate)	Qi Chen		
Contribution to the Paper	Proposed the ideas, conducted experiments, and wrote the manuscript.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	03/12/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Chaorui Deng		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Qi Wu		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Please cut and paste additional co-author panels here as required.

Statement of Authorship

Title of Paper	S4M: Generating Radiology Reports by A Single Model for Multiple Body Parts
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to Medical Image Analysis

Principal Author

Name of Principal Author (Candidate)	Qi Chen		
Contribution to the Paper	Proposed the ideas, conducted experiments, and wrote the manuscript.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	03/12/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Yutong Xie		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Biao Wu		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Minh-Son To

Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	James Ang		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Qi Wu		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Please cut and paste additional co-author panels here as required.

Chapter 3

Methodologies for Image Captioning

3.1 Introduction

Image captioning aims to generate natural descriptions for a given image. It is widely used in many real-world applications (*e.g.*, human-computer interaction and medical report generation) and hence has attracted lots of research attention. Here, for a comprehensive study, we review and seek to improve the image captioning methods in both the general natural image domain and the specific medical domain. For clarity, we discuss them separately below in this chapter.

3.1.1 Image Captioning in General Domain

Recently, many state-of-the-art (SoTA) methods [Huang et al., 2019; Li et al., 2020a; Zhou et al., 2020] have achieved promising results *w.r.t.* evaluation metrics like CIDEr [Vedantam et al., 2015], BLEU [Papineni et al., 2002], and SPICE [Anderson et al., 2016]. However, as discussed in [Wang et al., 2020], focusing on achieving higher scores on these metrics usually biases the image captioning models towards using only the common words, phrases and language patterns in the training corpus when describing the images (see Figure 3.2 for an example). In other words, the model automatically finds the most general mode to perform captioning. As a result, the generated captions are limited in diversity both semantically and syntactically. This is far away from the ability of human beings as humans are able to describe the image in various ways.

The cause of this phenomenon is widely known as the *mode collapse* problem and has been discussed in many prior works on generative modeling [Goodfellow et al., 2014; Zhou et al., 2019]. Formally, given an input x and a generative model $\mathcal{G}$, mode collapse appears when the estimated output distribution $P_{\mathcal{G}}(x)$ assigns most of its probability mass to a small region of the output space, despite that the real data distribution $P_{\text{data}}(y)$ has a much larger variance. As a consequence, the randomly sampled outputs $\{y_i \sim P_{\mathcal{G}}(x) \mid i \in \mathbb{N}\}$ tend to be very similar. While mode collapse is typically a side effect for generative modeling, it is somewhat "welcomed" in SoTA image captioning models as it usually facilitates a higher evaluation performance on reference-based metrics like CIDEr, BLEU and SPICE. For example, CIDEr optimization [Rennie et al., 2017] based methods [Cornia et al., 2020; Huang et al., 2019] have significantly pushed the performance of image captioning to a new level on mainstream reference-based evaluation metrics. However, as shown in [Luo and Shakhnarovich, 2020; Wang et al., 2020], the success of CIDEr optimization could be largely attributed to its ability of reducing the modes in the generated captions.



Captions generated by humans:
A man holding a white frisbee while standing on a field.
A man standing outside holding a frisbee in his hands.
A man is outside holding onto a frisbee disc.
A smiling man in the park with a frisbee.
A man is having a good time playing frisbee.

Captions generated by existing models:
Transformer: *A man holding a frisbee while standing in a field.*
AoANet: *A man holding a frisbee in a field.*

Captions generated by DML (Ours):
Mode-7: *There is a man playing with a frisbee.*
Mode-32: *Man in a blue shirt throwing a white frisbee.*
Mode-43: *A close up of a person with a frisbee.*
Mode-3: *A young man holding a white frisbee in his right hand.*
Mode-58: *A man is about to throw a frisbee.*

FIGURE 3.1. Captions written by humans vs. captions generated by existing models (Transformer [Vaswani et al., 2017] and AoANet [Huang et al., 2019]), and by our DML. The model-generated captions prefer to use common words or phrases while captions derived from humans are more informative and diverse. The proposed DML generates diverse image captions based on different mode embeddings, and some of the modes tend to yield certain language patterns. *E.g.*, mode-7 is likely to generate captions with the pattern of *"There is ..."*, mode-3 tends to produce complex sentences while mode-58 is prone to brief sentences.

Some recent researches in image captioning have attempted to tackle the mode collapse problem so as to improve the diversity of the generated captions. Specifically, [Aneja et al., 2019; Mahajan and Roth, 2020; Wang et al., 2017a] adopt conditional variational autoencoders (CVAE) to encode the conditional distribution of the image captions into a low-dimensional continuous latent space $\mathcal{Z}$. When performing inference, these methods randomly sample latent variables from $\mathcal{Z}$ and input them into the caption decoder to drive it towards using different language patterns to describe the image. In this sense, the sampled latent variables can be considered as the continuous representations of modes interpolated from all occurred modes during training, thus alleviating the mode collapse phenomena to some extent. Nevertheless, it is still difficult to interpret the underlying conditional distribution, *i.e.*, which part of $\mathcal{Z}$ corresponds to which kind of language patterns in the real world. Thus, for the captions derived from two sampled mode representations, it is uncertain how they differ from each other during the sampling process, as well as their qualities. To tackle this uncertainty, another line of works seeks to learn controllable image captioning models. For example, [Chen et al., 2018; Shuster et al., 2019] control the style of the image captions with learned sentiment or personality representations like "factual" or "humorous", "positive" or "negative", *etc.*. [Deshpande et al., 2019] controls the syntactic structure of the image captions through part-of-speech tags. [Cornia et al., 2019] focuses on different image regions to generate region-specific image captions. However, these methods rely on additional tools or annotations to supervise the learning of modes. More critically, this also restricts their modes within a pre-defined domain.

In this part, we tackle the above problems by learning distinct and representative modes for image captioning through a new Discrete Mode Learning (DML) paradigm. Different from CVAE-based diverse image captioning methods, DML learns a codebook, which is an embedding matrix consisting of a set of mode embeddings that spans a discrete latent space, thus the language patterns encoded in different modes can be easily evaluated and are more perceptible. Different from previous controllable image captioning methods, DML requires no additional supervision for mode and hence is

more convenient to use and is not limited to the pre-defined set of control signals.

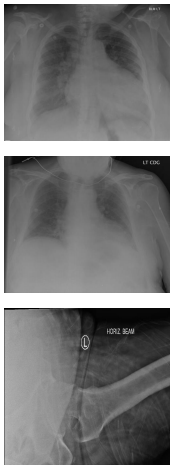
Specifically, we optimize a dual architecture in the proposed DML: **Firstly**, an image-conditioned discrete variational autoencoder (CdVAE) branch, where an encoder extracts the hidden states for all the reference captions paired with an image, and further quantizes them using their matched mode embeddings in the codebook according to Euclidean distance. For the matching algorithm, we choose Hungarian algorithm instead of the naive nearest neighbor look-up, which we find to be critical for increasing the number of effective modes. Afterward, a decoder is used to reconstruct the reference captions in a *fully non-autoregressive* manner, which breaks the sequential dependencies on previous tokens and enforces the decoder to rely purely on the mode embeddings to generate different captions based on the same image feature, leading to more distinct and representative mode embeddings. **Secondly**, a mode-conditioned image captioning (MIC) branch, which can be simply modified from an existing image captioning model by adding the mode embedding to the original word embeddings as a control signal. During inference, the CdVAE branch is dropped, and the MIC branch is used to generate image captions with various language patterns according to the mode embeddings in the codebook.

In the experiments, we evaluate the effectiveness of our proposed DML paradigm by applying it to the widely-used Transformer [Vaswani et al., 2017] and the state-of-the-art AoANet [Huang et al., 2019], denoted by Transformer-DML and AoANet-DML, respectively. We observe that some of the modes tend to yield clear language patterns in the generated captions, as shown in Figure 3.2, demonstrating that DML has successfully learned perceptible modes without explicit supervision. We also find that our models perform surprisingly well under diversity evaluation (using metrics like SelfCIDEr [Wang et al., 2020]) and oracle performance evaluation (on mainstream reference-based metrics like CIDEr [Vedantam et al., 2015]), achieving new state-of-the-art results. This shows that our learned modes are not only distinct but also effectively cover the rich modes that appeared within the dataset, which suggests that they are very representative. Moreover, we find that the models trained with DML outperform their original counterparts in terms of quality on some of the modes, meaning that DML can serve as a cost-free plugin for existing image captioning models.[1]

3.1.2 Image Captioning in Specific (Medical) Domain

The tasks of interpreting radiology images and producing diagnostic reports are both arduous and prone to errors. To reduce this burden, automatic report generation systems can provide candidate reports in natural language for the radiologist to verify. Besides, these systems can leverage data-hungry machine learning paradigms by learning directly from free-text reports, which is a significant advantage compared to other medical image analysis applications (*e.g.*, medical image segmentation [Isensee et al., 2021; Lee et al., 2023; Zhang et al., 2021b]) that often rely on large amounts of quality annotations. In general, radiologists commonly write reports based on radiology images covering different body parts, including the chest, abdomen, shoulder, elbow, wrist, hip, knee, ankle, and more. However, the existing report generation models [Li et al., 2023b; Liu et al., 2021a,b; Ma et al., 2021; Wang et al., 2023; Yan et al., 2023;

[1] It's important to note that at the time of publishing this image captioning work, the GPT4 vision model (GPT4-v) had not yet been introduced. Additionally, it's worth mentioning that GPT4-v tends to generate longer captions, whereas the captions in the COCO dataset are generally concise. As a result, we have deferred the comparison with GPT4-related works, particularly in narrative generation, to future research endeavours.



(Chest)

R2Gen: *The heart is normal in size. The mediastinum is stable. The aorta is tortuous.*

Ours: *Stable cardiomediastinal silhouette. Mild cardiomegaly. pulmonary vasculature is normal. No pneumothorax or pleural effusion. No acute bony abnormalities.*

GT: *There is moderate cardiomegaly. There are bilateral interstitial opacities increased since the previous exam. No focal airspace consolidation pleural effusions or pneumothorax. No acute bony abnormalities.*

(Chest)

R2Gen: *The heart is normal in size. The mediastinum is unremarkable. The lungs are clear but hypoinflated.*

Ours: *There are low lung volumes with bronchovascular crowding. There is no focal areas of consolidation. No pneumothorax . Degenerative changes of the thoracic spine.*

GT: *Low lung volumes with bibasilar subsegmental atelectasis. No focal consolidations pleural effusions or pneumothoraces. Cardiomediastinal silhouette is within normal limits. Degenerative changes of the thoracic spine.*

(Hip)

R2Gen: *The heart is normal in size. The lungs are clear. Pelvis and left hip the pelvic ring is intact. There is mild degenerative change of both hip joints. No pelvic fracture identified.*

Ours: *Pelvis and left hip there is a subcapital neck of femur fracture with shortening and angulation of the femoral head. No other fracture or dislocation seen.*

GT: *There is a transverse fracture involving the left neck of femur. there is mild impaction. There is external rotation. Both hips are enlocated. There is no significant degenerative disease of either hip. No further fracture is visualised.*

FIGURE 3.2. Reports written by radiologists vs. reports generated by the existing model (*e.g.*, R2Gen [Chen et al., 2020g] trained on our merged dataset), and our S4M. We observe that R2Gen remembers some commonly used descriptions across different body regions (highlighted in red) regardless of the semantic alignment with the given radiology images, *e.g.*, the correct diagnosis is *"there is moderate cardiomegaly"* (highlighted in green) while R2Gen keeps *"the heart is normal in size"* (highlighted in red).

Zhang et al., 2020] are designed by following the typical single-dataset training-and-testing paradigm, which inevitably suffers from severe performance drop issues, when these generation models are directly deployed to another dataset *w.r.t.* various body regions.

In this part, we seek to take a promising start in designing a single model for radiology report generation across multiple body parts. Unlike the existing works focusing on developing a framework verified within a single dataset, we aim to propose a versatile way to enable the report generation model to have the ability to learn from many diversified datasets regarding different body parts. To this end, we start by directly merging multiple datasets and training a single report generation model on the merged dataset, while observing that the performance of reports produced in such a simple way is only comparable with that trained separately on each specific dataset. We suspect the challenge lies in that increasing the diversity of the merged training data makes it harder to optimise the model, particularly if the model is complex and has a relatively large number of parameters, *e.g.*, a Transformer-based model [Vaswani et al., 2017]. This is because a more diverse dataset can make the optimisation problem more challenging, as the model needs to learn to generalise across a wider range of inputs and outputs. Thus, in general, even though increasing the diversity of the training data can improve the robustness and generalisation of the model and ultimately lead to better performance, optimising a model on a more diverse data space would require much more training data.

However, even with a merged dataset, the training data in radiology is still limited compared with that in natural images (*e.g.*, 400 million image-text pairs are used in CLIP [Radford et al., 2021] while 129 million images in BLIP [Li et al., 2022a] and BLIP-2 [Li et al., 2023a]). It makes models easily overfit and then results in poor generalisation to new or unseen data. From Figure 3.2, the existing model (*e.g.*, R2Gen) often generates the same sentence in the report even given different radiology images. Notably, not all the generated reports are semantically consistent with the images as the model tends to remember an "average" version that encompasses the

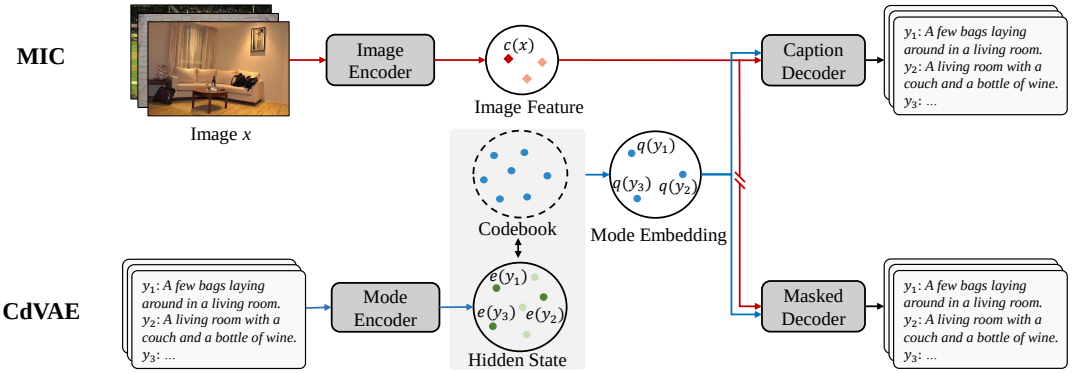


FIGURE 3.3. Overall of the DML paradigm. The "↔" denotes the matching operation by using Hungarian algorithm while the "\\" means there is no gradient flow.

frequently occurring words and phrases present in the training corpus [Chen et al., 2022a]. Such a phenomenon is not limited to the images within the same body part but is also observed in the data of different body parts (*i.e.*, chest vs. hip in Figure 3.2). Thus, this is particularly a concern with the models that have a high capacity to memorize the training data.

To address the above issue, we propose a single-for-multiple (S4M) framework for multi-part radiology report generation. In this framework, we use an explicit prior and an implicit prior to improve the understanding and reasoning capabilities of the model on limited but diverse data. Specifically, based on the widely used encoder-decoder backbone, our S4M introduces two additional branches, *i.e.*, a Radiology-informed Knowledge Aggregation (RadKA) branch and an Implicit Prior Guidance (IPG) branch. The RadKA branch incorporates pre-defined radiology-informed knowledge into the generation backbone, which provides a way to include domain-specific knowledge into the model to guide its behaviour. To further strengthen the region-aware knowledge during generation, the RadKA branch filters out the region-irrelevant knowledge using an indicator before aggregating the knowledge and image features. To do so, the model can prioritise the acquisition and utilisation of specific body region knowledge during the report generation process, and thus enhance the accuracy and relevance of the generated reports. Unlike RadKA, which uses explicit guidance, the IPG branch aims to enhance the recognition capability of the image encoder by employing a trainable implicit prior obtained from the ground-truth reports. This can significantly improve the image encoder's understanding of semantic meaning by enabling it to learn from another modality (*i.e.*, report), following the idea that one modality can provide diagnostic support but needs confirmation from another modality [Willemink et al., 2020]. We conduct the experiments on a merged dataset that consists of a public dataset (*i.e.*, IU-Xray) and five private datasets, covering the chest, abdomen, knee, hip, wrist and shoulder. The results show the superiority of our S4M model when generating radiology reports for multiple body parts compared with both specialised and generalist models.

3.2 Method and Experiments in General Domain

3.2.1 Method

The overall architecture of the proposed Discrete Mode Learning (DML) paradigm is illustrated in Figure 3.3. It consists of two branches: an image-conditioned discrete

variational autoencoder (CdVAE) branch, which contains a mode encoder $\mathcal{E}_m$ and a masked decoder $\mathcal{D}_m$; and a mode-conditioned image captioning (MIC) branch consists of an image encoder $\mathcal{E}_c$ and a caption decoder $\mathcal{D}_c$. The two branches are connected through a codebook Ω that contains a set of mode embeddings learned through the DML paradigm. During training, the model takes as inputs an image x and its paired reference captions $\{y_i\}_{i=1}^n$, and no additional supervision is required. During inference, the CdVAE branch is dropped, and the MIC branch is used to generate image captions with various language patterns according to the mode embeddings in the codebook. More details are as follows.

3.2.1.1 Discrete Mode Learning

In a typical image captioning model, the training objective is to maximize $\log p(y_i|x)$ for each y_i in $\{y_i\}_{i=1}^n$, which is ill-posed since the model is optimized to approach multiple different targets conditioned on the same input x, and usually leads to a mode collapse problem. Previous works alleviate this problem by introducing a latent variable z_i to the objective function, *i.e.*, $\mathbb{E}_{z_i\sim p(z)}[\log p(y_i|x, z_i)]$, which serves as an explicit indication of mode and drives the model to generate different targets conditioning on different z_i. *E.g.*, in CVAE-based diverse image captioning models [Mahajan and Roth, 2020; Wang et al., 2017a], z_i is randomly sampled from a continuous latent space, which leads to an uncertainty of the model behavior. While in some controllable image captioning models [Cornia et al., 2019; Deshpande et al., 2019], z_i needs to be pre-defined with the help of additional tools or annotations, which is usually restricted to a small set.

Unlike these methods, our DML samples latent variables from an embedding matrix $\Omega \in \mathbb{R}^{k\times d}$ which we call codebook. It defines a discrete latent space, where each entry in Ω corresponds to a potential mode embedding, and k is a hyper-parameter representing the total number of modes in the codebook. During training, the mode encoder $\mathcal{E}_m$ is used to extract the representation for each caption y_i, denoted by $e(y_i)$, and match the representation with one of the entries in Ω. The matched entry is then adopted as the mode embedding of y_i, denoted by $q(y_i)$, and the objective becomes:

$$\max_{\mathcal{E}_m,\mathcal{D}_m,\Omega} \log p(y_i|q(y_i), x), \quad \text{where } q(y_i) = \texttt{Match}(e(y_i), \Omega). \tag{3.1}$$

Here, $p(y_i|q(y_i), x)$ is the conditional distribution of the caption y_i given its mode embedding $q(y_i)$ and the paired image x, which is approximated with the masked decoder $\mathcal{D}_m$. $\texttt{Match}(\cdot,\cdot)$ indicates the matching operation. Note that, there is usually no real gradient defined for typical matching operators like the nearest neighbor look-up. Therefore, we follow [Van Den Oord et al., 2017] to first use the straight-through estimator to directly copy the gradients from $q(y_i)$ to $e(y_i)$, and then apply Vector Quantization algorithm to move $q(y_i)$ towards $e(y_i)$ through mean square loss so as to train the matched codebook entries. The new objective function is

$$-\log p(y_i|q(y_i), x) + \|\text{sg}[e(y_i)] - q(y_i)\|_2^2 + \beta\|e(y_i) - \text{sg}[q(y_i)]\|_2^2, \tag{3.2}$$

where the first term inherited from Eq. (3.1) is used to optimize $\mathcal{E}_m$ and $\mathcal{D}_m$. The second term is used to train the matched entries in Ω. The last term is a commitment loss that is responsible for bounding the embedding space of Ω. $\text{sg}[\cdot]$ refers to "stop gradient" and β is a hyper-parameter.

3.2.1.2 Mode Encoding and Assignment

We adopt a stack of N_e transformer encoder layers as our mode encoder $\mathcal{E}_m$. The input of $\mathcal{E}_m$ is y_i appended with a special `[MODE]` token to the start. The output of $\mathcal{E}_m$, $e(y_i) \in \mathbb{R}^d$, is the hidden state of the `[MODE]` token from the last layer. When performing the matching process in Eq. (3.1), a naive approach is to use the nearest neighbor look-up algorithm, *i.e.*,

$$q(y_i) = \Omega_\iota, \quad \text{where } \iota = \arg\min_j \|e(y_i) - \Omega_j\|_2. \tag{3.3}$$

However, we find in the experiments that in the models trained with this assignment strategy, the output of $\mathcal{E}_m$ quickly converges to the nearby of two or three mode embeddings in Ω, which is a clear sign of mode collapse (see Figure 3.6).

To avoid this, inspired by the object query assignment in [Carion et al., 2020], we treat the mode assignment of the captions as a Bipartite Graph Matching problem and solve it using Hungarian algorithm. Specifically, for all the reference captions $\{y_i\}_{i=1}^n$ paired with an image, we construct a bipartite graph between the output hidden states $\{e(y_i)\}_{i=1}^n$ from $\mathcal{E}_m$ and the mode embeddings in Ω. We first pad $\{e_i\}_{i=1}^n$ to the size of the codebook, k, with $\varnothing$ (assume $n < k$). Then, we search for a permutation of k elements $\tau \in \mathfrak{S}_k$ with the lowest assignment cost:

$$\hat{\tau} = \arg\min_{\tau \in \mathfrak{S}_k} \sum_i^k \|e(y_i) - \Omega_{\tau(i)}\|_2, \tag{3.4}$$

and the assigned mode embedding for each y_i is $q(y_i) = \Omega_{\hat{\tau}(i)}$. We find this assignment strategy greatly increases the number of effective mode embeddings in Ω (see Figure 3.6).

3.2.1.3 Fully Non-autoregressive Decoder

After getting the mode embeddings $q(y_i)$ for each caption y_i, we feed it together with the image features from x into a decoder to generate y_i and estimate the conditional distribution $p(y_i|q(y_i), x)$ in Eq. (3.2). In general, sequence generation models are often trained through the autoregressive Teacher Forcing scheme, which aims to maximize the likelihood of the ground-truth token w_t given all preceding ground-truth tokens $w_{j<t}$. In our setting, the objective function would be:

$$-\log p(y_i|q(y_i), x) = \sum_{t=1}^{T} -\log p(w_t|w_{j<t}, q(y_i), x) \tag{3.5}$$

where w_t is the t-th token in y_i and T is the total number of tokens in y_i. However, being able to access the previous tokens causes the decoder to ignore the mode embedding when reconstructing y_i, since the previous tokens already provide enough information for the mode of y_i. As a result, the training signals for the mode embeddings could be useless.

Therefore, we propose to use a fully non-autoregressive (NAT) objective for the CdVAE branch of DML to facilitate the training of the mode embeddings. Specifically, the masked decoder $\mathcal{D}_m$ is a stack of N_d transformer decoder layers that takes $q(y_i)$, x, and a sequence of T `[MASK]` tokens as input, and predicts each target token in y_i

in a conditionally-independent manner, *i.e.*,

$$-\log p(y_i|q(y_i), x) = \sum_{t=1}^{T} -\log p(w_t|\texttt{[MASK]}, q(y_i), x). \tag{3.6}$$

Apply this equation to Eq. (3.2) will result in our final objective function for the CdVAE branch. We find this leads to more distinct and representative mode embeddings (see Figure 3.6).

3.2.1.4 Learning MIC Models with DML

So far we have introduced how DML learns the mode embeddings with the CdVAE branch. Here we show how to attach the CdVAE branch to an existing image captioning model to make it aware of the mode of the generated captions. Specifically, in a standard encoder-decoder-based image captioning model, the encoder $\mathcal{E}_c$ encodes the information from the image x into a sequence of hidden states, then the decoder $\mathcal{D}_c$ attends to the encoder hidden states and predicts the ground-truth tokens in the reference caption y_i following the Teach Forcing scheme similar to Eq. (3.5).

To adapt it into our DML framework, we only need to make a minor modification to the token embedding layer of $\mathcal{D}_c$ by adding the mode embedding $q(y_i)$ to the word embeddings of each token in y_i (*i.e.*, $w_{j<t}$ in Eq. (3.5)) element-wisely, resulting in our MIC branch. The training of MIC and CdVAE is performed jointly, using the objective functions in Eq. (3.2) and Eq. (3.5), respectively. In practice, we find the CdVAE branch is much harder to train than the MIC branch, due to its non-autoregressive prediction manner. To make the convergence speed of the two branches compatible, we let them use different batch sizes, *i.e.*, for each image, the CdVAE branch takes as input all its paired captions $\{y_i\}_{i=1}^n$, while the MIC branch randomly samples just one caption from $\{y_i\}_{i=1}^n$. Moreover, the gradient from the CdVAE branch will not be back-propagated to the image encoder.

Inference

During inference, the CdVAE branch is dropped, and the inference of MIC follows a very similar procedure as the original image captioning model. The only difference is that it requires selecting a mode embedding from the codebook and adding it to the input token embeddings.

3.2.2 Experiments

3.2.2.1 Dataset and Evaluation Metrics

Dataset

We train and evaluate our method on MSCOCO dataset [Lin et al., 2014] that contains 123,287 images and each image is corresponding to at least 5 captions. For a fair comparison, we follow the previous works [Mahajan and Roth, 2020; Mahajan et al., 2020] in the area of diverse and controllable image captioning to use the m-RNN split [Mao et al., 2015] of the COCO dataset, which divides the data into 118,287, 4,000 and 1,000 for training, validation and testing, respectively.

Quality Evaluation

To assess the quality of the generated captions, we use five widely used evaluation metrics, *i.e.*, BLEU [Papineni et al., 2002], ROUGE [Lin, 2004], METEOR [Banerjee and Lavie, 2005], CIDEr [Vedantam et al., 2015], and SPICE [Anderson et al., 2016]. Moreover, to further evaluate the performance of the generated captions, we employ another CLIP-based [Radford et al., 2021] metric namely ClipScore [Hessel et al., 2021], which can assess whether the generated captions are semantically aligned with given images, even when they are totally different from the reference captions.

Diversity Evaluation

To investigate the diversity of the generated captions, we use SelfCIDEr [Wang and Chan, 2019], mBLEU, and n-gram diversity (*i.e.*, Div-n [Aneja et al., 2019]). All of these metrics evaluate the diversity by comparing the n-gram differences among the generated captions that belong to the same image.

3.2.2.2 Implementation Details

Choice of Base Models

The proposed DML is a general learning paradigm and we expect it can be easily applied in many existing image captioning models and improve their controllability and diversity. Thus, for the MIC branch, we choose two widely-used and representative architectures as our base models, *i.e.*, Transformer [Vaswani et al., 2017] and AoANet [Huang et al., 2019], denoted by Transformer-DML and AoANet-DML separately, to show the generalization ability of DML. Most current SoTA image captioning models, like M2Transformer [Cornia et al., 2020], XLAN [Pan et al., 2020] and many vision-language pre-training models, are based on the Transformer architecture. Moreover, the performance of Transformer and AoANet is also competitive with the SoTA models [Luo] when vision-language pre-training is not performed. Therefore, they are good baseline models to illustrate the performance of our DML paradigm.

Detailed Settings

In the CdVAE branch, the number of transformer layers for $\mathcal{E}_m$ and $\mathcal{D}_m$ is set to 6 and 2, respectively. We use 12 attention heads and a hidden size of 768 for all transformer layers. β in Eq. (3.2) is set to 0.25, following [Van Den Oord et al., 2017]. The number of mode embeddings in Ω is set to 64 by default. We convert each image to 100 object proposals by Faster RCNN [Ren et al., 2015] pre-trained on Visual Genome [Krishna et al., 2017]. We train the models for 100,000 iterations with a batch of 64 images and all the paired captions. We use AdamW [Loshchilov and Hutter, 2018] optimizer with a learning rate of 2e-4, and cosine decay it to 0. We use the label smoothing of 0.1 and the gradient clipping threshold of 1.0. We train Transformer-DML and AoANet-DML on one NVIDIA 3090 GPU with about 10 and 13 GPU hours, respectively.

3.2.2.3 Evaluation on Diversity

Quantitative Results

In this part, we perform a diversity analysis for the proposed DML paradigm based on Transformer-DML. We compare our model with previous SoTA methods as well as Beam Search (BS) and show the results in Table 3.1. From the table, our DML-based

TABLE 3.1. Diversity evaluation on the best-5 sentences obtained from consensus re-ranking.

Methods	LNFMM	COS-CVAE	Seq-CVAE	Transformer-BS	Transformer-DML (ours)
Div-1 (↑)	0.37	0.39	0.33	0.21	**0.43**
Div-2 (↑)	0.50	0.57	0.48	0.29	**0.59**
SelfCIDEr (↑)	-	0.79	-	0.57	**0.83**
mBLEU (↓)	0.64	**0.53**	0.64	0.78	0.54

Image			
Mode 7	*There is a man in the middle of a field playing with a frisbee*	*There is a man riding a motorcycle down the street*	*There is a man on a skateboard holding a can*
Mode 32	*Man in red shirt throwing a white frisbee*	*Man on a yellow motorcycle driving down the road*	*Man on skateboard with a beer on the ground*
Mode 43	*A close up of a person with a frisbee*	*A close up of a person riding a motorcycle on a road*	*A close up of a person on a skate board*
Mode 3	*A young man holding a white frisbee on top of a green field*	*A man riding a yellow motorcycle with a yellow helmet on*	*A young man holding a bottle of water while standing on a skateboard*
Mode 58	*A man getting ready to throw a frisbee*	*A man that is sitting on a motorcycle*	*A person on a court with a skateboard*

FIGURE 3.4. Samples of captions that are generated based on different modes. Mode-7, mode-32 and mode-43 tend to generate the captions with a certain pattern, *i.e.*, *"There is ..."*, *"Man ..."*, and *"A close up of ..."*. Mode-3 is apt to use long sentences while mode-58 is prone to brief sentences.

model achieves higher performance on most of the diversity evaluation metrics, which demonstrates its effectiveness in learning distinct modes.

Qualitative Results

We further provide several samples of the generated captions of our Transformer-DML model in Figure 3.4. We find that the learned modes exhibit clear and distinct language patterns. More specifically, mode-7, mode-32 and mode-43 tend to follow the patterns *"There is ..."*, *"Man ..."*, and *"A close up of ..."*, respectively. Mode-3 and mode-58 focus more on the semantic complexity of the caption, *i.e.*, mode-58 is likely to generate a brief caption while mode-3 is prone to the complicated one. These results show that for the same image, the mode embeddings learned through DML facilitate the model to generate captions in various and comprehensible ways.

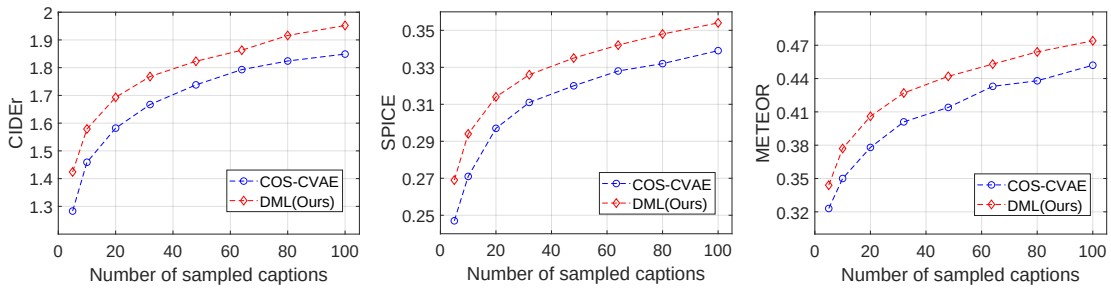
3.2.2.4 Evaluation on Quality

Oracle Results

To investigate whether the embeddings in the codebook are able to cover the rich modes that appeared within the dataset, we calculate the caption evaluation metrics in the oracle setting consistent with prior works [Aneja et al., 2019; Mahajan and Roth, 2020; Mahajan et al., 2020], *i.e.*, taking the maximum score for each quality metric over all the candidate captions for each image. Specifically, we train Transformer-DML and AoANet-DML with codebook sizes $k = 20$ and $k = 100$, and evaluate the oracle results of the captions generated by all modes. In Table 3.2, the models with DML obtain the best and second best results on all the evaluation metrics *w.r.t.* both 20 and

TABLE 3.2. Comparison with baselines *w.r.t.* oracle performance (*i.e.*, best-1 quality) on COCO dataset. "#Sample" refers to the number of generated captions for each image. The best and the second best results are highlighted with **bold** and underline, respectively.

Method	#Sample	B@1	B@2	B@3	B@4	R	M	C	S
Div-BS [Vijayakumar et al., 2018]	20	0.837	0.687	0.538	0.383	0.653	0.357	1.405	0.269
POS [Deshpande et al., 2019]		0.874	0.737	0.593	0.449	0.678	0.365	1.468	0.277
AG-CVAE [Wang et al., 2017a]		0.834	0.698	0.573	0.471	0.638	0.309	1.259	0.244
Seq-CVAE [Aneja et al., 2019]		0.870	0.727	0.591	0.445	0.671	0.356	1.448	0.279
COS-CVAE [Mahajan and Roth, 2020]		0.903	0.771	0.640	0.500	0.706	0.387	1.624	0.295
AoANet-DML (Ours)		**0.917**	**0.799**	**0.682**	**0.554**	**0.734**	**0.418**	**1.734**	**0.328**
Transformer-DML (Ours)		0.915	0.788	0.663	0.526	0.726	0.417	1.704	0.325
Div-BS [Vijayakumar et al., 2018]	100	0.846	0.698	0.555	0.402	0.666	0.372	1.448	0.290
POS [Deshpande et al., 2019]		0.909	0.787	0.672	0.550	0.725	0.409	1.661	0.311
AG-CVAE [Wang et al., 2017a]		0.883	0.767	0.654	0.557	0.690	0.345	1.517	0.277
Seq-CVAE [Aneja et al., 2019]		0.922	0.803	0.691	0.575	0.733	0.410	1.695	0.320
LNFMM [Mahajan et al., 2020]		0.920	0.802	0.695	0.597	0.729	0.402	1.705	0.316
COS-CVAE [Mahajan and Roth, 2020]		0.942	0.842	0.739	0.633	0.770	0.450	1.893	0.339
AoANet-DML (Ours)		**0.947**	**0.850**	**0.752**	**0.652**	**0.782**	**0.479**	**1.960**	**0.356**
Transformer-DML (Ours)		0.946	0.849	0.750	0.649	0.780	0.474	1.953	0.354



(A) (B) (C)

FIGURE 3.5. Comparison between our DML and SoTA baseline COS-CVAE. We show the oracle results of (A) CIDEr, (B) SPICE and (C) METEOR with different numbers of samples.

100 samples. Moreover, we compare our DML paradigm with the SoTA baseline COS-CVAE by calculating the oracle scores of CIDEr, SPICE and METEOR with different numbers of samples. In Figure 3.5, the proposed DML consistently outperforms COS-CVAE on all settings. The high gain in quality metrics demonstrates that the proposed DML successfully captures the rich and representative modes in the training corpus.

Results of Individual Mode Embedding

To assess the quality of captions generated using different modes, we calculate the evaluation scores of captions generated from some representative modes that are manually selected with distinct patterns. In Table 3.3, on mode-58, Transformer-DML outperforms the original Transformer on all reference-based metrics. Moreover, the Transformer-DML with different mode embeddings can achieve better or at least competitive performance compared with the original Transformer in terms of ClipScore. We further report the best-performed modes *w.r.t.* CIDEr, SPICE and ClipScore for AoANet-DML. In Table 3.3, the results show that the proposed DML can gain higher scores of CIDEr, SPICE and ClipScore by choosing the suitable mode embedding. These results suggest that our DML not only generates diverse results, but the results of each individual mode also yield high quality.

TABLE 3.3. The performance of image captions generated by different modes. For Transformer-DML, we select some representative modes that exhibit clear language patterns or semantic complexities as shown in Figure 3.4. For AoANet-DML, we show the best-performed modes in terms of CIDEr, SPICE and ClipScore, respectively (highlighted with underline). The original results of Transformer and AoANet are obtained using the code and settings in [Luo]. The best results of the two captioning methods on each metric are highlighted with **bold**.

Models	Mode	B@1	B@2	B@3	B@4	R	M	C	S	ClipScore
Transformer [Vaswani et al., 2017]	N/A	0.751	0.589	**0.448**	0.338	0.558	0.276	1.119	0.207	0.990
Transformer-DML	3	0.671	0.474	0.328	0.221	0.487	0.245	0.871	0.192	1.090
	7	0.677	0.497	0.359	0.253	0.499	0.247	0.958	0.189	0.967
	32	0.665	0.461	0.315	0.213	0.441	0.230	0.840	0.190	**1.121**
	43	0.658	0.481	0.339	0.238	0.479	0.234	0.909	0.175	1.098
	58	**0.752**	**0.590**	**0.448**	**0.340**	**0.559**	**0.277**	**1.124**	**0.208**	0.945
AoANet [Huang et al., 2019]	N/A	**0.770**	**0.613**	**0.476**	**0.368**	**0.571**	0.283	**1.173**	0.213	0.969
AoANet-DML	49	0.762	0.601	0.458	0.356	0.567	0.279	1.157	0.209	0.972
	64	0.658	0.461	0.315	0.212	0.529	**0.284**	0.768	**0.215**	1.131
	44	0.747	0.583	0.440	0.339	0.557	0.279	1.123	0.210	**1.195**

Further Comparison with COS-CVAE

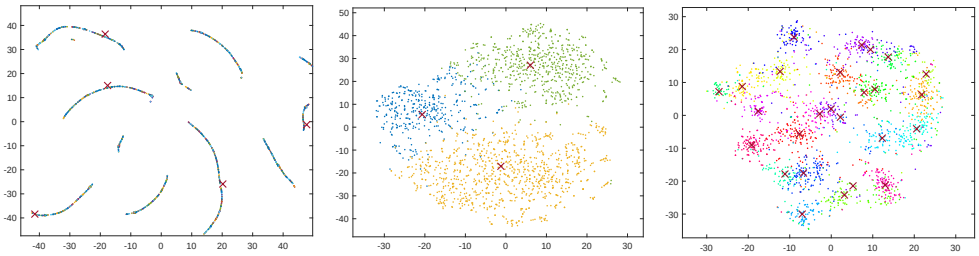
To further evaluate the effectiveness of our DML paradigm, we run experiments using the UpDown [Anderson et al., 2018a] model (a two-layer LSTM with a visual attention module) for our MIC branch, which is also the same language generation model used by COS-CVAE. The oracle performance of this model is 1.688 and 1.942 in terms of CIDEr for 20 and 100 samples, respectively, which still outperforms the COS-CVAE by a large margin. In fact, UpDown is a strong model that achieves compatible performance with a 6-layer Transformer model in a general image captioning setting (1.099 CIDEr vs. 1.114 CIDEr on Karpathy's test split [Karpathy and Fei-Fei, 2015]), which means that two-layer LSTMs may already have enough capacity for the COCO dataset. Moreover, considering that COS-CVAE requires a pre-processing step to construct pseudo supervisions with the help of a pretrained joint vision-language embedding model, the proposed end-to-end learning method could be more convenient to use than COS-CVAE.

3.2.2.5 Performance Analysis of the CdVAE branch

In this part, we investigate the effect of our mode assignment strategy (Section 3.2.1.2) and the fully non-autoregressive (NAT) objective (Section 3.2.1.3) used by the CdVAE branch of DML. Specifically, we train Transformer models with a codebook size of $k = 64$ on three different settings: 1) DML w/o NAT objective; 2) DML w/o Hungarian assignment; 3) the proposed DML, and visualize their mode embeddings in the codebook and caption embeddings from the mode encoder output using t-SNE.

As shown in Figure 3.6(a), DML w/o NAT only activates 5 of 64 mode embeddings, *i.e.*, all caption embeddings are assigned to the 5 modes. Moreover, the caption embeddings of different modes are mixed together (zoom in to see the details) and are distributed among the twisted spaces nearby the mode embeddings, which means that the mode embeddings and the caption embeddings may not contain useful information. The low oracle CIDEr score (1.207) also verifies this hypothesis.

In Figure 3.6(b), we show that DML w/o Hungarian assignment activates even fewer modes (*i.e.*, only three), but the caption embeddings now have a more distinct distribution compared with the result in Figure 3.6(a). This shows the importance



(A) (B) (C)

FIGURE 3.6. Effects of the Hungarian mode assignment and the fully non-autoregressive decoding objective. Here, (A) and (B) are DML without NAT objective and Hungarian assign, respectively. (C) is the final DML we design. We visualize the mode embeddings (red "×") and the caption embeddings ("·" with other colors) on the test set by t-SNE. The inactivated modes are not shown for clarity.

of the fully NAT objective in CdVAE. Still, due to the small number of modes, the oracle CIDEr score in this setting is only 1.211, indicating a limited diversity.

Lastly, in Figure 3.6(c), our DML obtains 29 effective mode embeddings, and the caption embeddings are tightly distributed around their corresponding mode embeddings. Moreover, our model achieves an oracle CIDEr of 1.871. These results demonstrate the effectiveness of the Hungarian mode assignment and the fully non-autoregressive objective for learning distinct and representative mode embeddings.

3.2.2.6 Ablation Studies on the CdVAE branch

In this section, we provide more analysis on the CdVAE branch of the proposed DML method. Our experiments are based on the Transformer-DML model with a codebook size of 64.

Asymmetric Batch Size

As we mentioned in Section 3.2.1.4, the training of the CdVAE branch is much harder than the training of the MIC branch. Thus, we adopt asymmetric batch sizes when training them jointly. Moreover, the gradient from the CdVAE branch will not be back-propagated to the image encoder, so that the whole MIC branch is trained with a batch size n times smaller than that of the CdVAE branch. Here, we show how this strategy benefits the training of our Transformer-DML model. Specifically, during training, we change the number of sampled captions for the MIC branch as well as the total batch size, and the rest settings are the same as in Section 3.2.2.2. The results are shown in Table 3.4.

The first row in Table 3.4 is the performance of our default setting, which yields the best results among all metrics, and also obtains a larger number of effective modes than other settings. When increasing the number of sampled captions from 1 to 5 for the MIC branch, the performance of Transformer-DML drops clearly. We believe this is due to the over-fitting of the MIC branch caused by the large batch size. When decreasing the total batch size from 64 to 16, we observe a clear reduction in the number of effective modes, which means the CdVAE branch is not sufficiently trained. Moreover, with such a small batch size, the MIC branch will also suffer from under-fitting if the number of sampled captions is small, *i.e.*, oracle CIDEr drops from 1.799 to 1.760 when reducing the number of sampled captions from 5 to 2. These results

TABLE 3.4. Oracle results of different batch sizes for CdVAE branch and MIC branch. "#effective modes" indicates the total number of modes that have ever been used during the whole training process. "#sampled caps" refers to how many captions per image are sampled to optimize the MIC branch.

Batch size	#sampled caps	B1	B2	B3	B4	R	M	C	S	#effective modes
64	1	0.932	0.826	0.719	0.608	0.761	0.454	1.871	0.342	29
64	2	0.929	0.820	0.712	0.599	0.756	0.445	1.833	0.336	28
64	5	0.921	0.805	0.692	0.574	0.743	0.430	1.779	0.331	24
16	2	0.917	0.799	0.683	0.561	0.737	0.427	1.760	0.327	15
16	5	0.925	0.810	0.698	0.580	0.749	0.441	1.799	0.335	17

TABLE 3.5. Oracle results of different masking strategies for CdVAE branch. "#effective modes" is the total number of modes that have ever been used during the whole training process. "$1.0 \rightarrow 0.0$" means we gradually reduce the masking probability from 1 to 0 throughout the training process.

Mask probability	B1	B2	B3	B4	R	M	C	S	#effective modes
1.0	0.932	0.826	0.719	0.608	0.761	0.454	1.871	0.342	29
0.5	0.833	0.675	0.513	0.342	0.633	0.337	1.384	0.262	6
$1.0 \rightarrow 0.0$	0.879	0.740	0.604	0.448	0.685	0.377	1.579	0.294	9
0.0	0.790	0.619	0.439	0.270	0.590	0.301	1.207	0.228	5

show the importance of balancing the training of the two branches in the proposed DML.

Masking Strategy of the Masked Decoder $\mathcal{D}_m$

When training the CdVAE branch, we use a fully non-autoregressive objective, where the input of the masked decoder $\mathcal{D}_m$ are all `[MASK]` tokens. This prevents the model from using the mode information leaked from the ground-truth tokens, and we find it greatly benefits the learning of modes. Here we evaluate the performance of several different masking strategies, including random masking the input caption using a fixed probability, and random masking with a dynamically changed probability. The results are shown in Table 3.5.

From the table, the full masking strategy, *i.e.*, a mask probability of 1.0, leads to the best performance, which is also the default setting in DML. Moreover, introducing any additional information to the CdVAE branch, *i.e.*, a mask probability less than 1.0, will severely hamper the learning of modes, where the number of effective modes drops clearly from 29 to less than 10, and the oracle performances also drop significantly.

3.3 Method and Experiments in Specific (Medical) Domain

While natural image captioning focuses on generating descriptions for general images, medical report generation targets the specific context of medical imaging data. Although solutions developed for natural image captioning provide a foundation, additional modifications are necessary to optimize performance in the medical domain. These modifications may include incorporating domain-specific knowledge, adapting models to handle medical terminology and semantics, and ensuring accurate and comprehensive reporting of medical findings and diagnoses. To bridge this gap and tailor

solutions to the medical context, we propose a new single-for-multiple (S4M) framework.

3.3.1 Method

The overview of our single-for-multiple (S4M) framework is shown in Figure 3.7, which contains a basic encoder-decoder report generation model (Base), a Radiology-informed Knowledge Aggregation (RadKA) branch, and an Implicit Prior Guidance (IPG). The RadKA branch enhances the base model by incorporating radiology-specific knowledge. The IPG branch improves the recognition capability of the image encoder by aligning multimodal semantics, which would be removed in inference. More details are in the following.

3.3.1.1 Preliminaries

Radiology Report Generation

In this part, we introduce a formal definition of radiology report generation. Let x denote a training radiology image while y denote the ground-truth report for x. During the training, our goal is to discover the optimal model $\hat{f}$ from a set of potential models $\mathcal{F}$, such that it minimises the average prediction loss. This loss can be determined by a specified loss function $\mathcal{L}$. Mathematically, the function can be defined as

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{x,y} \mathcal{L}(f(x), y), \tag{3.7}$$

where n is the number of training data. Typically, sequence generation models are trained using the autoregressive Teacher Forcing scheme, with the objective of maximising the probability of the ground-truth token w_t given all previous ground-truth tokens $w_{i<t}$. Formally, the corresponding loss function would be formulated as

$$\mathcal{L}(f(x), y) = -\log p(y|x) = \sum_{t=1}^{T} -\log p(w_t|w_{i<t}, x). \tag{3.8}$$

Here, w_t is the t-th token in the sequence y, and T is the total number of word tokens in y. Note that in inference, instead of relying on previous ground-truth tokens, we generate predictions for the next token based on the tokens that have been previously predicted in an autoregressive manner.

Building General Radiology-informed Knowledge Set

We seek to construct a general knowledge set $\mathfrak{S}$ that covers the most common abnormalities or findings in the radiology reports. For convenience, we call each item in this knowledge set a topic like [Liu et al., 2021a; Zhang et al., 2020]. While they develop a knowledge graph, the relations (edges) between topics (nodes) cannot be updated during training, which limits its effectiveness for exploiting implicit relationships. Furthermore, this knowledge graph focuses on chest X-rays only, which restricts its applicability to other body parts and broader use cases. Thus, we have reorganised the topics in our knowledge set such that they cover the medical terminologies relevant to a broad range of body parts without pre-defined and fixed relations.

To better build a generic knowledge base, following [Willemink et al., 2020], we adopt one of the most useful natural language processing methods, called topic modelling, on the database, that seeks to characterise the knowledge for each body part

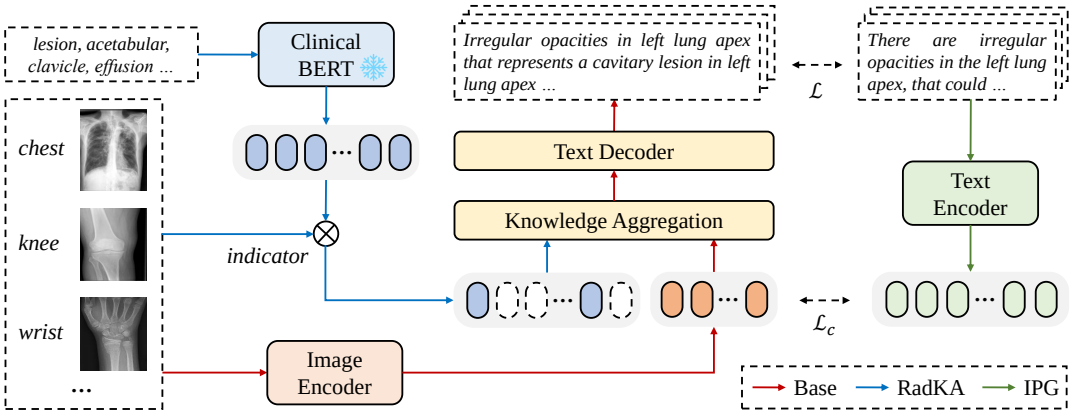


FIGURE 3.7. Overall of S4M. The framework of our base model is highlighted by the arrow in red. The blue and green arrows refer to Radiology-informed Knowledge Aggregation (RadKA) and Implicit Prior Guidance (IPG) branches, respectively. Notably, IPG would be removed after training, while RadKA would be retained. The "$\otimes$" denotes the indication function.

with a series of topics namely $\mathcal{G}$. Specifically, we first use spacy [Neumann et al., 2019], currently the most popular entity detection tool, to extract the medical entities and obtain the 50 most frequent words for each body part. Then $20 \sim 30$ most critical words are further filtered by radiologists to create the existing knowledge base. Mathematically, the general set $\mathfrak{S}$ can be defined as $\mathfrak{S} = \mathcal{G}_1 \cup \mathcal{G}_2 \cup ....$

3.3.1.2 Radiology-informed Knowledge Aggregation

Body Region-aware Knowledge Selection

Understanding the context and clinical nuances in radiology reports requires deep medical knowledge and expertise. Models may lack the comprehensive understanding needed to generate reports that incorporate relevant clinical information, leading to inaccuracies or missing crucial details. To address this, we introduce a condition signal c to selectively activate a set of topics based on the anatomical regions associated with the input image. By doing so, our model is able to filter out topics that are irrelevant to the specific region before conducting reasoning between the given image x and the general knowledge set $\mathfrak{S}$. Specifically, we devise an indicator $\mathbb{1}(\cdot)$, which enables the selection of the topics in $\mathfrak{S}$. Formally,

$$\mathcal{G} = \mathbb{1}(\varphi(\mathfrak{S})|c(x)), \tag{3.9}$$

where φ is the pre-trained knowledge feature extractor (*i.e.*, Clinical BERT) while $\mathcal{G}$ is the subset of the general knowledge $\mathfrak{S}$ (*i.e.*, $\mathcal{G} \subseteq \mathfrak{S}$), denoting the resulting selected topics. Here, $c(x)$ is a tag of the body part a given image x belongs to (*e.g.*, "*chest X-ray*"), which is manually predefined in advance. In the clinical context, images and their corresponding body parts are inherently paired. For instance, if the condition signal of an input image is predefined as a "*chest X-ray*", we choose a specific set of topics – namely, "*airspace disease, atelectasis, calcinosis, cardiomegaly, cicatrix, edema, effusion, emphysema, fractures, hernia, hypoinflation, lesion, medical device, normal, opacity, other, pneumonia, pneumothorax, scoliosis, thickening*" – to represent the knowledge base for this image. Notably, the topics selected for different samples belonging to the same body part are completely identical. In our work, we

define six different tags, including *chest X-ray*, *abdomen X-ray*, *knee X-ray*, *hip X-ray*, *wrist X-ray*, and *shoulder X-ray*.

Intra-region Knowledge Aggregation

For knowledge aggregation, our idea is to design a learnable aggregation function π with a capacity for cross-modal reasoning, allowing it to determine the most relevant topics between the knowledge set $\mathcal{G}$ and the input image x. A straightforward way is adopting scaled dot product attention, which enables the topics and images to interact with one another. Concretely, we design a knowledge-image co-attention module, *i.e.*, a three-layer Transformer. Each Transformer layer contains a multi-head attention (MHA) before a layer normalisation (LN) followed by a Feed-Forward Network (FFN), which can be defined as

$$\pi(x, \mathcal{G}) = \text{FFN}(\text{LN}([\phi(x); \mathcal{G}] + \text{MHA}([\phi(x); \mathcal{G}]))), \tag{3.10}$$

where ϕ refers to the image encoder and $[\cdot;\cdot]$ denotes the operation of concatenation.

Text Decoder

After the knowledge aggregation module, we obtain a series of aggregated tokens and adopt a Transformer-based text decoder for generating the final report. Concretely, the decoding process starts by feeding a special start token `[SOS]` to our text decoder, along with positional embeddings. The decoder uses a self-attention mechanism to process the start token, positional embeddings, and aggregated tokens. Mathematically, the whole process can be defined as

$$f(x, \mathfrak{S}) = \psi(\pi(x, \mathcal{G})) = \arg\max \prod_{t=1}^{T} p(w_t | w_{i<t}, x, \mathcal{G}), \tag{3.11}$$

where $f(x, \mathfrak{S})$ refers to the generated report and ψ is the text decoder. The decoder generates the output report auto-regressively, attending to the aggregated tokens and previously generated words at each step. Then, the decoder applies a softmax function to predict the next word's probability distribution over the entire vocabulary. The process is repeated until an end token is generated or a predefined maximum sequence length T is reached.

3.3.1.3 Implicit Prior Guidance

To enhance the recognition ability of the image encoder, we propose to guide the generation model to focus more on the image features that are considered important/contributive. To this end, we present a feature attribution method $\Phi(\theta_\phi, x)$ to assess the contribution of each image token for the generation task. Then, we define an attribution prior as a function $\Omega : \mathbb{R}^p \to \mathbb{R}$, where p is the number of features of image tokens. For simplicity, we omit the feature dimension d. With an input x, based on the given method $\Phi(\theta_\phi, x)$, Ω assigns penalties to the feature attributions for image encoder ϕ. Mathematically, the objective function can be defined as

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{x,y} \mathcal{L}(f(x, \mathfrak{S}), y) + \lambda \sum_{x} \Omega(\Phi(\theta_\phi, x)), \tag{3.12}$$

where λ is a hyper-parameter for the trade-off between them, and θ_ϕ refers to the learnable parameters of Φ.

Let us assume that the feature of each patch token in x is linked to a meta-feature represented by a scalar value. To represent these scalar-valued meta-features, we can use a vector $m \in \mathbb{R}^p$, where the i-th entry of m corresponds to the significance/-contribution for i-th tokens in image x. Assuming a hypothetical scenario where we have prior knowledge for the values in m, a reasonable option for Ω would be to impose a penalty on the discrepancy between predicted results from our model and the corresponding values in m. Formally, we can rewrite Eq. (3.12) to

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{x,y} \mathcal{L}(f(x, \mathfrak{S}), y) + \lambda \sum_{x} \|\Phi(\theta_\phi, x) - m\|_1. \tag{3.13}$$

Here, $\|\cdot\|_1$ refers to the L_1 norm. However, in most scenarios, it is unrealistic to have an m that is known in advance. Instead, we can take advantage of the function approximation capabilities of neural networks and replace m with deep one $\hat{h}$ from a set of models $\mathcal{H}$. We define such a model to be an Implicit Prior Guidance (IPG) branch. Thus, we jointly learn the generation model and IPG pair $(\hat{f}, \hat{h})$ by optimizing the following objective:

$$\begin{aligned}(\hat{f}, \hat{h}) = &\arg\min_{f \in \mathcal{F}, h \in \mathcal{H}} \frac{1}{n} \sum_{x,y} \mathcal{L}(f(x, \mathfrak{S}), y) \\ &+ \lambda \mathcal{L}_c(\sigma(\Phi(\theta_\phi, x) \circledast x), h(y)),\end{aligned} \tag{3.14}$$

where we define $\Phi(\theta_\phi, x) = \theta_\phi$, which can be regarded as the weights (*i.e.*, importance) of image tokens, and $\circledast$ is the convolution operation. Here, σ is a pooling operation for these image tokens, and we define $h(y)$[2] to be the vector in $\mathbb{R}^d$ resulting from feeding ground-truth report y into h. We are able to approximate a solution to this problem by a contrastive loss $\mathcal{L}_c$. We find empirically that, so long as $\hat{f}$ and $\hat{h}$ are convergent, this procedure accomplishes the following two goals: learning the semantic alignment between provided image x and report y; guiding the generation model to rely more heavily on the image features deemed important by these learned implicit priors.

3.3.2 Experiments

3.3.2.1 Datasets

In experiments, we use a merged dataset that contains paired data *w.r.t.* different body parts, including a widely used IU-Xray [Demner-Fushman et al., 2016] (chest) and five private datasets for abdomen, knee, hip, wrist and shoulder.

IU-Xray

IU-Xray [Demner-Fushman et al., 2016] is a widely used publicly accessible dataset in the radiology report generation. It consists of 3,955 fully de-identified radiology reports, each paired with frontal and/or lateral chest X-ray images. Following [Chen et al., 2020g; Li et al., 2023b], we remove cases that contain only a single image and then divide the entire dataset into train, validation, and test sets with a number of 2069/296/590, respectively.

[2] $h(\cdot)$ refers to the text encoder in Figure 3.7.

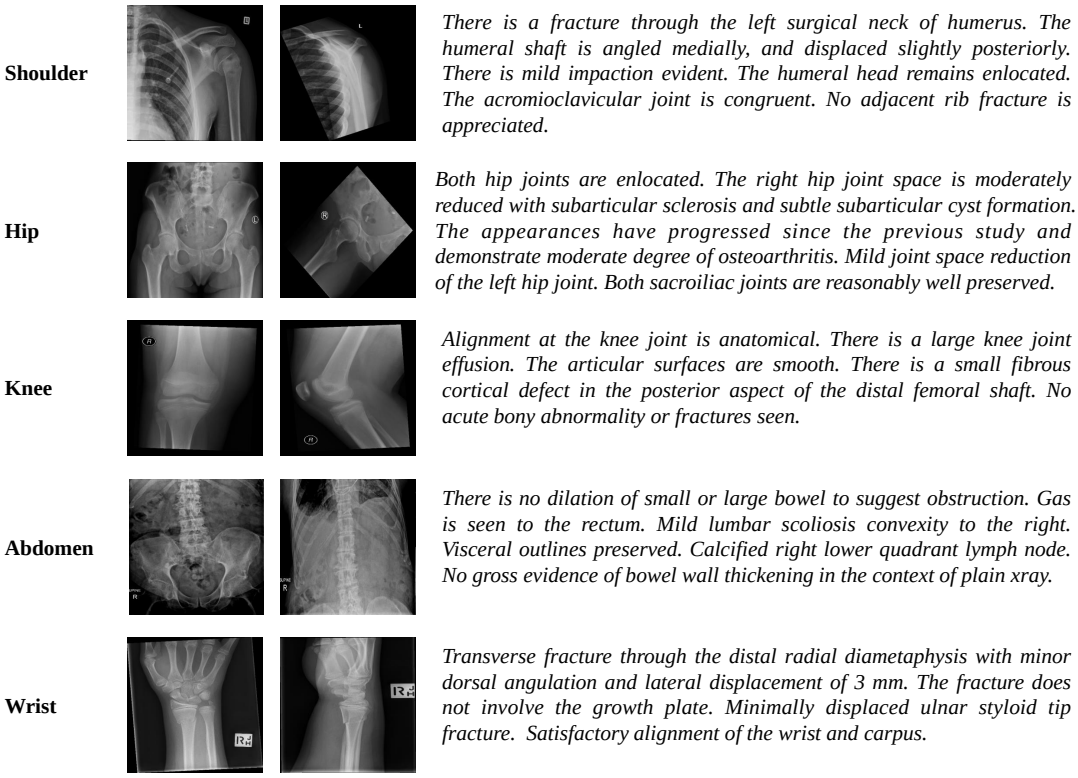


FIGURE 3.8. Examples on our private datasets. Each example contains a frontal image (first column) and another image (second column) with the corresponding radiology report.

Private Datasets

We collected private data on five additional body parts separately, each comprising paired radiology images and corresponding reports. Anonymous Human Research Ethics Committee provided ethics approval for private data used in this study. Unlike IU-Xray [Demner-Fushman et al., 2016], where one report strictly corresponds to two view images, we introduce private data, with each report pair corresponding to 1-8 images of different views. To be consistent with IU-Xray, we filter out the data with only a single radiology image while selecting two images (frontal and partial X-ray images) by radiologists if more than two. After removing the extremely long or short reports and keeping each report from 30 to 60 tokens, we finally select $\sim 3,000$ patients for each body part and the ratio of train/val/test is 70%/15%/15%. Here, we provide some samples on our private datasets in Figure 3.8.

Note that our work focuses on the problem of the dilemma between the diversity of body parts and the limited availability of medical data, which is practical and common in the medical domain. There are two reasons that we do not use the MIMIC-CXR dataset: 1) the size of the MIMIC-CXR dataset is notably large (approximately 300,000 pairs), which contrasts with our thesis's emphasis on a limited data scenario; 2) it only contains chest images and their associated reports while our thesis considers data for various body parts. Thus, the MIMIC-CXR dataset is not well-suited for the problem studied in this thesis.

3.3.2.2 Metrics and Implementation Details

Evaluation Metrics

To evaluate the quality of generated reports, we employ widely used evaluation metrics for image captioning, *i.e.*, BLEU [Papineni et al., 2002] and CIDEr [Vedantam et al., 2015]. We additionally use the CLIP-based metric called CLIPScore [Hessel et al., 2021], which can assess whether the generated reports are semantically aligned with given images, even when they are significantly different from the reference reports. To substantiate the enhanced capacity for semantic understanding offered by the image encoder, we undertake the linear classification probing evaluation using the CheXpert [Irvin et al., 2019] dataset, which contains five individual binary labels: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. For this process, we fix the image encoder, which has been trained on our S4M, and exclusively train a randomly initialised linear classification head.

Implementation Details

We adopt ResNet101 [He et al., 2016], pre-trained on ImageNet [Deng et al., 2009], serving as image encoder. Considering the domain disparity between medical and generic texts, we use the tokeniser and text encoder from MedClip [Wang et al., 2022d] for the purpose of embedding the report. The knowledge aggregation network consists of a three-layer Transformer. We configure the dimensions of input patches to 224×224 and incorporate online data augmentation techniques, such as random cropping and flipping, to expand the X-ray training dataset. We limit the maximum epochs to 100 and use the Adam optimiser [Kingma and Ba, 2014] with a weight decay parameter of 1e-4. The learning rates are set at 5e-5 and 1e-4 for the image and text encoders, and the remaining trainable parameters, respectively. Besides, based on the findings from our ablation study, we empirically set the hyper-parameter λ to 1.0. Our experiments are conducted using a single A100 GPU.

Details of Knowledge Base

Here, we used different colours to highlight shared topics across the six body parts. The results show that there are many topics commonly used, even across different body parts. This finding indicates that our knowledge set has a relatively general scope. Topics on our general knowledge set $\mathfrak{S}$ include:
{abdomen, acetabular, acromioclavicular, acute, airspace disease, anatomical, angulation, atelectasis, bilateral, bone, bony, bowel, calcification, calcinosis, cardiomediastinal, cardiomegaly, carpal, cast, change, changes, cicatrix, clavicle, colon, compartment, complication, consolidation, contours, cuff, degenerative, dislocation, displacement, distal, dorsal, edema, effusion, emphysema, enlocated, evidence, faecal, femoral, femur, fracture, fractures, gas, glenohumeral, glenoid, head, healing, hernia, hip, humeral, humerus, hypoinflation, identified, inferior, intact, interval, joint, knee, lateral, lesion, limits, loading, loops, lucency, lumbar, lung, material, medical device, mild, moderate, nonspecific, normal, obstruction, opacity, other, patella, patellar, pelvic, pelvis, periprosthetic, plate, pleural, pneumonia, pneumothorax, projection, prosthesis, proximal, pubic, quadrant, radial, radio-carpal, radius, rectum, replacement, ring, sacroiliac, satisfactory, scaphoid, sclerosis, scoliosis, shoulder, situ, soft, space, stomach, styloid, subacromial, subdiaphragmatic, supine, suprapatellar, surgical, swelling, symphysis, thickening, tissue, tissues, transverse, tuberosity, ulnar, visualised, wrist}

Topics on each body part namely $\mathcal{G}$ and we highlight the overlapped topics across different body parts in various colours:

- Chest = {*airspace disease, atelectasis, calcinosis, cardiomegaly, cicatrix, edema, effusion, emphysema, fractures, hernia, hypoinflation, lesion, medical device, normal, opacity, other, pneumonia, pneumothorax, scoliosis, thickening*}
- Abdomen = {*abdomen, bowel, cardiomediastinal, colon, consolidation, contours, degenerative, evidence, faecal, gas, limits, loading, loops, lumbar, lung, material, moderate, nonspecific, obstruction, pleural, projection, quadrant, rectum, stomach, subdiaphragmatic, supine, surgical, tissue*}
- Knee = {*acute, alignment, anatomical, changes, compartment, complication, degenerative, dislocation, effusion, evidence, femoral, fracture, gas, joint, knee, lateral, lucency, mild, moderate, patella, patellar, prosthesis, proximal, replacement, satisfactory, situ, soft, suprapatellar, swelling, tissue, tissues*}
- Hip = {*acetabular, acute, alignment, bilateral, bone, bony, degenerative, enlocated, femoral, femur, fracture, fractures, hip, identified, intact, joint, lucency, mild, moderate, pelvic, pelvis, periprosthetic, proximal, pubic, ring, sacroiliac, sclerosis, symphysis*}
- Wrist = {*acute, alignment, anatomical, angulation, bony, carpal, cast, degenerative, displacement, distal, dorsal, fracture, healing, intact, interval, lateral, mild, plate, radial, radio-carpal, radius, scaphoid, styloid, swelling, tissue, transverse, ulnar, wrist*}
- Shoulder = {*acromioclavicular, acute, alignment, bony, calcification, change, clavicle, cuff, degenerative, dislocation, fracture, fractures, glenohumeral, glenoid, head, humeral, humerus, identified, inferior, intact, joint, lateral, proximal, shoulder, space, subacromial, tissue, tuberosity, visualised*}

3.3.2.3 Comparison with State-of-the-arts

Specialized Baselines

We compare our S4M with the recent best radiology report generation methods, including R2Gen [Chen et al., 2020g], R2GenCMN [Chen et al., 2022b], MSAT [Wang et al., 2022c], DCL [Li et al., 2023b] and METransformer [Wang et al., 2023]. In addition, we also consider a widely used natural image captioning method (*i.e.*, Transformer [Vaswani et al., 2017]) as another baseline. To begin with, we individually optimise our model and each baseline in a separate training setting. For a fair comparison, we adopt the training batch size (bs) of 16, which is a commonly used setting in the report generation task[3]. In Table 3.6, we can observe that compared with the specialised baselines, our S4M achieves superior results in terms of average BLEU4 (B4) and CIDEr scores, with values of 0.103 and 0.269, respectively. This indicates that the two proposed branches have a beneficial effect even in the specialised setting.

[3]We also experiment with increasing the batch size of the baselines to improve their performance, but we observe that it only results in performance comparable to bs = 16.

TABLE 3.6. Comparison with the recent specialised models on six datasets covering different body parts. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance. We highlight the best results on specialised models with underline while the best results on all models (both specialised and generalist) with **bold**.

	Chest (IU-Xray)		Abdomen		Knee		Hip		Wrist		Shoulder		Ave	
	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr
specialized models														
Transformer	0.162	0.511	0.108	0.261	0.079	0.151	0.077	0.137	0.086	0.129	0.088	0.192	0.100	0.230
R2Gen	0.165	0.430	0.105	0.248	0.077	0.193	0.082	0.210	0.093	0.288	0.082	0.174	0.101	0.257
R2GenCMN	0.170	0.641	0.102	0.161	0.083	0.164	0.083	0.220	0.087	0.212	0.082	0.134	0.101	0.255
MSAT	0.171	0.394	0.105	0.275	0.082	0.135	0.081	0.235	0.081	0.180	0.080	0.173	0.100	0.232
DCL	0.163	0.586	-	-	-	-	-	-	-	-	-	-	-	-
METransformer	0.172	0.435	-	-	-	-	-	-	-	-	-	-	-	-
S4M (ours)	0.163	0.609	0.106	0.196	0.087	0.175	0.086	0.192	0.089	0.243	0.088	0.197	0.103	0.269
generalist models														
R2Gen† (bs=16)	0.084	0.289	0.104	0.280	0.064	0.154	0.074	0.203	0.085	0.217	0.082	0.186	0.082	0.222
R2Gen† (bs=96)	0.147	0.470	0.097	0.271	0.075	0.181	0.080	0.226	0.084	0.258	0.095	0.274	0.096	0.280
R2Gen† (bs=192)	0.114	0.359	0.100	0.271	0.089	0.204	**0.086**	0.238	**0.102**	0.296	0.096	0.277	0.098	0.274
S4M (ours, bs=16)	0.152	0.509	0.108	0.276	0.071	0.166	0.073	0.184	0.079	0.229	0.084	0.220	0.095	0.264
S4M (ours, bs=96)	0.161	**0.700**	0.110	0.292	0.077	0.188	0.084	**0.257**	0.090	0.255	**0.099**	0.272	0.104	0.327
S4M (ours, bs=192)	**0.177**	0.602	**0.118**	**0.327**	**0.093**	**0.242**	0.076	0.215	0.097	**0.305**	0.096	**0.287**	**0.110**	**0.330**

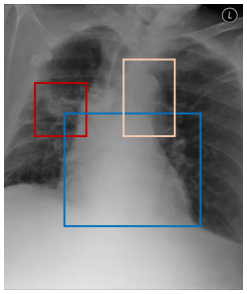


FIGURE 3.9. Reports generated by S4M (ours) and two baselines, *i.e.*, R2Gen and R2Gen†. Note that R2Gen is trained on IU-Xray only while R2Gen† optimised on our merged training dataset. We highlight the specific areas in the radiology image by coloured bounding boxes (red, blue and orange) as well as the corresponding descriptions in reports with the same colour.

Generalist Baselines

To further analyze the performance of S4M, we adapt specialised models into the joint training setting due to the lack of existing generalist baselines. Here, we use all the training data on different subsets for optimisation. To mitigate the impact of different architectures, we select R2Gen [Chen et al., 2020g] as the baseline. The main difference between R2Gen and our base model lies in the text decoder, where R2Gen has an additional Relational Memory (RM) module while our model does not include it. For a fair comparison, we adjust the batch size (bs) to match our setting. Specifically, we increase it from 16 to 96 and 192, which aligns with our own configuration, thereby mitigating the potential performance improvement attributed solely to the larger batch size.

Table 3.6 shows that regardless of bs = 96 or 192, our S4M consistently outperforms R2Gen in terms of both average B4 and CIDEr scores, which demonstrates its effectiveness in generating accurate and high-quality radiology reports. Furthermore, when comparing R2Gen to our S4M in both the specialised and generalist settings, we observe that R2Gen achieves an approximate 9% improvement in CIDEr (0.257 to 0.280) while achieving a comparable result in B4 (0.101 and 0.098). This indicates the positive impact of using diverse and increased training data. In contrast, our S4M achieves even larger improvements in both CIDEr ($\sim$ 22%) and B4 ($\sim$ 6%). These results demonstrate that the gains in performance are not solely attributed to the dataset, but also due to the benefits provided by the proposed two branches.

3.3.2.4 Qualitative Evaluation

To further test the performance of S4M, we assess the quality of reports generated by different methods, including our method and two baselines, *i.e.*, R2Gen and R2Gen†, trained on IU-Xray (chest) and our merged dataset, respectively. In Figure 3.9, we highlight the descriptions in different colours (red, blue and orange), which are semantically aligned with that in the ground-truth (GT) reports. Also, we enclose the corresponding area in the image with the same colour. When considering the prominent area, such as the heart (indicated by the blue box), all three models demonstrate the ability to provide (almost) accurate descriptions. However, R2Gen still tends to generate the "average" descriptions like "*the heart is normal in size*" while R2Gen†

TABLE 3.7. Performance analysis of each component. "Base" denotes the model containing only an image encoder and a text decoder. Both RadKA and RadKA* are applied on top of our Base model, where the RadKA* means RadKA without indicator $\mathbb{1}(\cdot)$ in Eq. (3.9). The IPG is applied only on top of the Base+RadKA.

	Chest (IU-Xray)		Abdomen		Knee		Hip		Wrist		Shoulder		Ave	
	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr
Base	0.129	0.426	0.106	0.268	0.081	0.171	0.076	0.255	0.089	0.258	0.091	**0.278**	0.095	0.276
+RadKA*	0.137	0.474	0.103	0.269	**0.085**	0.169	0.081	0.224	0.088	**0.277**	0.091	0.255	0.098	0.278
+RadKA	0.156	0.537	0.106	0.266	0.081	**0.197**	**0.085**	0.205	0.090	0.238	0.086	0.231	0.101	0.279
+IPG	**0.161**	**0.700**	**0.110**	**0.292**	0.077	0.188	0.084	**0.257**	**0.090**	0.255	**0.099**	0.272	**0.104**	**0.327**

shows improvement due to optimisation with a more diverse dataset. Furthermore, our S4M demonstrates the ability to identify specific and subtle areas (highlighted by red and orange boxes in the image) and generate descriptions that align semantically with the ground truth. For instance, while the baselines fail to capture the details, our model accurately describes "*an opacity in the right lung*", matching the GT description: "*a focal area of opacity in the right midlung zone*". However, despite our method achieving better performance compared to the baselines, it still faces challenges in accurately identifying and describing all tiny areas. From Figure 3.9, in the ground-truth report stating "*there are small bilateral pleural effusions*", our generated report states "*there is no pleural effusion*", indicating a slight mismatch. But it is worth noting that the same phenomenon is observed in both R2Gen and R2Gen† as well.

3.3.2.5 More Discussions

Ablation Study

In this part, we evaluate the performance of our base model with and without the RadKA and IPG branches. In Table 3.7, the results show that our base model with RadKA alone achieves better performance compared with that without any additional branches (*i.e.*, average B4: 0.095 $\rightarrow$ 0.101 while average CIDEr: 0.276 $\rightarrow$ 0.279), which verifies the significance of the radiology-relevant knowledge in report generation task. While RadKA* (without the indicator $\mathbb{1}(\cdot)$) can also achieve improved results compared to the base model (*e.g.*, average B4: 0.095 $\rightarrow$ 0.098), it is surpassed by RadKA (average B4: 0.101). This highlights the importance of region-specific guidance and demonstrates the necessity of incorporating such guidance for better performance. Finally, incorporating the IPG branch further enhances the performance, resulting in the best scores for both average B4 (0.104) and CIDEr (0.327). This demonstrates the contribution of IPG in improving the model performance by leveraging guidance from different modalities, including both images and reports.

Semantic Alignment between Image and Report

Besides the reference-based metrics like BLEU4, which may be influenced by semantically irrelevant factors (*e.g.*, writing style [Chen et al., 2022a]), we seek to directly assess the semantic alignment between the input images and the generated reports. Thus, we calculate a reference-free score, namely CLIPScore [Hessel et al., 2021], for R2Gen [Chen et al., 2020g] (trained on IU-Xray only), R2Gen† (trained on the merged dataset) and our S4M. Notably, as we use MedClip [Wang et al., 2022d] in CLIPScore, which is pre-trained on chest X-Ray datasets, we only evaluate this score on the IU-Xray (chest) dataset because it is open-sourced. Moreover, for a fair comparison, we

TABLE 3.8. CLIPScore on IU-Xray to assess semantic alignment between images and generated reports. The notation † means we optimise the model on our merged training dataset.

	CLIPScore on IU-Xray
R2Gen [Chen et al., 2020g]	77.670
R2Gen†	75.402
S4M (ours)	78.052

TABLE 3.9. Effect of each part for report-image alignment.

Method	CLIPScore on IU-Xray
Base	75.687
+ RadKA	76.865
+ RadKA + IPG	78.052

set the batch size to 96 for both R2Gen† (achieves the best results) and our model. For the specialised R2Gen, we keep the settings of the official code unchanged. In Table 3.8, our S4M outperforms R2Gen with a CLIPScore of 78.052, regardless of whether it is trained on IU-Xray only (77.670) or on our merged dataset (75.402).

Due to the use of multiple body parts, this can easily lead to confusion in the generated reports. To assess the effect of achieving a better report-image alignment aided by the proposed two branches, *i.e.*, Radiology-informed Knowledge Aggregation (RadKA) and Implicit Prior Guidance (IPG), similarly to Table 3.8, we assess the semantic alignment between the input images and the generated reports by calculating the CLIPScore to further validate these contributions. Table 3.9 reveals that RadKA can produce improved CLIPScore compared to the base model (from 75.687 to 76.865), and incorporating the IPG further enhances the performance, resulting in the best CLIPScore. This proves the capability of our RadKA and IPG modules in mitigating model confusion when dealing with datasets from multiple body parts, therefore generating more accurate and realistic reports.

Performance Analysis of Image Encoder

To test the effect of our IPG, we seek to directly investigate the recognition ability of image encoders in report generation models. To this end, we directly add a classification head on top of the image encoder (*i.e.*, linear probing) for CheXpert [Irvin et al., 2019], a multi-label classification dataset. As shown in Table 3.10, our S4M model obtains an 81.252 AUC score that outperforms both R2Gen (77.435) and R2Gen† (79.213), which indicates the ability of our model to correctly recognise and classify different medical diseases within the input radiology images. Furthermore, to delve deeper into the influence of our IPG branch, we evaluate the performance of the S4M without incorporating IPG during training. In this case, the AUC score decreases to 80.405, further demonstrating the effectiveness of IPG in enhancing the recognition ability of our model. Interestingly, in Table 3.8, R2Gen performs better than R2Gen†, while the results are reversed in Table 3.10. This observation suggests that R2Gen trained solely on the IU-Xray dataset is susceptible to overfitting and exhibits relatively weaker recognition capability compared to the model trained on a more diverse and merged dataset, which further underscores the importance of training a single model on a broader and merged dataset to achieve better performance.

TABLE 3.10. Results of linear probing on CheXpert to evaluate the recognition ability of the image encoder. The notation † means we optimise the model on our merged training dataset.

	AUC score
R2Gen [Chen et al., 2020g]	77.435
R2Gen†	79.213
S4M w/o IPG	80.405
S4M (ours)	81.252

TABLE 3.11. Impact of hyper-parameter λ in Eq. (3.14)

λ	B4	CIDEr
0.5	0.108	0.317
1.0	0.110	0.330
1.5	0.101	0.272

Impact of Hyper-parameter λ in Eq. (3.14)

In Table 3.11, When the value of λ is small, such as $\lambda = 0.5$, the performance of our model is suboptimal. This is because there is insufficient exploitation of the semantic alignment across different modalities (*i.e.*, images and reports). As we increase the value of λ, the performance of our S4M reaches its peak at $\lambda = 1.0$. However, beyond that point, the performance starts to degrade. To strike a balance between the contrastive loss and the captioning loss, we set the weighting parameter λ to a value of 1.0 for all our experiments.

Effect of Feature Extractors

In our S4M framework, the knowledge feature extractor is initialised using Clinical BERT. This choice is informed by the need to effectively extract features from specific medical terms. Clinical BERT, trained extensively on a vast corpus of clinical text, offers a robust choice for such feature extraction. Meanwhile, within the IPG, the text encoder is initialised with MedCLIP. The core objective here is to learn from reports and discern implicit priors that will guide the model's focus towards important image regions. Given that MedCLIP has undergone rigorous training on a comprehensive dataset of medical images and reports, it presents an ideal foundation for initiating the representation learning required for these reports.

To empirically assess the contributions of the two medical-specific pre-training models, we modified our S4M, substituting these two pre-training models with a generic BERT pre-training [Devlin et al., 2019]. For a fair comparison, we set all the batch sizes to 96. Table 3.12 indicates that when initialised with this general-domain BERT, our S4M model experiences a performance degradation of approximately 22% in CIDEr scores (declining from 0.324 to 0.302) and a 4% decrease in B4 scores (from 0.104 to 0.100). These findings underscore the significance of medical-specific initialisation. Nevertheless, even without it, our S4M significantly outperforms the base model. This suggests that the performance gains of the S4M framework are attributed not only to domain-specific initialisation but also to the RadKA and IPG modules we introduced.

TABLE 3.12. Effect of different feature extractors.

Method	B4	CIDEr
Base	0.095	0.276
S4M with BERT init.	0.100	0.302
S4M	0.104	0.327

TABLE 3.13. Effect of different feature extractors.

Method	B4	CIDEr
R2Gen	0.096	0.280
R2Gen with predefined tags	0.097	0.284

Risk of Information Leakage from $c(x)$

To examine the absence of information leakage, we feed the tag information $c(x)$ of each input image x into the existing well-known R2Gen method and observe the impact of the performance. From Table 3.13, the inclusion of input tags does not lead to much-improved performance for R2Gen, which implies that the presence of input tags does not result in information leakage. On the contrary, they can be considered as medical-specific priors, but need a well-designed approach (*e.g.*, our RadKA) to unleash their inherent potential.

Effectiveness for Limited Data Scenarios

The proposed S4M framework is specifically designed to be effective in scenarios with limited data. To further verify this claim, we include experiments that compare the performance of our S4M framework under varying sizes of training samples. As shown in Table 3.14, our S4M consistently outperforms the base model no matter how many percentages of training samples are used. Meanwhile, we find that the B4 gain achieved by our S4M increases from 0.9% to 1.8% when the number of training samples decreases from 100% to 1%. It suggests that our S4M is particularly advantageous in limited data scenarios. Besides, it is really encouraging to see that our S4M is able to achieve even superior performance with fewer training samples, even one-fifth. For example, S4M with 20% training samples achieves 10% B4 scores, which is better than the 9.5% of the base method with 100% training samples. This further showcases the potential of S4M for handling data-limited medical data analysis scenarios.

More Quantitative Results

To assess the quality of the generated captions, we use four widely used evaluation metrics, *i.e.*, BLEU Papineni et al. [2002], ROUGE Lin [2004], METEOR Banerjee and Lavie [2005] and CIDEr Vedantam et al. [2015]. As shown in Table 3.15, we report the average scores of all the above evaluation metrics. The results exhibit that regardless of bs = 96 or 192, our S4M consistently outperforms R2Gen in terms of all the average scores (except for ROUGE-L), which demonstrates its effectiveness in generating accurate and high-quality radiology reports. Specifically, when comparing R2Gen to our S4M in both the specialized and generalist settings, we observe that the improvements of R2Gen are 2.1%, −0.4%, −2.6%, −2.9%, 5.6%, −2.2% and 8.9%

TABLE 3.14. Results of B4 using partial training data.

Method	1%	20%	100%
Base	0.042	0.086	0.095
S4M	0.060 (+0.018)	0.100 (+0.014)	0.104 (+0.009)

for BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L and CIDEr, respectively[4]. In contrast, our S4M achieves even larger improvements in these evaluation metrics about 8.3%, 7.4%, 6.7%, 6.8%, 6.9%, −0.6% and 22.7% separately. Moreover, we also report the values of all the evaluation metrics on these six datasets from Tables 3.16 to 3.21.

3.4 Conclusion

In this chapter, we address biases in image captioning models that often lead to the generation of "average" captions dominated by common words or phrases. We introduce a Discrete Mode Learning (DML) paradigm to mitigate this issue, wherein we explore diverse modes within the training caption corpus. The DML paradigm involves learning a codebook with "mode embeddings", enabling image captioning models, including Transformer and AoANet, to generate diverse and high-quality captions by leveraging multiple rich modes.

Simultaneously, in the medical area, we contribute to radiology report generation, aiming to streamline the diagnostic process for radiologists. Our proposed Single-for-Multiple (S4M) framework demonstrates its capability to produce coherent reports across various body parts. By integrating general radiology knowledge through Radiology-informed Knowledge Aggregation (RadKA) and enhancing cross-modal alignment via Implicit Prior Guidance (IPG), our framework outperforms both specialized and generalist baselines. The experiments, conducted on a merged dataset encompassing public and private datasets across different body regions, affirm the effectiveness of our S4M approach in generating superior reports.

[4]Note that for a fair comparison, we compare the highest results for both R2Gen and ours.

TABLE 3.15. **Average** results on the six datasets compared with the recent specialized models. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1 (Ave)	BLEU-2 (Ave)	BLEU-3 (Ave)	BLEU-4 (Ave)	METEOR (Ave)	ROUGE-L (Ave)	CIDEr (Ave)
specialized models							
Transformer	0.368	0.223	0.147	0.100	0.134	0.305	0.230
R2Gen	0.374	0.229	0.149	0.101	0.141	**0.312**	0.257
R2GenCMN	0.371	0.229	0.150	0.101	0.138	0.307	0.255
MSAT	0.393	0.237	0.151	0.100	0.139	0.302	0.232
S4M (ours)	0.370	0.227	0.150	0.103	0.144	**0.312**	0.269
generalist models							
R2Gen† (bs=16)	0.345	0.200	0.126	0.082	0.133	0.289	0.222
R2Gen† (bs=96)	0.382	0.228	0.145	0.096	0.149	0.301	0.280
R2Gen† (bs=192)	0.369	0.225	0.145	0.098	0.146	0.305	0.274
S4M (ours, bs=16)	0.363	0.217	0.140	0.095	0.144	0.296	0.264
S4M (ours, bs=96)	0.383	0.231	0.151	0.104	0.149	0.306	0.327
S4M (ours, bs=192)	**0.401**	**0.244**	**0.160**	**0.110**	**0.154**	0.310	**0.330**

TABLE 3.16. Comparison with the recent specialized models on IU-Xray. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialized models							
Transformer Vaswani et al. [2017]	0.459	0.298	0.215	0.162	0.188	0.362	0.511
R2Gen Chen et al. [2020g]	0.470	0.304	0.219	0.165	0.187	0.371	0.430
R2GenCMN Chen et al. [2022b]	0.475	0.309	0.222	0.170	0.191	0.375	0.641
MSAT Wang et al. [2022c]	0.481	0.316	0.226	0.171	0.190	0.372	0.394
DCL Li et al. [2023b]	-	-	-	0.163	0.193	**0.383**	0.586
METransformer Wang et al. [2023]	**0.483**	**0.322**	**0.228**	0.172	0.192	0.380	0.435
S4M (ours)	0.441	0.285	0.208	0.163	0.184	0.361	0.609
generalist models							
R2Gen† (bs=16)	0.306	0.175	0.117	0.084	0.134	0.316	0.289
R2Gen† (bs=96)	0.433	0.275	0.196	0.147	0.184	0.355	0.470
R2Gen† (bs=192)	0.349	0.217	0.153	0.114	0.154	0.332	0.359
S4M (ours, bs=16)	0.444	0.287	0.202	0.152	0.190	0.365	0.509
S4M (ours, bs=96)	0.454	0.290	0.210	0.161	0.187	0.361	**0.700**
S4M (ours, bs=192)	0.466	0.306	0.225	**0.177**	**0.199**	0.367	0.602

TABLE 3.17. Comparison with the recent specialized models on Abdomen. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialized models							
Transformer Vaswani et al. [2017]	0.409	0.247	0.161	0.108	0.142	0.314	0.261
R2Gen Chen et al. [2020g]	0.389	0.241	0.156	0.105	0.143	0.309	0.248
R2GenCMN Chen et al. [2022b]	0.361	0.231	0.151	0.102	0.135	0.310	0.161
MSAT Wang et al. [2022c]	0.410	0.246	0.157	0.105	0.140	0.286	0.275
S4M (ours)	0.373	0.228	0.154	0.106	0.137	0.314	0.196
generalist models							
R2Gen† (bs=16)	0.386	0.238	0.154	0.104	0.144	0.297	0.280
R2Gen† (bs=96)	0.407	0.244	0.150	0.097	0.155	0.297	0.271
R2Gen† (bs=192)	0.397	0.240	0.151	0.100	0.153	0.296	0.271
S4M (ours, bs=16)	0.395	0.243	0.159	0.108	0.152	0.305	0.276
S4M (ours, bs=96)	0.409	0.252	0.162	0.110	0.159	0.313	0.292
S4M (ours, bs=192)	**0.432**	**0.269**	**0.175**	**0.118**	**0.161**	**0.322**	**0.327**

TABLE 3.18. Comparison with the recent specialized models on Knee. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialized models							
Transformer Vaswani et al. [2017]	0.304	0.177	0.116	0.078	0.115	0.288	0.169
R2Gen Chen et al. [2020g]	0.308	0.191	0.121	0.077	0.130	0.300	0.193
R2GenCMN Chen et al. [2022b]	0.329	0.201	0.130	0.083	0.120	0.284	0.164
MSAT Wang et al. [2022c]	**0.366**	0.203	0.128	0.082	0.134	0.282	0.135
S4M (ours)	0.339	0.207	0.133	0.087	0.135	0.295	0.175
generalist models							
R2Gen† (bs=16)	0.321	0.170	0.100	0.064	0.119	0.255	0.154
R2Gen† (bs=96)	0.343	0.197	0.120	0.075	0.134	0.284	0.181
R2Gen† (bs=192)	0.333	0.207	0.134	0.089	**0.139**	**0.308**	0.204
S4M (ours, bs=16)	0.315	0.180	0.111	0.071	0.124	0.276	0.166
S4M (ours, bs=96)	0.331	0.193	0.120	0.077	0.130	0.277	0.188
S4M (ours, bs=192)	0.359	**0.219**	**0.141**	**0.093**	**0.139**	0.291	**0.242**

TABLE 3.19. Comparison with the recent specialized models on Hip. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialized models							
Transformer Vaswani et al. [2017]	0.334	0.193	0.118	0.077	0.116	0.264	0.137
R2Gen Chen et al. [2020g]	0.358	0.211	0.131	0.082	0.131	0.288	0.210
R2GenCMN Chen et al. [2022b]	0.362	0.214	0.133	0.083	0.133	0.286	0.220
MSAT Wang et al. [2022c]	0.362	**0.218**	0.131	0.081	0.125	0.282	0.235
S4M (ours)	0.356	0.216	**0.135**	**0.086**	0.138	**0.294**	0.192
generalist models							
R2Gen† (bs=16)	0.351	0.199	0.120	0.074	0.132	0.275	0.203
R2Gen† (bs=96)	0.361	0.209	0.126	0.080	0.137	0.281	0.226
R2Gen† (bs=192)	**0.367**	0.214	0.133	**0.086**	**0.139**	0.285	0.238
S4M (ours, bs=16)	0.332	0.187	0.113	0.073	0.129	0.263	0.184
S4M (ours, bs=96)	0.366	0.211	0.130	0.084	0.137	0.281	**0.257**
S4M (ours, bs=192)	**0.367**	0.206	0.122	0.076	0.133	0.277	0.215

TABLE 3.20. Comparison with the recent specialized models on Wrist. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialized models							
Transformer Vaswani et al. [2017]	0.339	0.203	0.133	0.086	0.120	0.301	0.129
R2Gen Chen et al. [2020g]	0.359	0.214	0.139	0.093	0.135	0.299	0.288
R2GenCMN Chen et al. [2022b]	0.351	0.210	0.134	0.087	0.129	0.290	0.212
MSAT Wang et al. [2022c]	0.374	0.216	0.134	0.081	0.124	0.295	0.180
S4M (ours)	0.358	0.214	0.137	0.089	0.142	0.302	0.243
generalist models							
R2Gen† (bs=16)	0.351	0.207	0.133	0.085	0.136	0.293	0.217
R2Gen† (bs=96)	0.375	0.215	0.133	0.084	0.144	0.291	0.258
R2Gen† (bs=192)	0.389	**0.238**	**0.154**	**0.102**	0.148	**0.312**	0.296
S4M (ours, bs=16)	0.342	0.199	0.124	0.079	0.133	0.280	0.229
S4M (ours, bs=96)	0.368	0.217	0.138	0.090	0.144	0.298	0.255
S4M (ours, bs=192)	**0.390**	0.232	0.148	0.097	**0.149**	0.299	**0.305**

TABLE 3.21. Comparison with the recent specialized models on Shoulder. † means we optimize the model on our merged training dataset while the "bs" is the training batch size. All evaluations are conducted on the test set, and a higher value indicates better performance.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
specialized models							
Transformer Vaswani et al. [2017]	0.363	0.219	0.138	0.088	0.123	0.301	0.192
R2Gen Chen et al. [2020g]	0.358	0.213	0.130	0.082	0.122	0.307	0.174
R2GenCMN Chen et al. [2022b]	0.348	0.210	0.129	0.082	0.119	0.297	0.134
MSAT Wang et al. [2022c]	0.364	0.221	0.131	0.080	0.123	0.297	0.173
S4M (ours)	0.353	0.211	0.133	0.088	0.129	0.304	0.197
generalist models							
R2Gen† (bs=16)	0.355	0.212	0.131	0.082	0.132	0.299	0.186
R2Gen† (bs=96)	0.374	0.225	0.142	0.095	0.142	0.297	0.274
R2Gen† (bs=192)	0.380	0.231	0.145	**0.096**	**0.144**	0.299	0.277
S4M (ours, bs=16)	0.350	0.207	0.128	0.084	0.133	0.288	0.220
S4M (ours, bs=96)	0.369	0.225	0.145	0.099	0.139	**0.304**	0.272
S4M (ours, bs=192)	**0.389**	**0.234**	**0.146**	**0.096**	0.141	0.302	**0.287**

Statement of Authorship

Title of Paper	V2C: Visual Voice Cloning
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Chen, Q., Tan, M., Qi, Y., Zhou, J., Li, Y. and Wu, Q., 2022. V2C: visual voice cloning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21242-21251).

Principal Author

Name of Principal Author (Candidate)	Qi Chen		
Contribution to the Paper	Proposed the ideas, conducted experiments, and wrote the manuscript.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a t ion in this thesis. I am the primary author of this paper.		
Signature		Date	03/12/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Mingkui Tan		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Yuankai Qi		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Jiaqiu Zhou

Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Yuanqing Li		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Qi Wu		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Please cut and paste additional co-author panels here as required.

Chapter 4

Multi-modal Synergy in Visual Voice Cloning

4.1 Introduction

Voice Cloning (VC) [Arik et al., 2018; Chen et al., 2019b; Jia et al., 2018; Nachmani et al., 2018] aims to convert a paragraph of text to speech with the desired voice from a reference audio. However, there exist many applications in the real world that require the generated speeches not only using a template voice but also with rich emotions (*e.g.*, angry, happy, and sad), such as movie dubbing. This is beyond the scope of conventional VC tasks (Figure 4.1(a)), as no extra guiding information can be used to generate desired tones and rhythms. Considering that we humans accomplish the movie dubbing task with the most reference from visual observations (*e.g.*, watching the movie to grasp the emotion of characters), we propose an extension task of VC, namely Visual Voice Cloning (V2C).

An example of the proposed V2C task is shown in Figure 4.1(b). Unlike the conventional VC task, which converts text to speech only aided by reference audio, our V2C task takes a triplet (text/subtitle, reference audio, reference video) as input and expects a resulting speech with the same voice but varying emotions derived from the reference video. The text/subtitle is the content that the generated speech needs to cover. The reference audio includes a pre-recorded voice of the target speaker from a different clip. We aim to generate a speech with the voice in the reference audio and the character's visual emotion from the reference video, speaking the content in the given text.

The new task poses several novel challenges. First, the conventional Voice Cloning (VC) methods [Arik et al., 2018; Blaauw et al., 2019; Chen et al., 2019b; Jia et al., 2018; Nachmani et al., 2018] cannot well solve the V2C task as they focus only on how to convert the input text to speech with the voice/tone exhibited in the reference audio, without considering the emotion and context of the new speech. However, in our V2C task (*e.g.*, movie dubbing) the voice emotion is crucial for generating human-like speech. Second, in our V2C task, the voice emotion should be derived from the reference video rather than the reference audio from an irrelevant clip. Taking movie dubbing as an example, it requires humans to grasp the emotions of characters by watching the corresponding movie clips and observing their performances (*e.g.*, facial expressions or actions). Although several improved VC methods [Skerry-Ryan et al., 2018; Wang et al., 2018b] also try to inject voice emotion into their generated speech, they capture both emotion and voice from the reference audio, which cannot satisfy the requirements of the V2C task. In our V2C task, an ideal method should be able to disentangle voice and emotion from the reference audios and the reference videos[1],

[1]Notably, the reference videos refer to the image sequences without audio.

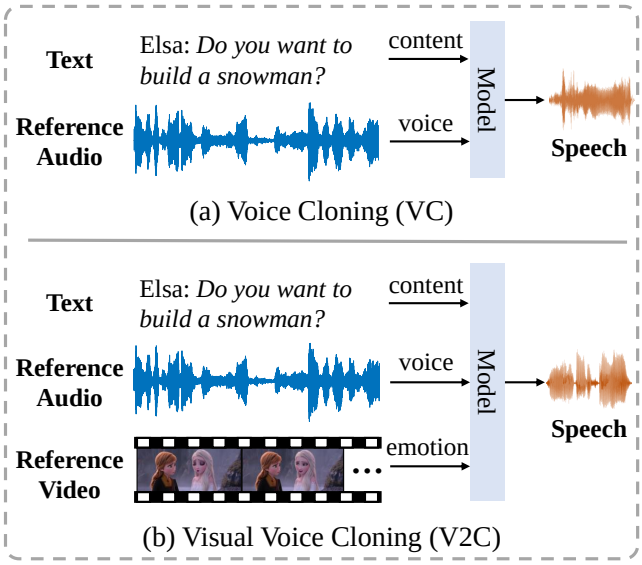


FIGURE 4.1. (a) Voice Cloning (VC) vs. (b) Visual Voice Cloning (V2C). Given an input triplet (*i.e.*, subtitle/text, reference audio, and target video), our V2C task seeks to convert the text into a speech, which should be with the voice of reference audio and the emotion derived from reference video. Note that the reference audio only provides an expected voice while its content is irrelevant.

respectively.

As there is no off-the-shelf dataset suitable for the V2C task, we collect the first V2C-Animation dataset to facilitate the research in this field. It comprises 10,217 video clips with audio and subtitles, covering 26 animated movies with 153 characters (*i.e.*, speakers) in total. Our V2C dataset covers three modalities (*i.e.*, text, audio and video) unlike the existing text-to-speech datasets [Ito and Johnson, 2017; Panayotov et al., 2015; Yamagishi et al., 2019; Zen et al., 2019] or movie description datasets [Rohrbach et al., 2015; Tapaswi et al., 2016], which only focus on text and audio, or text and video. Besides, we also provide emotional annotation (*e.g.*, happy or sad) for each audio and video clip like [Goodfellow et al., 2013]. To alleviate the impact of background music, we only extract the sound channel of the centre speaker, which mainly focuses on the sound of the speaking character. In this way, we ensure that all the audio clips only contain the sound from speaking characters.

To address the above challenges of the V2C task, based on the widely used Text-to-Speech (TTS) framework (*i.e.*, FastSpeech2 [Ren et al., 2021]), we propose a new method called Visual Voice Cloning Network (V2C-Net), considering the emotion information derived from the reference video frames. Moreover, based on MCD [Kubichek, 1993], we design an evaluation metric, called MCD-DTW weighted by Speech Length (MCD-DTW-SL), seeking to evaluate the generated speech effectively and automatically.

In summary, our contribution include:

- We propose a new task, namely Visual Voice Cloning (V2C). Given a triplet (*i.e.*, text/subtitle, reference audio and reference video), the task seeks to convert the text into a speech with voice and emotion derived from reference audio and reference video, respectively.

- We collect the first V2C-Animation dataset, consisting of 26 animated movies, 153 characters, and 10,217 video clips with aligned audios and subtitles, covering



FIGURE 4.2. An example of how to cut a movie into a series of video clips according to SRT subtitle files. Note that the SRT files contain both starting and ending time stamps for each video clip. No. 1340 refers to the sequential number of the current utterance.

three modalities (text, audio, video) and speakers' emotions.

- We design a new method, called Visual Voice Cloning Network (V2C-Net). Besides, to evaluate the generated speech automatically, we provide an advanced automatic evaluation metric, named MCD-DTW-SL.

4.2 Task and Dataset

4.2.1 Problem Definition for V2C Task

Given a triplet $\mathcal{Z} = \{Z_{text}, Z_{audio}, Z_{video}\}$ (*i.e.*, text, reference audio and reference video), our Visual Voice Cloning (V2C) task aims to generate a speech $\mathcal{Y}_{wave}$ (*i.e.*, a waveform in time-domain) from text Z_{text}, which should use the voice of reference audio Z_{audio} and be with the emotion derived from reference video Z_{video}. In Figure 4.1, we take movie dubbing as an example. Given a movie clip (*i.e.*, reference video), a subtitle (*i.e.*, text) and a reference audio, we seek to synthesize a speech from a subtitle according to both the character's emotion derived from a movie and the voice from the reference audio.

4.2.2 Dataset Construction for V2C Task

The dataset for the V2C task should cover all three modalities and the samples from different modalities need to be aligned with each other. As there is no off-the-shelf dataset suitable for this new task, we collect the first V2C dataset, called V2C-Animation.

Data Collection

We search for Blu-ray animated movies with the corresponding subtitles and then select a set of 26 movies of diverse genres. Specifically, we first cut the movies into a series of video clips according to the subtitle files. Here, we use an SRT-type subtitle file. In addition to subtitles/texts, the SRT file also contains starting and ending time-stamps to ensure the subtitles match with video and audio, and the sequential number of subtitles (*e.g.*, No. 1340 in Figure 4.2), which indicates the index of each video clip. Based on the SRT file, we cut the movie into a series of video clips using the FFmpeg toolkit [Tomar, 2006] (an automatic audio and video processing toolkit) and then extract the audio from each video clip by FFmpeg as well. Note that, to alleviate the impact of the background music, we only extract the sound channel of the centre speaker, which mainly focuses on the sound of the speaking character.

Dataset	Text	Audio	Video	Identity	Emotion	#Movies	#Video Clips	#Audio Clips	#Speakers	Avg. S	Avg. A/V (s)
LJ Speech	√	√		√		-	-	13100	1	17.23	6.57
LibriSpeech	√	√		√		-	-	250698	2484	32.55	14.10
VCTK	√	√		√		-	-	44070	108	7.41	3.59
LibriTTS	√	√		√		-	-	375086	2456	16.86	5.62
MPII-MD	√		√			94	68337	-	-	-	3.88
MovieQA	√		√			140	6771	-	-	6.20	202.67
LRS2-main	√	√	√			-	48164	48164	-	7.13	1∼2
LRS2-pretrain	√	√	√			-	96318	96318	-	21.43	∼10
V2C-Animation (ours)	√	√	√	√	√	26	10217	10217	153	6.51	2.40

TABLE 4.1. We compare our V2C-Animation dataset with several existing multi-modal datasets. "Identity" and "Emotion" indicate whether the datasets contain annotations corresponding to the speaker's identity and emotion. The notations "#Movies", "#Video Clips", "#Audio Clips" and "#Speakers" refer to the number of movies, videos, audios and speakers/characters, respectively. "Avg. S" indicates the average length of subtitle while "Avg. A/V" is the average duration of audio/video.

Data Annotation and Organization

Inspired by the organization of LibriSpeech dataset [Panayotov et al., 2015], we categorize the obtained video clips, audios and subtitles into their corresponding characters (*i.e.*, speakers) via a crowd-sourced service. To ensure that the characters appearing in the video clips are the same as the speaking ones, we manually remove the data example that does not satisfy the requirement. Then, following the categories of FER-2013 [Goodfellow et al., 2013] (a dataset for human facial expression recognition), we divide the collected video/audio clips into 8 types including angry, happy, sad, *etc.*. In this way, we totally collect a dataset with 10,217 video clips with paired audio and subtitles. All of the annotations, time-stamps of the mined movie clips and a tool to extract the triplet data will be released. We randomly split 60% samples as training data, 10% samples as validation data and 30% samples as testing data.

4.2.3 V2C-Animation Dataset vs. Related Datasets

We compare our V2C-Animation dataset against ones of the VC, TTS, and PT tasks. In addition, we consider some Movie Description (MD) datasets and Lip Reading Sentence (LRS) datasets, which contain both video and text like ours. Specifically, the VC/TTS/PT datasets include LJ Speech [Ito and Johnson, 2017], LibriSpeech [Panayotov et al., 2015], VCTK [Yamagishi et al., 2019] and LibriTTS [Zen et al., 2019], the MD datasets involve MPII-MD [Rohrbach et al., 2015] and MovieQA [Tapaswi et al., 2016], while LRS datasets contain LRS2 [Chung et al., 2017][2]. From Table 4.1, the statistic results demonstrate that our V2C-Animation dataset is unique, covering all three modalities (*i.e.*, text, audio and video) with both identity and emotion annotations, while most others only focus on the two modalities and all of them are without emotion annotations.

To further compare the differences between our V2C-Animation dataset and related datasets, following [Skerry-Ryan et al., 2018], we visualize the pitch tracks of the samples from our dataset and others. Specifically, we randomly select an audio sample from LJ Speech, LibriSpeech and LibriTTS, respectively. Due to the varying lengths of audio, for a fair comparison, we crop two seconds of audio from each compared sample. As shown in Figure 4.3, the audio pitches from the existing datasets are smoother and the ranges of frequency (Hz) are narrower than ours. Moreover, we provide average and variance values of the pitch tricks. Table 4.2 shows that the variance of our V2C-Animation is the largest, which further demonstrates that our proposed dataset has a wider range of frequency (Hz). Both the visual and statistical results demonstrate that our V2C-Animation dataset is more challenging due to the varied prosody.

4.2.4 More Analysis of V2C-Animation Dataset

Word Cloud and Count

In Figure 4.4, we visualise the texts/subtitles of our V2C-Animation dataset as Venn-style word cloud [Coppersmith and Kelly, 2014], where the size of each word refers to the harmonic mean of its count.

Besides, we also provide the top 30 words on our V2C-Animation dataset along with their counts in Figure 4.5. More results (top 100) are in the following:

[2]The LRS2 dataset has two subsets: LRS2-main and LRS2-pretrain. On LRS2-pretrain, the utterance of each video may contain multiple sentences. But each video only corresponds to a single sentence on LRS2-main. There is some overlap between LRS2-pretrain and LRS2-main sets.

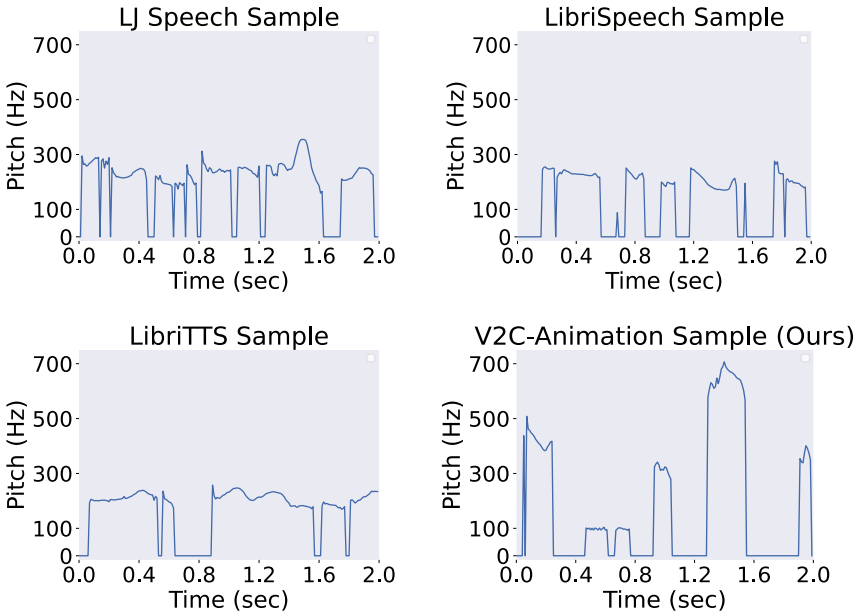


FIGURE 4.3. Examples from the existing Text-to-Speech (TTS) datasets (*i.e.*, LJ speech, LibriTTS and LibriSpeech) and our V2C-Animation dataset. A pitch of 0 Hz refers to an unvoiced segment.

Dataset	LJ Speech	LibriSpeech
Avg. P (Hz)	127.27 ± 11800.96	88.15 ± 7313.39
Dataset	LibriTTS	V2C-Animation (Ours)
Avg. P (Hz)	93.97 ± 9295.67	117.99 ± 16910.77

TABLE 4.2. We compare the average and variance of the pitch from our V2C-Animation dataset and the related datasets. "Avg. P" refers to the average values of pitches with the corresponding variances.

('know', 437), ('oh', 305), ('right', 255), ('one', 254), ('now', 250), ('well', 250), ('go', 233), ('okay', 217), ('come', 210), ('want', 201), ('look', 196), ('got', 181), ('going', 173), ('think', 167), ('will', 165), ('thing', 163), ('gonna', 163), ('need', 159), ('see', 155), ('back', 153), ('never', 151), ('us', 147), ('time', 141), ('say', 139), ('hey', 138), ('mean', 137), ('let', 137), ('good', 135), ('yeah', 131), ('guy', 128), ('really', 124), ('make', 124), ('thank', 124), ('little', 112), ('way', 108), ('love', 108), ('ye', 108), ('find', 104), ('help', 97), ('tell', 96), ('wait', 95), ('take', 93), ('kid', 92), ('please', 91), ('sorry', 88), ('something', 87), ('great', 87), ('dad', 87), ('friend', 84), ('day', 82), ('game', 80), ('stop', 75), ('even', 75), ('Uh', 74), ('big', 67), ('work', 66), ('Ralph', 66), ('much', 62), ('give', 62), ('first', 61), ('everything', 60), ('new', 59), ('still', 58), ('life', 58), ('keep', 58), ('dragon', 58), ('family', 57), ('sure', 56), ('made', 56), ('talk', 55), ('world', 53), ('place', 53), ('heart', 53), ('every', 53), ('maybe', 53), ('stay', 52), ('wanna', 51), ('better', 51), ('people', 50), ('huh', 50), ('anything', 50), ('getting', 49), ('thought', 48), ('man', 48), ('mom', 48), ('listen', 48), ('guess', 47), ('fine', 47), ('around', 47), ('gotta', 46), ('believe', 46), ('two', 45), ('someone', 45), ('home', 45), ('call', 45), ('boy', 45), ('son', 44), ('put', 43), ('fix', 43), ('always', 43)

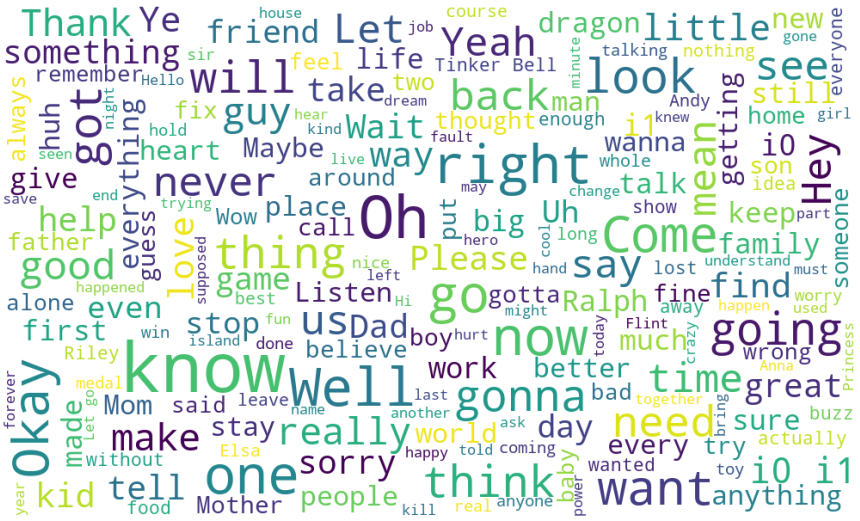


FIGURE 4.4. Word cloud of the texts on our V2C-Animation dataset.

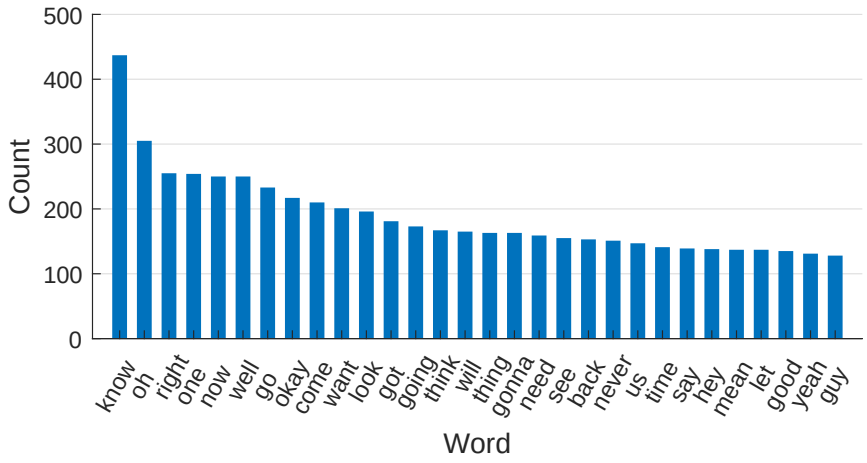


FIGURE 4.5. Top 30 words on V2C-Animation along with the counts.

Distribution of Emotion Labels

Following the categories of FER-2013 [Goodfellow et al., 2013] (a dataset for human facial expression recognition), we divide the collected video/audio clips into 8 types (*i.e.*, 0: angry, 1: disgust, 2: fear, 3: happy, 4: neutral, 5: sad, 6: surprise, and 7: others). The number and distribution of each emotion label can be found in Table 4.3 and Figure 4.6, respectively.

Distribution of Utterance Length

Figure 4.7 exhibits the distribution of utterance/text length on the V2C-Animation dataset, which shows that most utterances range from 3 to 8 words. Besides, we also list the number of utterances/texts and their corresponding percentages in the following (format: *length, count, percentage*):
(1, 594, 5.81%), (2, 708, 6.93%), (3, 914, 8.95%), (4, 1116, 10.92%), (5, 1232, 12.06%), (6, 1213, 11.87%), (7, 1040, 10.18%), (8, 919, 8.99%), (9, 783, 7.66%), (10, 615, 6.02%), (11, 423, 4.14%), (12, 270, 2.64%), (13, 192, 1.88%), (14, 105, 1.03%), (15, 54, 0.53%), (16, 24, 0.23%), (17, 11, 0.11%), (18, 3, 0.03%), (19, 1, 0.01%)

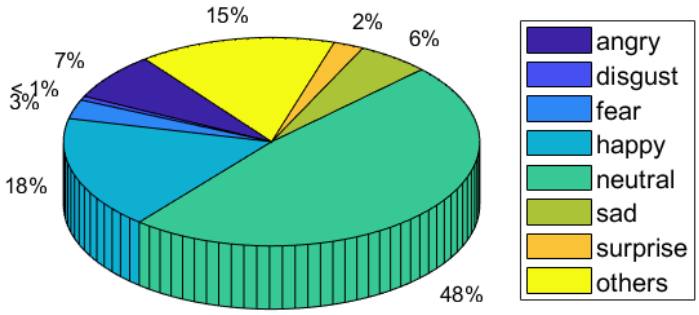


FIGURE 4.6. Distribution of emotion labels on V2C-Animation.

Emotion	angry	disgust	fear	happy
Count	756	64	305	1799
Emotion	neutral	sad	surprise	others
Count	4919	572	240	1562

TABLE 4.3. Counts of the emotion labels on V2V-Animation dataset.

More Examples of Subtitle and Video Clip

We show several examples of how to crop movies based on a corresponding subtitle file. Here, we use an SRT-type subtitle file. Besides the subtitles/texts, the SRT file also contains starting and ending time stamps to ensure the subtitles match with video and audio. The sequential number of subtitles (*e.g.*, No. 726 and No. 1340 in Figure 4.8) indicates the index of each video clip. Based on the SRT file, we cut the movie into a series of video clips via FFmpeg toolkit [Tomar, 2006] (an automatic audio and video processing toolkit).

Samples of Character's Emotion

Figure 4.9 shows some samples of the reference videos on the V2C-Animation dataset with their corresponding emotions.

List of Animated Movies and Characters

As shown in Figure 4.4, we report all the names of our collected animated movies with their corresponding characters/speakers on the V2C-Animation dataset.

4.3 Method

For the V2C task, we propose a baseline model, called Visual Voice Cloning Network (V2C-Net), which is based on a widely used TTS framework FastSpeech2 [Ren et al., 2021]. As shown in Figure 4.10, our model contains three main components: a multi-modal encoder, a synthesizer and a vocoder. We input a triplet (*i.e.*, text, reference audio, and reference video) into the encoder and output three types of features (*i.e.*, phoneme, speaker, and emotion). Based on these features, we use the synthesizer to generate a mel-spectrogram (see Figure 4.10 right side), which is a time-frequency representation of the audio signal. Last, we convert the generated mel-spectrogram into waveform (*i.e.*, speech, see Figure 4.10 right-bottom) by the vocoder.

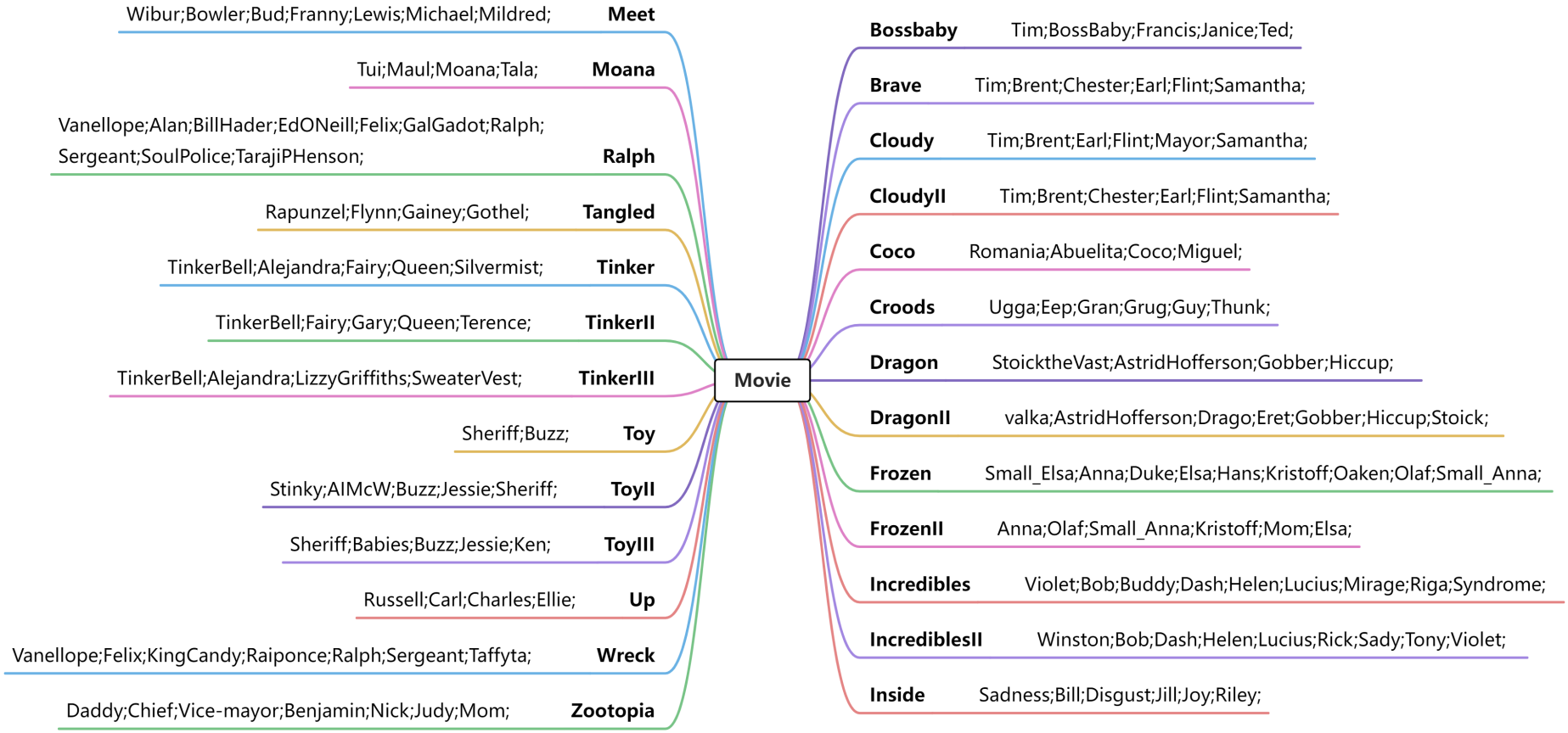


TABLE 4.4. Movies with the corresponding speakers/characters on the V2C-Animation dataset.

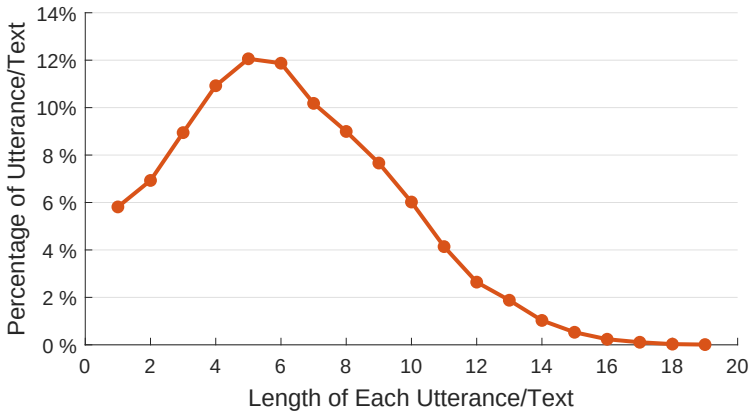


FIGURE 4.7. Distribution of utterance/text length.



FIGURE 4.8. Examples of how to cut a movie into a series of video clips according to subtitle files. Note that the subtitle files contain both starting and ending time stamps for each video clip.

4.3.1 Multi-modal Encoder for Feature Extraction

Given a triplet $\mathcal{Z} = \{Z_{text}, Z_{audio}, Z_{video}\}$, the output feature/embedding $\mathcal{X}$ from the multi-modal encoder is

$$\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_L\} = f(\mathcal{Z}), \tag{4.1}$$

where L is the length of input sentence (*i.e.*, the number of phoneme). $f(\cdot)$ is the multi-modal encoder, mainly containing three sub-modules: a text encoder f_{txt}, a speaker encoder f_{spk}, and an emotion encoder f_{emo}. Here, Z_{text}, Z_{audio} and Z_{video} indicate the input text, reference audio and reference video, respectively. We obtain the i-th output feature $\mathbf{x}_i = \mathbf{o}_i \oplus \mathbf{u} \oplus \mathbf{v}$, where $\mathbf{o}_i$ is the embedding of the i-th phoneme derived from text Z_{text}. The embeddings $\mathbf{u}$ and $\mathbf{v}$ are the outputs of f_{spk} and f_{emo}, respectively. The notation $\oplus$ indicates the operation of element-wise add.

Text Encoder

Following the structure of FastSpeech2, we take 4 Feed-Forward Transformer (FFT) blocks [Ren et al., 2019b] as our text encoder. Based on such a text encoder, we produce a series of phoneme embeddings $\mathcal{O} = \{\mathbf{o}_1, ..., \mathbf{o}_L\}$ from an input text Z_{text}. Mathematically, the process can be defined as $\mathcal{O} = f_{txt}(Z_{text})$.

Speaker Encoder

To explore the voice characteristics of different speakers, we adopt a speaker encoder f_{spk}, which has the same architecture as [Wan et al., 2018], comprising 3 LSTM layers and a linear layer. The speaker encoder first converts a sequence of mel-spectrogram frames, derived from the reference audio, to a series of hidden embeddings by LSTM,



FIGURE 4.9. Samples of the character's emotion (*e.g.*, happy and sad) involved in the reference video. Here, we take *Elsa* (a character in the movie *Frozen*) as an example.

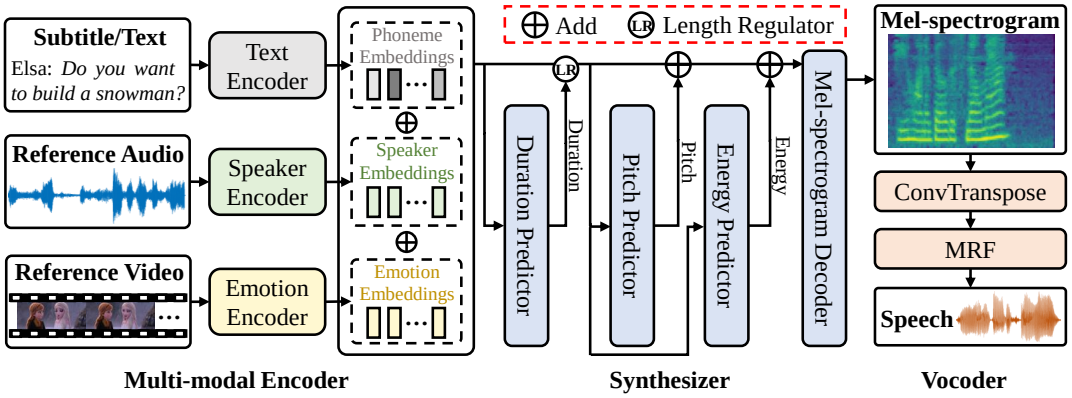


FIGURE 4.10. Overview of V2C-Net. It consists of three main components: a multi-modal encoder, a synthesizer and a vocoder. A triplet (*i.e.*, text, reference audio and reference video) is fed into the multi-modal encoder (Sec. 4.3.1) and it outputs three types of embeddings. Based on these embeddings, the synthesizer (Sec. 4.3.2) generates a mel-spectrogram. Finally, the mel-spectrogram is converted into the waveform (*i.e.*, speech) by the vocoder (Sec. 4.3.3).

and then maps the last hidden embedding to a fixed-dimensional vector via the linear layer. For convenience, we define the process as $\mathbf{u} = f_{spk}(\sigma(Z_{audio}))$, where $\mathbf{u}$ refers to the speaker embedding while σ is a mapping function, converting the audio from waveform to mel-spectrogram.

Emotion Encoder

To exploit the emotion from the video, we design an emotion encoder E_E, which captures the embedding of the whole video clip Z_{video}. To be specific, we use the I3D model [Carreira and Zisserman, 2017] as our emotion encoder f_{emo} and calculate the emotion embedding $\mathbf{v}$ by $\mathbf{v} = f_{emo}(Z_{video})$.

4.3.2 Synthesizer for Mel-Spectrogram Generation

To generate the mel-spectrogram from the conditional phoneme embedding $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_L\}$, we introduce a synthesizer inspired by FastSpeech2 [Ren et al., 2021], and obtain the predicted mel-spectrogram frames $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_T\}$, where T is the number of mel-spectrogram frames. Here, the synthesizer mainly contains four parts: a duration predictor, a pitch predictor, an energy predictor, and a mel-spectrogram

decoder. The loss function of the synthesizer is

$$\mathcal{L}_S = \mathcal{L}_{mel} + \lambda_1 \mathcal{L}_{dur} + \lambda_2 \mathcal{L}_{pitch} + \lambda_3 \mathcal{L}_{energy}, \tag{4.2}$$

where $\mathcal{L}_{mel}$, $\mathcal{L}_{dur}$, $\mathcal{L}_{pitch}$ and $\mathcal{L}_{energy}$ refer to the losses of mel-spectrogram, duration predictor, pitch predictor and energy predictor, respectively. The λ_1, λ_2 and λ_3 are hyper-parameters, and we set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in practice. The details are depicted in the following.

Duration Predictor

To alleviate the problem of length mismatch between the input embedding $\mathcal{X}$ and mel-spectrogram frames $\mathcal{Y}$ (*i.e.*, $L \neq T$), we introduce a duration predictor S_d, which takes the embedding $\mathcal{X}$ as input and predicts the duration $\mathcal{D} = \{d_1, ..., d_i..., d_L\} = \{S_d(\mathbf{x}_1), ..., S_d(\mathbf{x}_i)..., S_d(\mathbf{x}_L)\}$ of each phoneme embedding. The i-th phoneme duration d_i indicates the number of copies for i-th phoneme embedding $\mathbf{x}_i$. Then, we use a Length Regulator ($\mathcal{LR}$):

$$\mathcal{X}_{mel} = \mathcal{LR}(\mathcal{X}, \mathcal{D}) = \mathcal{LR}(\mathcal{X}, S_d(\mathcal{X})), \tag{4.3}$$

where $\mathcal{X}_{mel}$ is extended phoneme embedding[3] with length T. To optimize the duration predictor, we use Montreal forced alignment (MFA) [McAuliffe et al., 2017] tools to obtain the ground-truth phoneme duration sequence, and then calculate a mean square error (MSE) loss between ground-truth $\hat{\mathcal{D}}$ and predicted $\mathcal{D}$. Formally, the loss can be defined as

$$\mathcal{L}_{dur} = \frac{1}{L} \sum_{i=1}^{L} (\hat{d}_i - d_i)^2. \tag{4.4}$$

Pitch and Energy Predictors

To affect the prosody and volume of speech, following [Ren et al., 2021], we employ a pitch predictor S_p and an energy predictor S_e, respectively. Specifically, to predict the pitch contour, we use continuous wavelet transform (CWT) to convert the continuous pitch into pitch spectrogram [Hirose and Tao, 2015; Suni et al., 2013], and take it as ground-truth to optimize the pitch predictor S_p by MSE loss:

$$\mathcal{L}_{pitch} = \frac{1}{T} \sum_{t=1}^{T} (\hat{p}_t - p_t)^2, \tag{4.5}$$

where $\hat{p}_t$ and $p_t = S_p(\mathbf{x}_t)$ denote the t-th ground-truth and predicted pitch value, respectively. For energy, we follow the operation in [Ren et al., 2021] that calculates an L2-norm of the amplitude of each short-time Fourier transform (STFT) frame and takes it as energy. The corresponding loss function is

$$\mathcal{L}_{energy} = \frac{1}{T} \sum_{t=1}^{T} (\hat{e}_t - e_t)^2, \tag{4.6}$$

where $\hat{e}_t$ and $e_t = S_e(\mathbf{x}_t)$ are the t-th ground-truth and predicted energy value, respectively. Last, we encode each pitch and energy value into the corresponding embedding by the embedding layers ϕ and φ separately, and then add the pitch and energy embeddings into the extended phoneme embedding $\mathcal{X}_{mel}$. Mathematically, the

[3] For example, if $d_i = 2$, $\mathcal{X}_{mel}$ would be $\mathcal{X}_{mel} = \{..., \mathbf{x}_i, \mathbf{x}_i, ...\}$. For simplicity, we redefine $\mathcal{X}_{mel}$ as $\mathcal{X}_{mel} = \{\mathbf{x}_1, ..., \mathbf{x}_t, ...\mathbf{x}_T\}$

Methods	MCD ↓	MCD-DTW ↓	MCD-DTW-SL ↓	Id. Acc. ↑	Emo. Acc. ↑	MOS-naturalness ↑	MOS-similarity ↑
Ground Truth	00.00	00.00	00.00	90.62	84.38	4.61 ± 0.15	4.74 ± 0.12
SV2TTS	21.08	12.87	49.56	33.62	37.19	2.03 ± 0.22	1.92 ± 0.15
SV2TTS*	17.41	11.16	15.92	38.21	41.24	3.20 ± 0.20	3.09 ± 0.33
FastSpeech2	12.08	10.29	10.31	59.38	53.13	3.86 ± 0.07	3.75 ± 0.06
V2C-Net (ours)	**11.79**	**10.09**	**10.05**	**62.50**	**56.25**	**3.97 ± 0.06**	**3.90 ± 0.06**

TABLE 4.5. Comparison with the state-of-the-art methods. We provide the results of both objective (*i.e.*, MCD, MCD-DTW and MCD-DTW-SL) and subjective evaluation metrics (*i.e.*, MOS-naturalness and MOS-similarity). "Id. Acc." and "Emo. Acc." are the identity and emotion accuracy of the generated speech, respectively. The method with "*" refers to a variant taking video (emotion) embedding as an additional input. "Ground Truth" denotes the results on ground-truth samples. ↑ (↓) means that the higher (lower) value is better.

mel-spectrograms $\mathcal{Y}$ can be generated by

$$\mathcal{Y} = g\left(\mathcal{X}_{mel} \oplus \phi(S_p(\mathcal{X}_{mel})) \oplus \varphi(S_e(\mathcal{X}_{mel}))\right), \tag{4.7}$$

where $g(\cdot)$ refers to a mel-spectrogram decoder, consisting of 6 FFT blocks [Ren et al., 2019b]. To optimise the predicted mel-spectrogram, we use the loss function:

$$\mathcal{L}_{mel} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{\mathbf{y}}_t - \mathbf{y}_t\|, \tag{4.8}$$

where $\hat{\mathbf{y}}_t$ denotes the t-th frame of the ground-truth mel-spectrogram while $\mathbf{y}_t \in \mathcal{Y}$ is the predicted one.

4.3.3 Vocoder for Speech Synthesis

In Figure 4.10, to convert the generated mel-spectrogram $\mathcal{Y}$ into time-domain waveform $\mathcal{Y}_{wave}$, we use HiFi-GAN [Kong et al., 2020] as our vocoder, which mainly focuses on the raw waveform generation from mel-spectrogram by GANs [Goodfellow et al., 2014]. The generator of HiFi-GAN can be divided into two major modules: a transposed convolution (ConvTranspose) network and a multi-receptive field fusion (MRF) module. Specifically, we first upsample the mel-spectrogram $\mathcal{Y}$ by ConvTranspose, which seeks to take an alignment between the length of the output features and the temporal resolution of raw waveforms. Then, we feed the upsampled features into the MRF module, which consists of multiple residual blocks [He et al., 2016], and take the sum of outputs from these blocks as our predicted waveform. Here, we follow the settings of [Kong et al., 2020] that use the residual blocks with different kernel sizes and dilation rates to ensure different receptive fields. We optimize the vocoder via the objective function that contains an LSGAN-based loss [Mao et al., 2017], a mel-spectrogram loss [Isola et al., 2017], and a feature matching loss [Kumar et al., 2019]. In practice, we use the "universal" checkpoint provided in the official code of HiFi-GAN without fine-tuning on any of the synthesised mel-spectrograms.

4.4 Experiments

We evaluate the quality of generated speech in terms of three aspects: 1) objective evaluation, 2) subjective evaluation, and 3) identity and emotion accuracy. The objective and subjective evaluation metrics aim to assess the quality of generated speeches

by comparing them with ground-truth ones. By contrast, identity accuracy and emotion accuracy focus on whether the generated speeches involve the desired voice (*i.e.*, identity) and emotion, respectively. We provide both quantitative and qualitative results on the V2C-Animation dataset. More details are in the following.

4.4.1 Evaluation Metrics

Objective Evaluation Metric

To assess the quality of generated speech, we use Mel Cepstral Distortion (MCD) [Kubichek, 1993] metric, which compares Mel Frequency Cepstral Coefficient (MFCC) vectors $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_i, ..., \mathbf{c}_M\}$ and $\mathcal{C}' = \{\mathbf{c}'_1, \mathbf{c}'_2, ..., \mathbf{c}'_j, ..., \mathbf{c}'_N\}$ derived from the generated speech and ground truth, respectively. We sum the Euclidean distance over the first $K(K = 13)$ MFCC values:

$$\text{MCD}(\mathcal{C}, \mathcal{C}') = \frac{1}{T}\sum_{t=1}^{T} d(\mathbf{c}_t, \mathbf{c}'_t) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{k=1}^{K}(c_{t,k} - c'_{t,k})^2}, \tag{4.9}$$

where $M = N = T$ refers to the number of speech/audio frames. The $c_{t,k}$ and $c'_{t,k}$ denote the k-th MFCC value of the t-th speech frame from generated and ground-truth speeches, respectively, while $\mathbf{c}_t = (c_{t,1}, c_{t,2}..., c_{t,K})$ and $\mathbf{c}'_t = (c'_{t,1}, c'_{t,2}..., c'_{t,K})$.

Note that the MCD metric requires the lengths of two input speeches to be the same (*i.e.*, $M = N$). When $M \neq N$, the existing voice cloning methods like [Skerry-Ryan et al., 2018] simply extend the shorter speech to the length of longer one by padding 0 for the time-domain waveform. In this way, the value of MCD may become extremely large if the mismatching occurs at the beginning of two speeches. To avoid this issue, Battenberg *et al*[Battenberg et al., 2020] use an improved MCD metric, called MCD-DTW, which adopts the Dynamic Time Warping (DTW) [Müller, 2007] algorithm to find the minimum MCD between two speeches. However, MCD-DTW would achieve a better value as long as there is a match between the two speeches. This is not reasonable as a better-generated speech should have a similar length to the ground truth.

To alleviate the above issues, we propose an MCD-DTW weighted by Speech Length (MCD-DTW-SL), which evaluates both the length and the quality of alignment between two speeches. In MCD-DTW-SL, to evaluate whether the two speeches (*i.e.*, $\mathcal{C}$ and $\mathcal{C}'$) are aligned, we still use the DTW algorithm to calculate the minimum distance among them. Specifically, we compute the cumulative distance $\gamma_{i,j} = d(\mathbf{c}_i, \mathbf{c}'_j) + \min(\gamma_{i-1,j-1}, \gamma_{i-1,j}, \gamma_{i,j-1})$, where $\gamma_{i,j}$ is the minimum cumulative distance from index $(1, 1)$ to (i, j). Then, we obtain the objective minimum distance $\gamma_{M,N}$ by accumulating R distances in total, where $\max(M, N) \leq R < M + N - 1$. Besides, considering the influence of lengths, we design a simple but effective coefficient $\eta = \frac{\max(M,N)}{\min(M,N)}$. Formally, we calculate the metric

$$\text{MCD-DTW-SL}(\mathcal{C}, \mathcal{C}') = \frac{\eta}{R} \cdot \gamma_{M,N}. \tag{4.10}$$

Subjective Evaluation Metric

To further evaluate the quality of generated speech, we conduct a human study by using a subjective evaluation metric. Specifically, following the settings in [Jia et al., 2018], we use a Mean Opinion Score (MOS) evaluation approach based on subjective listening tests. In this approach, we use the Absolute Category Rating (ACR) scale [Rec, 1996] with rating scores from 1 to 5 (*i.e.*, from "Bad" to "Excellent") in

0.5 point increments. Based on such an approach, we mainly evaluate the generated speeches with respect to naturalness and similarity. 1) *MOS-naturalness*: to assess the naturalness of the generated speech, we randomly sample 100 generated audios from the testing set and divide them into 4 groups. Each group is rated by a single rater. 2) *MOS-similarity*: to evaluate whether the generated speech is well aligned with the desired voice and prosody, we compare each generated speech with the ground-truth one from the same speaker. We use the same samples as when evaluating MOS-naturalness above. Each pair is rated by the rater according to the similarity between the two speeches.

Identity and Emotion Accuracy

To evaluate whether the generated speech carries proper speaker identity and emotion, we propose identity accuracy and emotion accuracy, respectively. The identity accuracy aims to verify whether the generated speech can be recognized as the same speaker as the input reference audio. Similarly, the emotion accuracy reflects whether the generated speech contains the same emotion as the reference video. To this end, we first use GE2E [Wan et al., 2018] model as our audio encoder to obtain a fixed-dimensional audio embedding for each speech. Then, based on the audio embeddings $\{\mathbf{a}_{i1}, ..., \mathbf{a}_{iH}\}$[4] of the i-th speaker, we obtain the centroid $\mathbf{a}'_i$ of the i-th speaker by $\mathbf{a}'_i = \mathbb{E}_d[\mathbf{a}_{ih}] = \frac{1}{H}\sum_{d=1}^{H}\mathbf{a}_{ih}$, where H refers to the number of audio belonging to the i-th speaker. Last, we compare the cosine similarity between the embedding of generated speech and each centroid, and then classify it into the category the most similar centroid belongs. Similarly, we can calculate the emotion accuracy in the same way as well.

4.4.2 Implementation Details

For the speaker encoder f_{spk}, we use the same architecture as [Wan et al., 2018], comprising three LSTM layers. The audio encoder maps a sequence of mel-spectrogram frames, derived from the reference audio, to a vector with a fixed dimension of 256. We optimize the model with a generalized end-to-end speaker verification loss, which ensures features from the same speaker are more similar than ones from different speakers. For the emotion encoder f_{emo}, we use a conventional I3D model [Carreira and Zisserman, 2017], trained on our V2C-Animation dataset [Carreira and Zisserman, 2017] with 64×10^3 iterations and final output a vector with 1024 dimensions. For our synthesizer, we train the text encoder f_{txt} and the synthesizer in an end-to-end manner with 16 batch size and 2×10^6 iterations on our proposed V2C-Animation dataset. We train all models on a single GPU device (GeForce RTX 3090).

Details of Vocoder

To synthesise the waveform of the speech from our generated mel-spectrogram, we use HiFi-GAN [Kong et al., 2020] as our vocoder. The HiFi-GAN model is based on Generative Adversarial Networks (GANs) [Goodfellow et al., 2014], which consists of one generator and two discriminators, *i.e.*, a multi-period discriminator (MPD) and a multi-scale discriminator (MSD).

The generator of HiFi-GAN can be divided into two major modules: a transposed convolution (ConvTranspose) network and a multi-receptive field fusion (MRF) module. Specifically, we first upsample the mel-spectrogram by ConvTranspose, which

[4]The embeddings have been normalized using L2 norm.

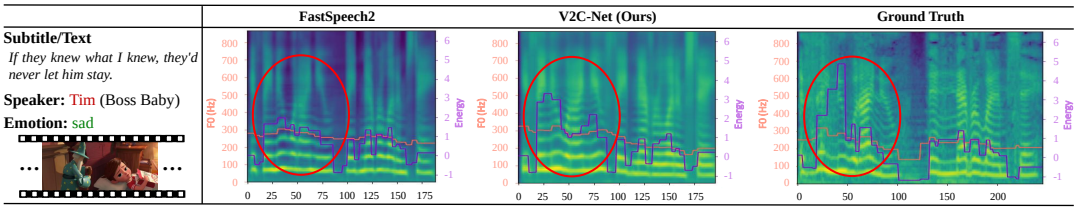


FIGURE 4.11. Mel-spectrogram of generated and ground-truth audios. Orange curves are F_0 contours, where F_0 is the fundamental frequency of audio. Purple curves refer to the energy (volume) of audio. The horizontal axis is the duration of the audio. We highlight the main difference via a red circle.

seeks to take an alignment between the length of the output features and the temporal resolution of raw waveforms. Then, we feed the upsampled features into the MRF module, which consists of multiple residual blocks [He et al., 2016], and take the sum of outputs from these blocks as our predicted waveform. Here, the residual blocks with different kernel sizes and dilation rates are used to ensure different receptive fields.

For the two discriminators, the multi-period discriminator (MPD) contains several sub-discriminators, where each sub-discriminator handles a specific periodic part of the input audio. By contrast, the multi-scale discriminator (MSD) proposed in MelGAN [Kumar et al., 2019], consisting of three sub-discriminators, tries to capture the consecutive patterns and long-term dependencies from the input audio.

The generator and discriminators are trained adversarially, aiming to improve the training stability and the model performance. Specifically, the vocoder (*i.e.*, HiFi-GAN) is optimized via the objective function that contains an LSGAN-based loss [Mao et al., 2017], a mel-spectrogram loss [Isola et al., 2017], and a feature matching loss [Kumar et al., 2019]. In practice, we use the vocoder (*i.e.*, HiFi-GAN) pretrained on the LibriSpeech dataset [Panayotov et al., 2015].

4.4.3 Quantitative Evaluation

To evaluate the performance of our method, we compare V2C-Net with several state-of-the-art methods. In Table 4.5, our V2C-Net consistently outperforms the existing VC models (*e.g.*, the MCD-DTW-SL of our V2C-Net is 10.05 while the SV2TTS model only achieves 49.56). Besides, we propose a variant of the SV2TTS model, called SV2TTS*, which takes the video embedding derived from our method as an additional input. From Table 4.5, SV2TTS* achieves better performance compared with SV2TTS, which further demonstrates the effectiveness of our video component. Note that the Id. Acc. and Emo. Acc. on the ground truth are not 100%. This is because our pre-trained identity and emotion classification models are not perfect.

4.4.4 Qualitative Evaluation

To further assess the quality of generated speech, we show the visualized results of the proposed methods, baseline method and ground-truth, respectively. In Figure 4.11, compared with FastSpeech2, both the energy (volume) curve and the fundamental frequency curve (*i.e.*, F_0 curve) of the mel-spectrogram generated by our V2C-Net are more similar to the ground-truth (GT) ones. Notably, as the duration of audio should be predicted as well, the lengths of the generated audio and the GT one may be different.

	ref. A	ref. V	Id. Acc. ↑	Emo. Acc. ↑
V2C-Net		√	25.00	47.61
	√		59.38	53.13
	√	√	**62.50**	**56.25**

TABLE 4.6. Effect of reference audio and video. "ref. A" and "ref. V" denote the reference audio and reference video separately. "Id. Acc." and "Emo. Acc." are the identity and emotion accuracy of the generated speech, respectively. ↑ means a higher value is better.

Task	Dataset	MCD ↓	MOS-naturalness ↑
VC	VCTK [Yamagishi et al., 2019]	-	4.07 ± 0.06
	LibriSpeech [Panayotov et al., 2015]	-	3.98 ± 0.06
	Multi-speaker [Skerry-Ryan et al., 2018]	12.37	-
V2C	V2C-Animation (Ours)	21.08	2.03 ± 0.22

TABLE 4.7. Comparisons on the difficulties of the conventional Voice Cloning (VC) and Visual Voice Cloning (V2C) tasks. We show the performance of SV2TTS [Jia et al., 2018] trained on different datasets. VCTK, LibriSpeech and Multi-speaker are widely used in VC task. ↑ (↓) means higher (lower) value refers to better performance.

4.4.5 More Discussions

Effect of Reference Audio and Video

To investigate the effect of reference audio and video, we compare the generated speeches by removing them alternately and show the quantitative results (*i.e.*, identity and emotion accuracy) in Table 4.6. The results show that with the control of reference audio, our V2C-Net achieves higher identity accuracy obviously than the counterpart without reference audio (*i.e.*, from 25.00% to 59.38%). After further incorporating the information from the reference video, the model obtains the best performance on both two metrics.

Comparing Difficulties of V2C and VC Tasks

To investigate whether our V2C task is more challenging than the conventional VC task, we compare results of the SV2TTS method [Jia et al., 2018] on both our V2C-Animation dataset and the existing VC datasets (*e.g.*, VCTK [Yamagishi et al., 2019], LibriSpeech [Panayotov et al., 2015] and Multi-speaker [Skerry-Ryan et al., 2018]). In Table 4.7, SV2TTS obtains 4.07±0.06 and 3.89±0.06 MOS-naturalness on VCTK and LibriSpeech, respectively, which are higher than that on our V2C-Animation dataset (2.03 ± 0.22 in Table 4.7). Besides, the SV2TTS model achieves 12.37 MCD value on Multi-speaker dataset [Skerry-Ryan et al., 2018], which is better than the results of the same model on our V2C-Animation dataset (*i.e.*, 21.08 in Table 4.7). It demonstrates that the proposed V2C task is more non-trivial as the same VC model obtains worse results on our V2C-Animation dataset than others.

More Visual Results of Mel-spectrogram

We provide more visualized results of our V2C-Net with comparisons against the baseline method and ground truth. As shown in Figure 4.12, the mel-spectrograms generated by the proposed V2C-Net are more similar to the ground-truth ones. Note that

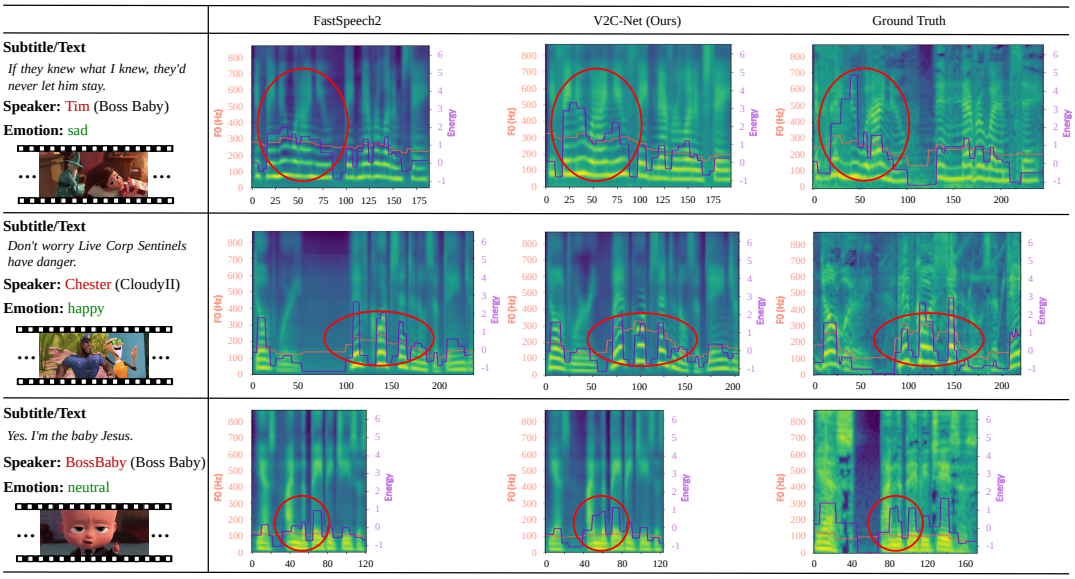


FIGURE 4.12. More visualized mel-spectrograms of generated and ground-truth audios. The orange curves are F_0 contours, where F_0 denotes the fundamental frequency of audio. The purple curves refer to the energy (volume) of audio. The horizontal axis is the duration of the audio. We highlight the main difference via a red circle.

the baseline method FastSpeech2 does not take the reference videos (*i.e.*, emotions) as inputs, which may lead to some misses of the prosody involved in the videos. The results further demonstrate the effect of the reference video when generating speech with rich emotions. Besides, the ranges of the pitch for the mel-spectrograms are various due to the different emotions. For example, the pitch of the mel-spectrogram would be more drastic with the emotions "*happy*" or "*sad*", while it would be more smooth if the emotion is "*neutral*".

4.5 Conclusion

In this paper, we propose a novel task, Visual Voice Cloning (V2C), extended from conventional Voice Cloning. It seeks to convert a paragraph of text to speech with the desired voice and emotion from reference audio and video, respectively. To facilitate the research of this new task, we collect the first V2C-Animation dataset. We also design a V2C baseline method, namely Visual Voice Cloning Network (V2C-Net), based on FastSpeech2 (a widely used TTS framework). Moreover, to assess the quality of the generated speech, we propose a variant of MCD-DTW, called MCD-DTW-SL, which is weighted by speech length. The experimental results demonstrate the effectiveness of our V2C-Net, but it is still far from saturation.

Statement of Authorship

Title of Paper	Likelihood-Based Text-to-Image Evaluation with Patch-Level Perceptual and Semantic Credit Assignment
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to IEEE Transactions on Image Processing

Principal Author

Name of Principal Author (Candidate)	Qi Chen		
Contribution to the Paper	Proposed the ideas, conducted experiments, and wrote the manuscript.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a t in this thesis. I am the primary author of this paper.		
Signature		Date	03/12/2023

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Chaorui Deng		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Zixiong Huang		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Bowen Zhang

Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Mingkui Tan		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Name of Co-Author	Qi Wu		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	03/12/2023

Please cut and paste additional co-author panels here as required.

Chapter 5

Likelihood-Based Text-to-Image Evaluation

5.1 Introduction

Recent works [Ramesh et al., 2022; Saharia et al., 2022; Tan et al., 2020; Yang et al., 2021; Yu et al., 2022] have made promising achievements on the text-to-image generation task, which can produce photo-realistic images or impressive artworks according to the text inputs. However, automatic evaluation of the text-to-image generation performance has always been a challenging problem, especially in complex, open-vocabulary scenarios. Most existing text-to-image generation models, including those state-of-the-art models, still rely on the popular evaluation metrics in general image generation, such as the Inception Score (IS) [Salimans et al., 2016] and Fréchet Inception Distance (FID) [Heusel et al., 2017].

The IS leverages a pre-trained image recognition model (Inception-V3 [Szegedy et al., 2016]) to provide the recognition confidence of the generated images to assess the generative performance. A higher confidence score may indicate that the generated images are more similar to those used for training the Inception-V3 (*i.e.*, ImageNet [Deng et al., 2009]). However, the critical issue is that the similarity with the pre-training images does not necessarily reflect the perceptual quality or semantic alignment of the generated images, since there exists a significant domain shift between the pre-training data (which has relatively limited semantic diversity) and the open-vocabulary natural languages. Besides, the pre-trained model can also be over/under-fitted. Thus, the confidence scores may be unreliable. Likewise, the followers of IS, such as Classification Accuracy Score (CAS) [Ravuri and Vinyals, 2019] and Semantic Object Accuracy (SOA) [Hinz et al., 2020], could also suffer from these problems.

The FID, on the other hand, computes the distribution distance between the real images and the generated images in a latent space, where the backbone of the Inception-V3 is used to map those images into the latent space, and a multivariate Gaussian assumption is imposed on the latent space to estimate the distribution statistics. Although it addresses the domain shift problem of IS (to some extent) by considering real images during evaluation, FID still faces several issues. First, the assumption of the multivariate Gaussian distribution is groundless, making it a poor indicator of the perceptual quality of generated images. Second, FID is designed only to consider the overall similarity of two sets of images, which cannot measure the semantic alignment between each input text and the corresponding generated image. Finally, both IS and FID is inefficient, as they have to process lots of image samples using the Inception V3 model to stabilize their evaluation results. Most of these drawbacks can also be found in Kernel Inception Distance (KID) [Bińkowski et al., 2018], a kernelized variant of FID. CLIPScore [Hessel et al., 2021] typically produces

a similarity score between images and text embeddings. However, it evaluates the semantic similarity between the content of images and textual descriptions but does not provide a likelihood estimation for generated images.

In this thesis, we propose a new metric for text-to-image evaluation based on likelihood estimation. Specifically, given n image descriptions $X = \{x_i\}_{i=1}^n$ and the corresponding generated images $Y = \{y_i\}_{i=1}^n$, we leverage a pre-trained likelihood-based text-to-image model $\mathcal{G}$ to estimate the likelihood of each y_i given x_i as condition, *i.e.*, $P_{\mathcal{G}}(y_i|x_i)$. Ideally, when the model-estimated density is closely aligned with the real-world data density, then $P_{\mathcal{G}}(y_i|x_i)$ would effectively indicate how likely the image y_i exists in real-world when the text x_i is given [Theis et al., 2016]. In this sense, to achieve a high $P_{\mathcal{G}}(y_i|x_i)$, y_i should not only be semantically aligned with x_i, but also be perceptually good enough as the real-world images. Otherwise, we will get a low likelihood score for y_i. In fact, average log-likelihood has been the *de facto* standard and widely considered as the default measure for detecting out-of-distribution data [Ren et al., 2019a] or quantifying generative modeling performance [Theis et al., 2016; Van Den Oord et al., 2016]. However, existing text-to-image evaluation methods have to resort to the aforementioned alternatives due to the lack of a well-trained likelihood estimator $\mathcal{G}$.

We address this problem using the recently emerged likelihood-based autoregressive image generation models [Ramesh et al., 2021; Wang et al., 2022a], which have shown impressive text-to-image generation performance. Concretely, they first rely on a discrete variational autoencoder (dVAE) [Esser et al., 2021; Van Den Oord et al., 2017] to learn a codebook so that an image can be converted into a sequence of visual codes and vice versa. The generative modeling is then accomplished by a transformer model [Vaswani et al., 2017], which takes the text x_i as input and predicts the likelihood of the visual codes autoregressively for the image y_i. In this way, we can obtain $P_{\mathcal{G}}(y_i|x_i)$ as the product of the likelihood of the visual codes. In the evaluation setting, y_i is given, and its visual codes can be directly obtained from the dVAE. Thus, their likelihoods can be efficiently predicted in parallel with just one forward pass of the transformer, instead of the step-by-step autoregressive manner.

Nevertheless, we observe that even with such a good estimator, $P_{\mathcal{G}}(y_i|x_i)$ can still be dominated by: 1) The visual codes whose corresponding patterns occasionally appear in the generated images but rarely exist in real-world scenes. We refer to them as visual codes with low perceptual significance. Although they may not affect the overall perceptual quality of y_i as they generally account for only a small ratio in the whole image, they tend to have extremely small likelihoods and drag down the overall likelihood $P_{\mathcal{G}}(y_i|x_i)$; 2) The visual codes that are semantically irrelevant to the text, *i.e.*, those corresponding to the background. which exhibits low semantic significance, but usually accounts for the largest portion of the image and thus affects $P_{\mathcal{G}}(y_i|x_i)$ mostly.

To tackle these issues, we explicitly introduce a credit assignment strategy based on the perceptual and semantic significance of each visual code. Specifically, we first design an indicator function to eliminate the factors of the rare visual codes from $P_{\mathcal{G}}(y_i|x_i)$, so that only the perceptually significant codes are considered. Then, we design a semantic scoring function using a pre-trained image-text matching model CLIP [Radford et al., 2021] to down-weight the visual codes with low semantic significance. Since the original CLIP can only predict an alignment score for the whole image and the text, we make non-trivial modifications to make it suitable for patch-level image-text matching, including a new inference-time model architecture and a new scoring function. We denote our proposed metric as **L**ikelihood-based t**E**xt-to-**I**mage evaluation with **C**redit **A**ssignment (LEICA).

We verify the effectiveness of the proposed LEICA score on 11 state-of-the-art text-to-image generation models, including three Generative Adversarial Network (GAN) based models [Hinz et al., 2020; Ye et al., 2022; Zhou et al., 2022], four autoregressive models [Ding et al., 2021, 2022; Ramesh et al., 2021, 2022], and four diffusion models [Gu et al., 2022; Nichol et al., 2022; Rombach et al., 2022; Saharia et al., 2022]. We also evaluate LEICA on multiple text-to-image datasets, including one open-domain dataset COCO [Lin et al., 2014], and two specific-domain datasets CUB [Wah et al., 2011] and Oxford-Flower [Nilsback and Zisserman, 2008]. We provide extensive analysis, including human studies, to show the superiority of our evaluation metric.

5.2 Method

5.2.1 Overview

The overall pipeline of our evaluation metric LEICA consists of a likelihood estimation step and two credit assignment steps on the perceptual and semantic significance, respectively. First of all, the likelihood of the generated image y_i is estimated by

$$P_{\mathcal{G}}(y_i|x_i) = \prod_{t=1}^{m} P_{\mathcal{G}}(c_{i,t}|c_{i,<t}, x_i), \tag{5.1}$$

where $\{c_{i,t}\}_{t=1}^{m}$ are the visual codes obtained from the dVAE given y_i as input. Each $c_{i,t}$ is mainly responsible for representing the information near a specific spatial (patch) location in y_i, and t is the flattened index of the patch. m is the total number of patches, and $c_{i,<t}$ indicates the codes before t, *i.e.*, $c_{i,1}, ..., c_{i,t-1}$. Note that Eq. (5.1) serves as the basis of the proposed evaluation metric, where this vanilla likelihood estimation can provide a generally unbiased but unstable evaluation due to the existence of perceptually & semantically insignificant visual codes.

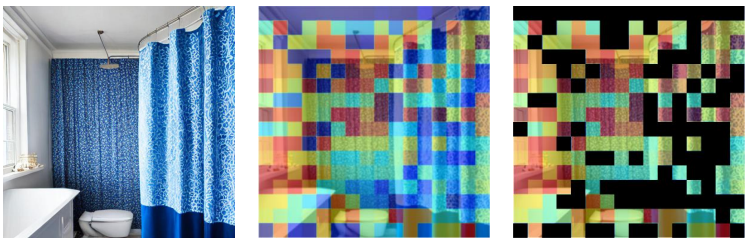
Then, we introduce an indicator function H to explicitly eliminate the visual codes with low perceptual significance from Eq. (5.1), *i.e.*,

$$\frac{1}{m}\sum_{t=1}^{m} H(P_{\mathcal{G}}(c_{i,t}|c_{i,<t}, x_i)). \tag{5.2}$$

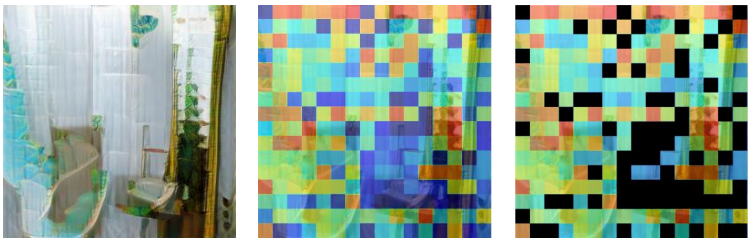
We convert the likelihoods into log-likelihoods inside H so the summation sign is used here, and the scores are averaged among all codes. Lastly, we leverage the CLIP [Radford et al., 2021], a well-trained image-text matching model (denoted by $\mathcal{M}$) to focus on the locations with high semantic significance. As the CLIP can only provide the global matching score between an image-text pair, we modify its inference-time architecture to get $\mathcal{M}'$, and further devise a scoring function S, to assign a semantic alignment score for each location t. The final evaluation score of LEICA on a sample (x_i, y_i) is then computed by

$$\text{LEICA}(x_i, y_i) = \frac{1}{m}\sum_{t=1}^{m} S(x_i, y_i, t) H(P_{\mathcal{G}}(c_{i,t}|c_{i,<t}, x_i)). \tag{5.3}$$

More details are in the following.



(A) Stable Diffusion



(B) OPGAN

FIGURE 5.1. Generated images for the caption *"A bathroom with a blue shower curtain and walls"* with OPGAN and stable diffusion (the first column) and their visualized likelihood maps (the second column). Although stable diffusion produces a much better result, its average likelihood is 1.401×10^{-4}, less than OPGAN (8.020×10^{-4}). After removing the rare codes (the last column), their likelihoods increase to 3.297×10^{-3} and 1.670×10^{-3}, respectively, where stable diffusion is considered to be better.

5.2.2 Evaluation via Likelihood Estimation

To directly estimate the likelihood of a generated image y_i conditioned on the image description x_i, we leverage a well-trained likelihood-based text-to-image generator $\mathcal{G}$, which consists of a dVAE model and a transformer model.

Specifically, the dVAE is originally learned to convert an image into a sequence of visual codes and vice versa, so that the generation of images can be reformulated as the generation of their code sequences. It is a convolutional encoder-decoder architecture accompanied by a learnable codebook that contains K visual codes. In our evaluation setting, we can directly use the trained dVAE to tokenize the generated images. Suppose $y_i \in \mathbb{R}^{3\times H\times W}$, the encoder of dVAE first maps it into a feature map $\boldsymbol{f}_i \in \mathbb{R}^{d\times h\times w}$ where $h, w < H, W$. Then, each spatial location in $\boldsymbol{f}_i$ will be represented by its nearest code retrieved from the codebook, and y_i is thus represented by $m = h \times w$ visual codes, *i.e.*, $\{c_{i,t}\}_{t=1}^{m}$. The decoder of dVAE is used for reconstructing y_i from $\{c_{i,t}\}_{t=1}^{m}$, which is not required in our setting.

The transformer is trained to model the joint distribution between image descriptions and the code sequences of the corresponding images. The joint distribution is explicitly optimized to have a causal chain structure, where each visual code is predicted conditioning only on previous visual codes and the text. Thus, during the generation procedure, the visual codes are predicted in an autoregressive manner. When it comes to our evaluation setting, both x_i and $\{c_{i,t}\}_{t=1}^{m}$ are given, we thus can obtain the conditional likelihood $P_{\mathcal{G}}(\{c_{i,t}\}_{t=1}^{m}|x_i)$ by one forward pass of the transformer, which serves as a lossless approximate to $P_{\mathcal{G}}(y_i|x_i)$. Due to its causal chain structure, $P_{\mathcal{G}}(\{c_{i,t}\}_{t=1}^{m}|x_i)$ can be factorized as the product of the likelihood of each $c_{i,t}$, leading to the form in Eq. (5.1). We further consider using log-likelihood instead of likelihood to avoid too small values, *i.e.*, $\ln P_{\mathcal{G}}(y_i|x_i) = \sum_{t=1}^{m} \ln P_{\mathcal{G}}(c_{i,t}|c_{i,<t}, x_i)$.

5.2.3 Credit Assignment of Perceptual Significance

While it is feasible to directly assess the text-to-image generation performance using the likelihood estimated above, we observed that there usually exists a small portion of codes in $\{c_{i,t}\}_{t=1}^{m}$ that have extremely small likelihoods (typically less than 10^{-9}) and severely drag down the overall likelihood of y_i. As shown in Figure 5.1, this phenomenon occurs no matter what the actual qualitative performance is. As a result, a well-synthesized image may be rated lower than a poorly-synthesized image just because it contains more rarely-used visual codes.

To handle this problem, we introduce an indicator function H on the likelihood of each visual code into Eq. (5.1) to explicitly focus on those with high perceptual significance:

$$\begin{aligned} H(P_{\mathcal{G}}(c_{i,t}|c_{i,<t}, x_i)) = \mathbb{1}(\ln P_{\text{data}}(c_{i,t}) - \lambda) \\ * \ \max(\ln P_{\mathcal{G}}(c_{i,t}|c_{i,<t}, x_i) - \lambda, 0), \end{aligned} \tag{5.4}$$

Here, $\mathbb{1}(\cdot)$ equals 1 for positive inputs and 0 otherwise. $P_{\text{data}}(c_{i,t})$ is the prior probability of $c_{i,t}$ estimated from real images. $\lambda < 0$ is a threshold that controls whether the effect of $c_{i,t}$ should be considered or not according to both its prior probability $P_{\text{data}}(c_{i,t})$, *i.e.*, the perceptual significance of $c_{i,t}$ in real images, and the predicted likelihood $P_{\mathcal{G}}(c_{i,t}|c_{i,<t}, x_i)$, which indicates its perceptual significance in the generated image. In this way, only the visual codes whose predicted likelihoods are larger than e^{λ} are rewarded with positive values, and those values will only take effect when the visual codes have higher prior probability than e^{λ}. Otherwise, $H(P_{\mathcal{G}}(c_{i,t}|c_{i,<t}, x_i))$ will be zeroed out from the summation in Eq. (5.2). By default, λ is set to a small value, *e.g.*, $\ln 10^{-9}$. Thus, H will only affect a few codes in $\{c_{i,t}\}_{t=1}^{m}$ that have extremely small likelihoods.

5.2.4 Credit Assignment of Semantic Significance

Since the background information (that is irrelevant to the input description x_i) usually constitutes a large part of the generated image y_i, it can easily dominate the overall likelihood, as shown in Figure 5.2. Therefore, it would be beneficial to focus on the semantically significant part of y_i, which is also very close to human behavior. To achieve this, an intuitive idea is to use a pre-trained image-text matching model with zero-shot recognition abilities to estimate the semantic significance. We choose CLIP [Radford et al., 2021], a multi-modality model trained on web-scale image-text pairs.

Adapt CLIP for Patch-Level Scoring

The CLIP model we adopted (denoted by $\mathcal{M}$) has two transformer encoders to process images and texts, respectively. A feature vector is extracted from each modality with the corresponding encoder and is linearly projected into the same dimensional space and $L2$-normalized. The optimization of $\mathcal{M}$ is achieved via contrastive learning, where the feature vectors that belong to the same image-text pair are pulled close, and the others are pushed away. However, when used for our evaluation setting, although we can assess how the generated y_i and the text x_i are semantically close using the cosine similarity of their feature vectors, we are unfortunately unable to assess the alignment at a more detailed level, *i.e.*, how each spatial location in y_i aligns with x_i. This is caused by the feature extraction mechanism of the image encoder. It takes $s \times s$ non-overlapping patches divided from the image and an extra `[CLS]` token as

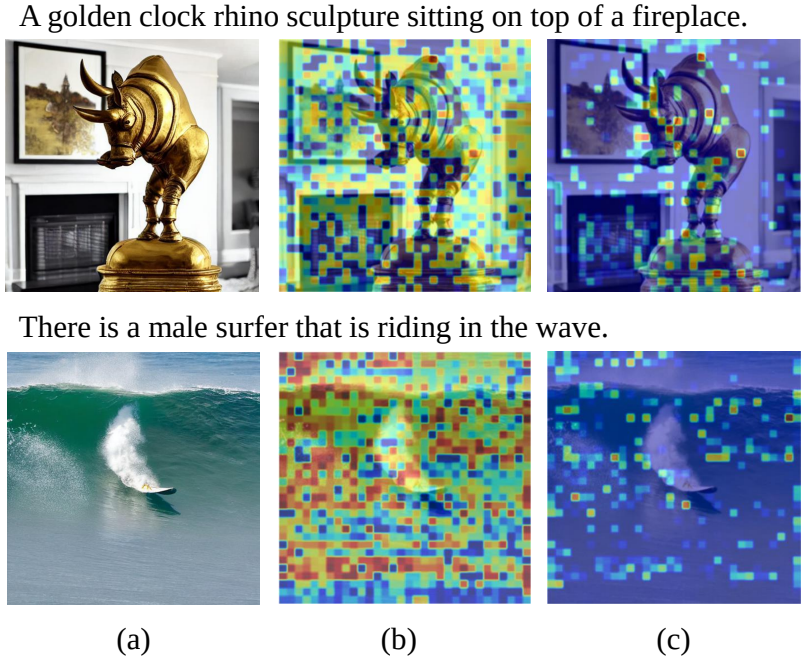


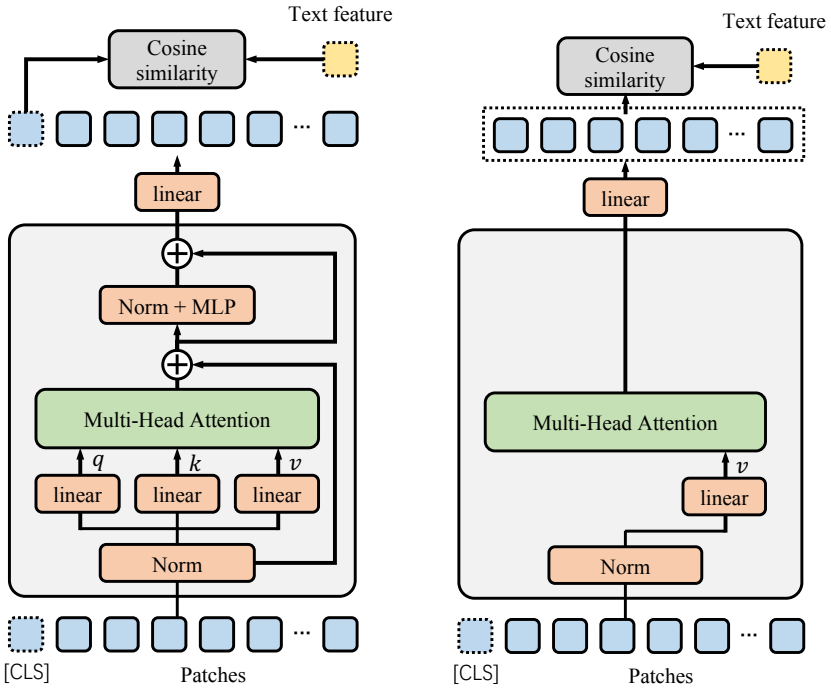
FIGURE 5.2. (a) are the images generated by the stable diffusion model. (b) and (c) are the likelihood maps before and after the semantic credit assignment, respectively. For these two images, their overall likelihoods are 2.422×10^{-4} and 4.123×10^{-4}, respectively, where the second image which misses the foreground object *"male surfer"* is rated higher than the first image. This can be corrected after the semantic credit assignment, where the overall likelihoods become 1.049×10^{-4} and 4.704×10^{-5} in (c).

input, and the final embedding of the `[CLS]` token is adopted as the output image feature vector, while the final patch embeddings are all ignored. Consequently, the contrastive loss is only applied on the `[CLS]` token, making the similarity between the text feature vector and the final patch embeddings uncertain during inference. We address this problem by proposing a new inference-time architecture for $\mathcal{M}$, termed as $\mathcal{M}'$.

Our basic idea is to look for the patch representations in the computation graph of $\mathcal{M}$ that is closest to the loss computation node, which will also be the most relevant patch representations to the text feature vector. As shown in Figure 5.3, the last interaction between the patch embeddings and the `[CLS]` token embedding occurs within the last self-attention operation of the image encoder. Thus, all the operations afterwards, *e.g.*, the last feed-forward network, the last layer normalization, *etc.*, are only used for the `[CLS]` token and have no benefit to the modeling of patch-level image-text relation. We thus remove them from $\mathcal{M}$. Moreover, inside the last self-attention operation, the `[CLS]` token is updated with the patch representations output by the value projection layer, denoted by $\{\boldsymbol{v}_{i,t}\}_{t=1}^{s^2}$, while the query and key projection layers are only used for computing the self-attention map. Therefore, we leverage $\{\boldsymbol{v}_{i,t}\}_{t=1}^{s^2}$ to compute the patch-level alignment based on their cosine similarities with the feature vector of x_i, denoted by $\boldsymbol{x}_i$.

$$\phi_i(t) = \frac{(\boldsymbol{W}^\top \boldsymbol{v}_{i,t})^\top \boldsymbol{x}_i}{\|\boldsymbol{W}_1^\top \boldsymbol{v}_{i,t}\|_2^2 \cdot \|\boldsymbol{x}_i\|_2^2}, \ t = 1, ..., s^2 \tag{5.5}$$

where $\boldsymbol{W}$ is the final linear project layer of the image encoder that transforms $\boldsymbol{v}_{i,t}$



(A) CLIP (B) Ours

FIGURE 5.3. The computation graph of the last layer in the original CLIP and our modified CLIP.

into the same dimensional space as $\boldsymbol{x}_i$. As shown in Figure 5.7, with our modified CLIP model $\mathcal{M}'$, the obtained patch-level similarity scores are closely aligned with the foreground part of the images.

Scoring Function for Semantic Significance

To obtain the final semantic significance scores for the predicted visual codes $\{c_{i,t}\}_{t=1}^{m}$, we first apply a bilinear interpolation on ϕ_i to resize its sequence length from s^2 to m. We then clip the negative values in ϕ_i to zeros so that the unrelated parts in y_i will not affect the final evaluation score. Finally, we can further consider the global image-text alignment score with the original CLIP model $\mathcal{M}$, to enhance the ability of our proposed LEICA metric on the semantic alignment evaluation. Our scoring function for semantic significance is defined as follows:

$$S(x_i, y_i, t) = e^{\psi_i/\tau} \max(\phi_i(t), 0), \tag{5.6}$$

where ψ_i is the cosine similarity between the feature vectors of y_i and x_i. As the cosine similarities obtained from CLIP are generally within $(0.2, 0.4)$, we use the exponential function to scale the values to a broader range to distinguish the semantically aligned/unaligned results better. τ is a temperature hyper-parameter that controls the scaling effect, which is set to 0.07 following [Chen et al., 2020e]. The final LEICA score is calculated via Eq. (5.3).

5.3 Experiments

5.3.1 Meta Evaluation

To assess the quality of the proposed evaluation metrics, following [Yuan et al., 2021], we use four measures for the meta-evaluation: 1) **Accuracy**, in our experiments, measures the percentage of correct judgment made by the metrics on whether a text is matched with an image or is mismatched/altered. 2) **Kendall's Tau Correlation** measures the ordinal association between two sets of data. 3) **Pearson Correlation** measures the linear correlation between two variables. 4) **Spearman Correlation** assesses the consistency between the ranking of two variables.

5.3.2 Implementation Details

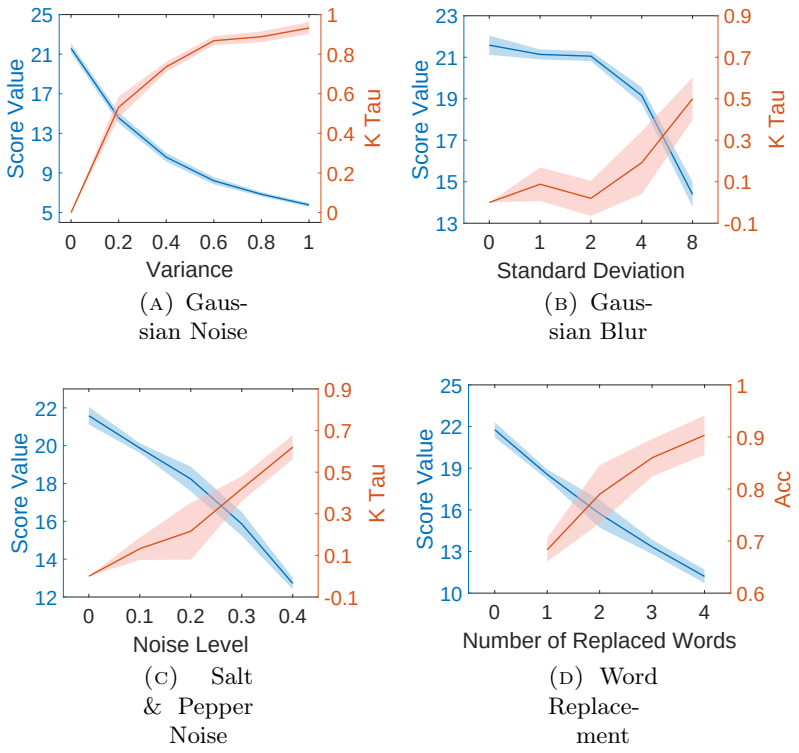
For the pre-trained likelihood-based text-to-image generation model G, we use OFA [Wang et al., 2022a], which is the best-performed and open-sourced autoregressive text-to-image generation model on MSCOCO [Lin et al., 2014] dataset. Still, we can always update it to the most powerful model in the future for better evaluation capability. For the images used for evaluation, we generate them using the captions from the validation set of Karpathy's split on MSCOCO, where the images and captions are semantically diverse enough and are also commonly used to evaluate the performance of text-to-image generation models in previous works [Gu et al., 2022; Xu et al., 2018; Zhang et al., 2021a; Zhou et al., 2022]. Besides, we consider two datasets with specific domains: CUB [Wah et al., 2011] and Oxford-Flower [Nilsback and Zisserman, 2008] to test the effectiveness of our score. We set the default resolution for the generated images to 256×256, following the common practice in text-to-image generation.

5.3.3 Results on Manually Synthesized Data

We first manually introduce artifacts into the real-world image-text pairs from the MSCOCO dataset, so that we have oracle judgments for those data that can be used to assess the performance of LEICA. Specifically, for the effectiveness of LEICA on perceptual image quality, we add noises with multiple degrees into the images while keeping the paired texts unchanged. We consider several types of noises, including Gaussian Noise (GN), Gaussian Blur (GB), and Salt & Pepper Noise (SPN). For more complicated distortions, we combine a "funny mirror" distortion [fun] with these three noises, named GN+, GB+, and SPN+. Besides, we reconstruct images using the dVAE model of DALL-E [Ramesh et al., 2021] to obtain a type of model-based distortion. For semantic text-image alignment, we gradually replace the keywords (*i.e.*, those with strong semantic meanings) in the text with random words, or directly use mismatched text for the image, where the paired image stays unaltered. To ensure the reliability of our results, we run each experiment five times and calculate their mean values and standard deviations.

Perceptual Quality

As shown in Figure 5.4(A)-(C), as the disturbance degree gradually increases, the LEICA score of these images decreases accordingly, showing that the proposed LEICA score can reflect the perceptual quality of the given images. Furthermore, for each text x, we take its paired clean image y_{clean} as well as the corresponding noised one y_{noise} as a triplet, *i.e.*, $(x, y_{\text{noise}}, y_{\text{clean}})$. For each triplet, if the LEICA score of y_{clean} is higher than that of y_{noise}, it is concordant with the oracle judgment. Otherwise,



(A) Gaussian Noise (B) Gaussian Blur (C) Salt & Pepper Noise (D) Word Replacement

FIGURE 5.4. (A)-(C): The trend of the proposed LEICA score (Score Value) and the corresponding values of Kendall' Tau (K-Tau) when adding different noises with different disturbance degrees to the images. (D): The trend of the LEICA score and Accuracy (Acc) of its judgments *w.r.t.* semantic alignment when gradually replacing the keywords in texts.

the triplet is discordant. Based on this, we calculate Kendall's Tau coefficient for the triplets. From Figure 5.4(A)-(C), Kendall's Tau coefficient becomes larger as the disturbance becomes heavier, which further shows that LEICA can reliably measure the perceptual quality of the image.

Semantic Alignment

See the result in Figure 5.4(D), as the number of replaced words increases, the LEICA scores on the paired images drop clearly, which demonstrates that our proposed metric is sensitive to the semantic alignment between text and image. Moreover, the accuracy of making a correct judgment, *i.e.*, predicting a larger score for the paired image with clean text than that with noised text, increases steadily from 70% to more than 90%. This indicates that LEICA is able to perceive the mild changes in the semantic meaning of the text, and can also correctly assess the alignment degree between the image-text pairs.

Performance Analysis

We analyze the effect of components of LEICA, *i.e.*, the indicator function H for perceptual significance and the scoring function S for semantic significance. Notably, when all these components are removed, LEICA is annealed to the log version of Eq. (5.1). We calculate the accuracy of LEICA on discriminating matched or mismatched text-image pairs with a $1:1$ ratio. We set GN & GN+ with a variance of

TABLE 5.1. Ablation studies on the components of LEICA. The first row of results indicates the performance of LEICA. The last row is the performance of the averaged log-likelihood. "+" denotes combining the "funny mirror" distortion with the current distortion. dVAE: distortion using the dVAE model of DALL-E. Best results are highlighted with **bold**.

H	S	Acc	Kendall's Tau (↑)						
			GN	GB	SPN	GN+	GB+	SPN+	dVAE
√	√	**1.00**	**0.86**	**0.66**	**0.73**	**0.70**	**0.60**	**0.72**	**0.44**
	√	0.96	0.60	0.40	0.60	0.66	0.32	**0.72**	0.20
√		0.86	0.73	**0.66**	0.60	0.54	0.56	0.64	0.42
		0.54	0.33	-0.20	0.40	0.43	0.20	0.40	-0.04

TABLE 5.2. Pearson, Spearman and Kendall's Tau correlations between different evaluation metrics and human judgment. The highest correlation is **bold**.

Methods	Pearson (↑)	Spearman (↑)	K-Tau (↑)
IS [Salimans et al., 2016]	0.7175	0.4636	0.3091
FID [Heusel et al., 2017]	0.5932	0.2363	0.1636
KID [Bińkowski et al., 2018]	0.3039	0.2181	0.2000
SOA-C [Hinz et al., 2020]	0.8449	0.7363	0.5636
SOA-I [Hinz et al., 2020]	0.8301	0.7909	0.6000
CLIP-FID [Kynkäänniemi et al., 2023]	0.6419	0.5636	0.4545
LEICA w/o H	0.8640	0.8727	0.7454
LEICA w/o S	0.8918	0.8636	0.7091
LEICA w/o $e^{\psi/\tau}$	0.9101	0.9091	0.7818
LEICA (base)	0.8979	0.9181	0.7818
LEICA (large)	**0.9365**	**0.9181**	**0.7818**

0.2, GB & GB+ with a standard deviation of 8, and SPN & SPN+ with a noise level of 0.4.

As shown in Table 5.1, when discarding the perceptual significance indicator H from LEICA, the results reduce clearly in all the meta-evaluation metrics, *e.g.*, the accuracy drops from 1.00 to 0.96 and the Kendall's Tau coefficient drops from 0.86 to 0.60 when adding Gaussian noise (GN). Note that the performance drops more severely in the settings related to the image disturbances, indicating that H mainly focuses on the perceptual image quality. Besides, when removing the patch-level scoring function S, the performance of LEICA drops dramatically in terms of accuracy (1.00 to 0.86), demonstrating its importance in assessing the semantic text-image alignment. Lastly, we show the performance of a naive likelihood-based metric, *i.e.*, the averaged log-likelihood, in the last row of the table. We find that even with such a naive strategy, it can also achieve the positive Kendall's Tau correlations with human judgment in most settings. Overall, the above results have shown the effectiveness of the basic likelihood-based evaluation paradigm and the proposed credit assignment strategies *w.r.t.* perceptual and semantic significance.

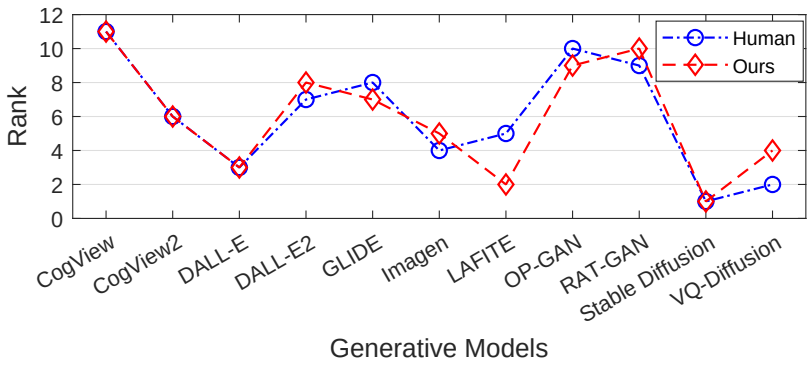


FIGURE 5.5. The ranks of each generative model derived by human raters and the proposed evaluation metric. Rank-1 means the best model while rank-11 is the worst one.

5.3.4 Results on Model Generated Data

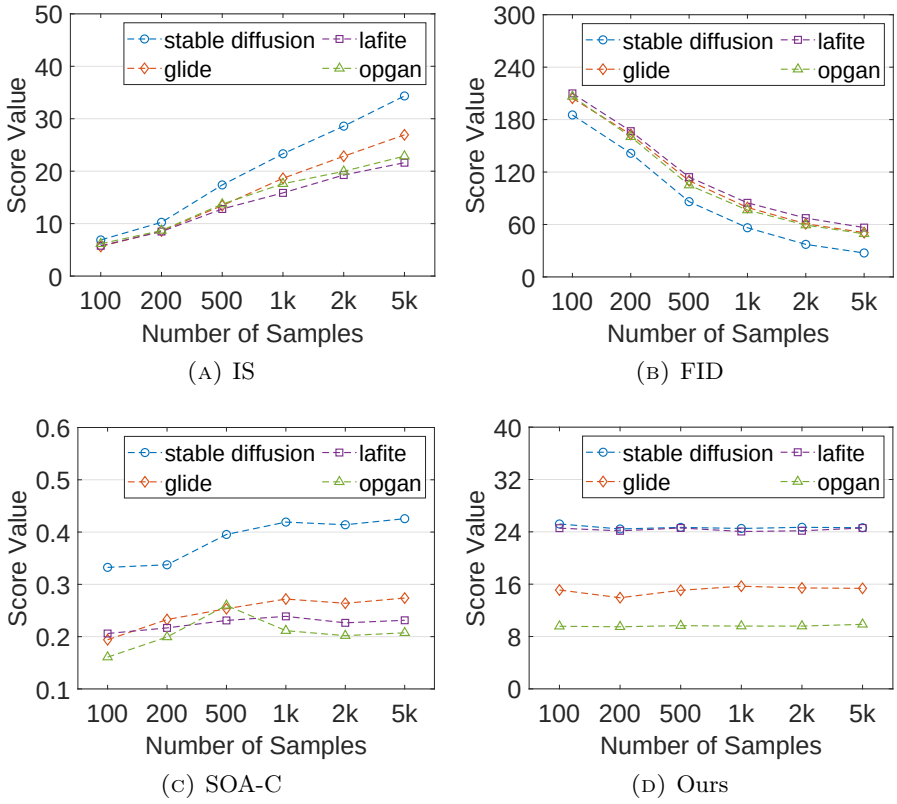
We investigate the effectiveness of the proposed metric on generated images. We consider 11 well-known generative models of various types, including Generative Adversarial Networks, Autoregressive models and Diffusion models: CogView [Ding et al., 2021], CogView2 [Ding et al., 2022], DALL-E [Ramesh et al., 2021], DALL-E2 [Ramesh et al., 2022], GLIDE [Nichol et al., 2022], Imagen [Saharia et al., 2022], LAFITE [Zhou et al., 2022], OP-GAN [Hinz et al., 2020], RAT-GAN [Ye et al., 2022], Stable Diffusion [Rombach et al., 2022], VQ-Diffusion [Gu et al., 2022]. Each model synthesizes $5,000$ images, the same as the size of the validation set of MSCOCO.

Alignment with Human Judgment

We conduct a human study to test the alignment between our evaluation metric and human judgment. Specifically, we randomly choose 100 texts from Karpathy's validation split, and obtain corresponding generated images from all 11 models, so that each testing sample for our human study contains one text and 11 images. We use a Mean Opinion Score (MOS) evaluation approach to score the test samples, where the score ranges from 1 to 10 in 1-point increments. For each testing sample, we ask the raters to score the generated image according to the perceptual image quality and semantic text-image alignment, and emphasize that these two aspects should be equally considered. Finally, we collect $8,479$ scores for all the $1,100$ testing samples, where each sample has 7.7 human judgments on average.

After we obtain the human study results, we conduct a meta-evaluation that compares our metric with six commonly used evaluation metrics in the field of text-to-image generation, *i.e.*, Inception Score (IS) [Salimans et al., 2016], Fréchet Inception Distance (FID) [Heusel et al., 2017], CLIP-FID [Kynkäänniemi et al., 2023], Kernel Inception Distance (KID) [Bińkowski et al., 2018] and Semantic Object Accuracy (SOA) [Hinz et al., 2020] for class average (SOA-C) and image average (SOA-I). For each metric, we compute the Pearson, Spearman, and Kendall's Tau correlation coefficient between the model rankings under its judgment and the human judgment.

As shown in Table 5.2, our evaluation metric consistently outperforms the baseline metrics on all meta-evaluations by a large margin. Besides, we also investigate the impact of each component in LEICA as in Section 5.3.3. From this table, we can draw a similar conclusion as in Table 5.1, where discarding any of these components leads to a decrease in performance. Nevertheless, all the ablation results of our method are still superior to the compared metrics in terms of the correlation with human judgment.



(A) IS (B) FID

(C) SOA-C (D) Ours

FIGURE 5.6. Impacts of the number of evaluated samples. We test three commonly used evaluation metrics (*i.e.*, IS, FID and SOA-C) as well as the proposed LEICA for the generated images derived from four different generative models.

Besides, in Table 5.2, we also show the impact of the term $e^{\psi/\tau}$ in Eq. (5.6): remove it decreases the performance.

Moreover, in Figure 5.5, we plot the model ranking derived from human raters and the ranking based on the proposed LEICA. We observe that the two rankings yield very similar trends, which means our metric is well aligned with human judgment and would not be biased to any specific model. Lastly, comparing the performance of LEICA (OFA-base) and LEICA (OFA-large), in Table 5.2, we find that a better likelihood estimator will improve the reliability of LEICA, suggesting its scalability to future generative foundation models.

Robustness/Reliability of Used Pre-trained Models

Indeed, the pre-trained models (*e.g.*, OFA) have limitations in accurately estimating likelihood, which can hinder their performance in evaluating more powerful generative models. To alleviate this, we propose two credit assignment strategies based on perceptual and semantic significance. They potentially mitigate the constraints imposed by the inherent imperfections of the pre-trained model and improve its evaluation performance (*i.e.*, results in Table 5.2), which also outperforms existing evaluation metrics in all the experiments.

Besides, using pre-trained models can lead to prediction errors, and may not be perfectly aligned with human perception. Thus, we propose to consider the semantic

TABLE 5.3. Pearson, Spearman and Kendall's Tau correlations between different evaluation metrics and human judgment on CUB & Oxford-Flower. The highest correlation is **bold**.

Methods	Pearson (↑)	Spearman (↑)	K-Tau (↑)
CUB			
IS [Salimans et al., 2016]	0.2636	0.3	0.2
FID [Heusel et al., 2017]	0.7033	**0.7**	**0.6**
KID [Bińkowski et al., 2018]	0.6806	0.6	0.4
LEICA (Ours)	**0.8189**	**0.7**	**0.6**
Oxford-Flower			
IS [Salimans et al., 2016]	0.3166	0.2	0.2
FID [Heusel et al., 2017]	0.4719	0.5	**0.4**
KID [Bińkowski et al., 2018]	-0.3113	-0.1	0.0
LEICA (Ours)	**0.4880**	**0.6**	**0.4**

and perceptual significance of different areas in the generated image, which down-weight areas with extremely small prior or posterior likelihood values, and also highlight areas corresponding to foreground objects. Therefore, our metric is more robust and reliable than the naive likelihood estimated using a pre-trained model, and is better aligned with human judgments. This can be supported by the visualizations in Figures 5.1 and 5.2, and the results in Tables 5.1 and 5.2.

Moreover, prediction error is a common issue in all learning-based methods. In fact, all previously popular metrics like FID, IS, and CLIPScore suffer from this problem. To mitigate it, we carefully choose the pre-trained models, *i.e.*, OFA is the best-performed open-sourced autoregressive model recently, and CLIP is the best text-image alignment model available. Thus, they may be less influenced by this issue. Besides, as the primary focus of this thesis is to build a new text-to-image evaluation metric/paradigm rather than being limited to any particular pre-trained model, it can further evolve with more powerful text-to-image likelihood estimators and alignment models.

Impacts of Number of Evaluated Samples

In this part, we investigate how the number of evaluation samples influences the evaluation results for our LEICA metric and three baseline metrics, including IS, FID and SOA-C. We calculate the scores of these metrics using different numbers of generated samples ranging from 100 to $5,000$ for four different generative models. In Figure 5.6, the values of IS and FID are severely affected by the number of samples. SOA-C performs more stable than IS and FID, but still suffers from fluctuation when the number of samples is changed. Therefore, these metrics have to use a very large number of samples to make their evaluation results stable. In contrast, the scores from the proposed LEICA are very stable, which means that our metric is able to evaluate the performance of the generative models using only a few (one hundred) generated samples, making it very efficient in practice.

Evaluating Samples from the Same Model & Text

Besides accessing different generative models, we further investigate LEICA in another useful scenario, *i.e.*, assessing different samples generated from the same model using the same text. Specifically, we randomly choose 20 texts from COCO and generate 5

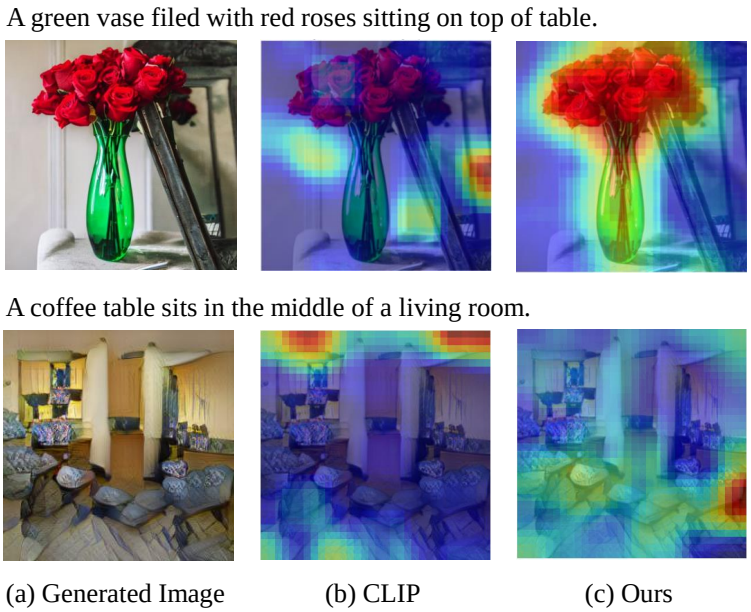


FIGURE 5.7. The semantic significance maps obtained from the original CLIP and our modified CLIP on generated images.

images for each text with Stable Diffusion (SD) and DALL-E (DE), respectively. Two raters are asked to score these images, and the Pearson correlations between their judgments are 0.203 (SD) and 0.298 (DE), showing that this is a very challenging setting even for humans. Nevertheless, LEICA still achieves positive correlations with these raters (Pearson: 0.105 for SD and 0.257 for DE), indicating that our proposed metric has the ability to evaluate individual samples in this setting.

Evaluation Performance on Specific Domains

In addition to COCO, in Table 5.3, we also provide results on the CUB dataset, where LEICA achieves a 0.819 Pearson correlation with the human judgment, clearly better than IS (0.264), FID (0.703) and KID (0.681). SOA is not compared as it is designed only for COCO. We further evaluate LEICA on Oxford-Flower using the same setting, and it still yields the best performance (*e.g.*, Pearson: 0.488), while IS, FID and KID achieve 0.317, 0.472, and -0.311, respectively. This demonstrates the generalization ability of LEICA, especially in class-specific domains.

Visualization Results

We visualize the patch-level alignment performance of the original CLIP and our modified model. Figure 5.7 shows that our modified CLIP is able to capture the patches semantically aligned with the given texts in a well-synthesized image (row 1) while suppressing all the regions for a low-quality image (row 2). However, the original CLIP fails to capture the semantically significant regions of images described in the given texts. These results verify the effectiveness of our modified CLIP.

5.3.5 More Discussions

More Visual Results

We provide more visual results for the impact of each component in our LEICA. As shown in Figure 5.8, by using $H(\cdot)$ for perceptual significance, the likelihoods of

extremely small values have been removed observably. While we further incorporate the $S(\cdot)$ for semantic significance, the proposed LEICA is able to clearly suppress the background (*i.e.*, the regions that are semantically irrelevant to the given text). These results demonstrate the effectiveness of each proposed component in the LEICA.

Rank Comparison with Baseline Evaluation Metrics

Like Figure 5.5, we plot the rank of each generative model from human beings as well as that from different evaluation metrics in Figure 5.9, including Inception Score (IS) [Salimans et al., 2016], Fréchet Inception Distance (FID) [Heusel et al., 2017], Kernel Inception Distance (KID) [Bińkowski et al., 2018], Semantic Object Accuracy (SOA) [Hinz et al., 2020] for class average (SOA-C) & image average (SOA-I), and CLIP-FID [Kynkäänniemi et al., 2023]. As shown in Figure 5.9, the visual results show that compared with baseline evaluation metrics, our LEICA is more consistent with human judgment.

Cheating Metric by Overfitting Metric Model

It is a common issue in learning-based evaluation metrics, including widely used metrics such as FID, IS and CLIPScore. While our LEICA serves as a more effective benchmark for evaluation, a robust evaluation process should involve multiple metrics and validation techniques to ensure the reliability of the results. Cheating LEICA (or any other metric) is meaningless as it may have adverse effects on other evaluation metrics or generate poor-quality images that are easily noticeable by humans.

5.4 Conclusion

In this thesis, we propose a likelihood-based metric for text-to-image evaluation based on the intuition that a good likelihood estimator will reflect how likely a generated image is to exist in the real world given the text. Besides, we propose two credit assignment strategies *w.r.t.* perceptual significance and semantic significance. The former eliminates the effect of the perceptually insignificant image regions, which may not affect the overall perceptual quality of a generated image but severely drag down its overall likelihood. The latter down-weights the effect of the background regions with low semantic significance, which are usually not related to the text but account for the largest portion of the image. We evaluate our metric on 11 models and 3 datasets, showing that it is effective, efficient, and stable.

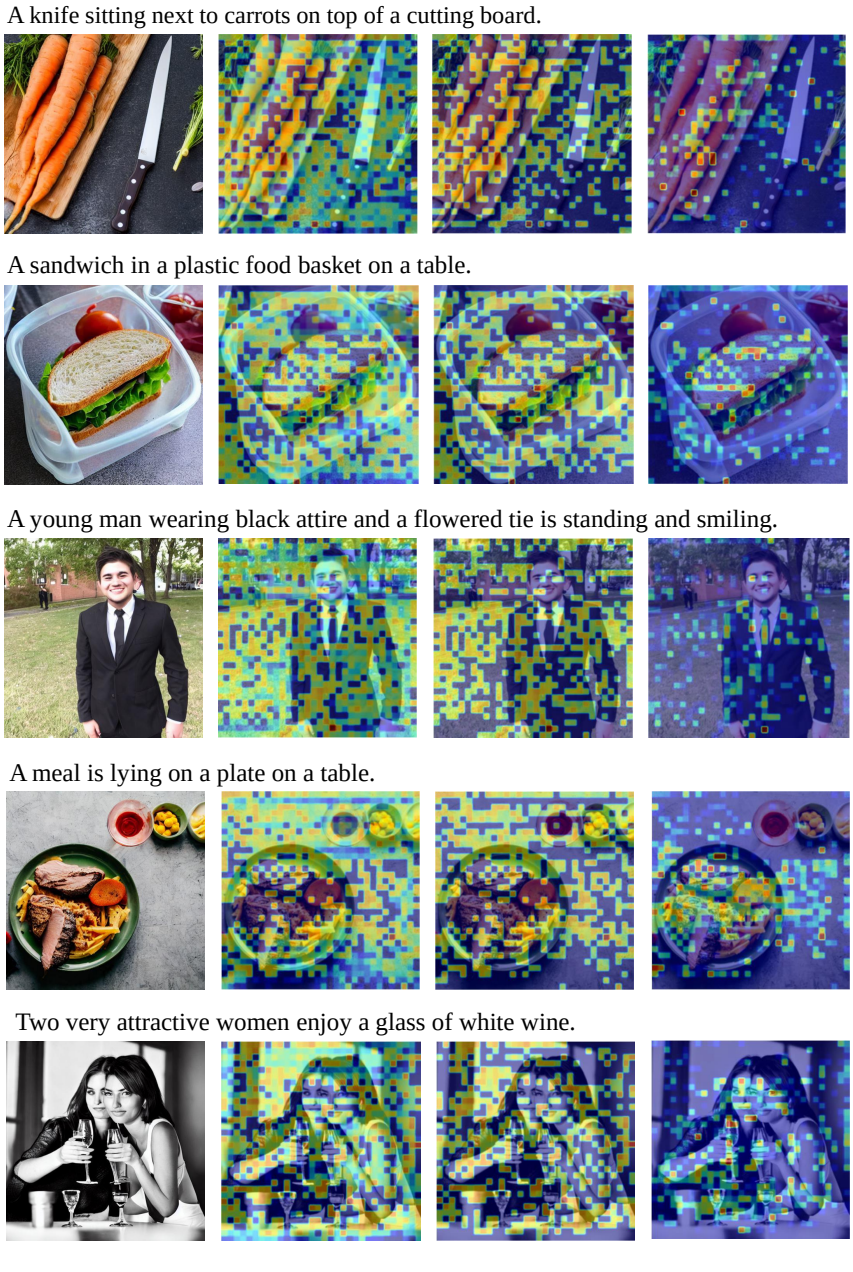
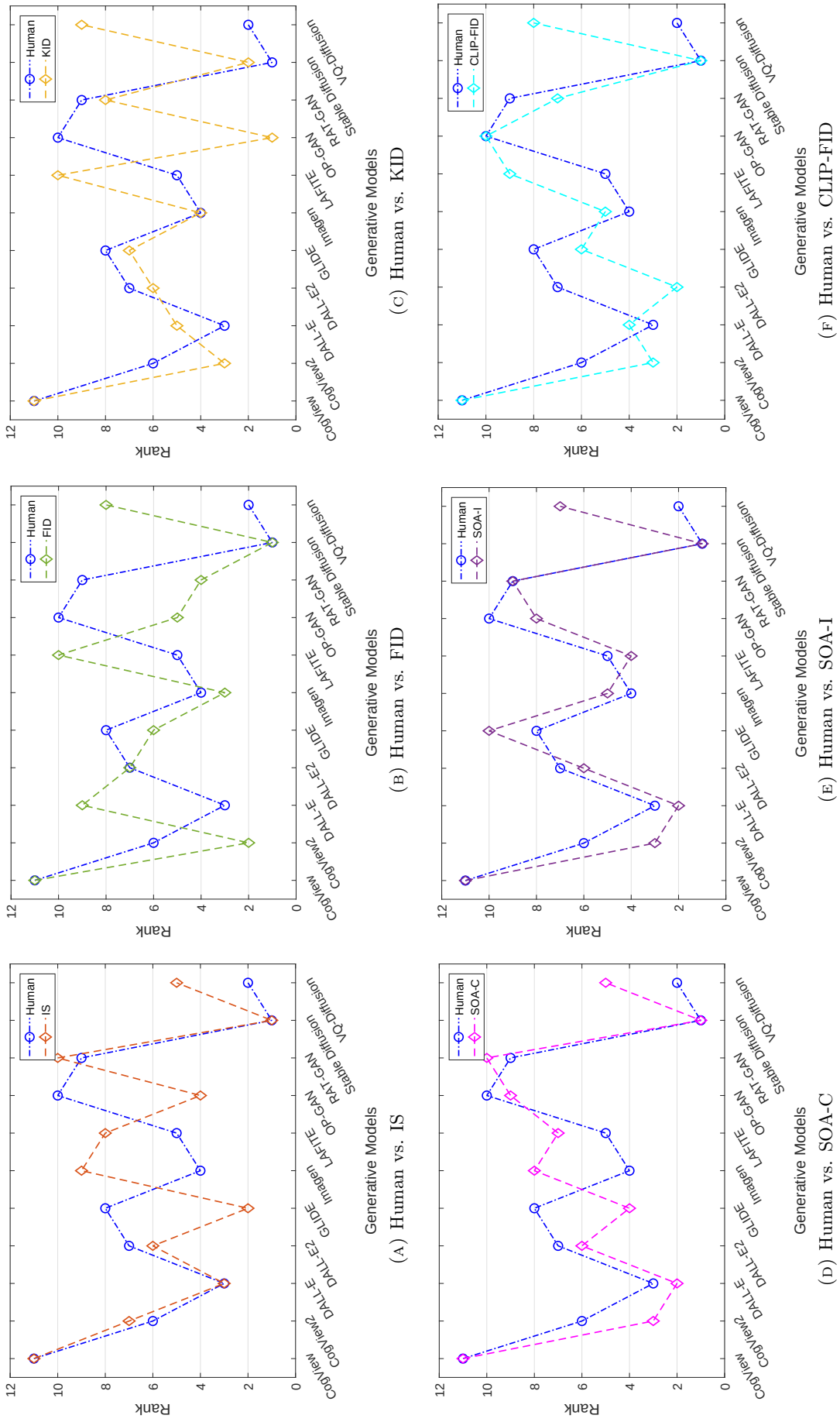


FIGURE 5.8. More visual results of the impact of each component in our LEICA. The generated images are from stable diffusion.



(A) Human vs. IS (B) Human vs. FID (C) Human vs. KID

(D) Human vs. SOA-C (E) Human vs. SOA-I (F) Human vs. CLIP-FID

FIGURE 5.9. Comparison between the rank of each generative model derived by human raters and that from various automatic evaluation metrics, including IS, FID, KID, SOA-C, SOA-I and CLIP-FID on the MS-COCO dataset. Rank-1 means the best model while rank-11 is the worst one.

Chapter 6

Conclusion

In this chapter, we provide a thorough overview of our significant contributions and outline prospective avenues for future work in the field of multi-modal generation, synergy, and evaluation.

In conclusion, multi-modal content generation stands at the forefront of contemporary AI research, offering innovative solutions across diverse modalities and domains such as natural language processing, computer vision, and healthcare. This thesis comprehensively explores its framework through the lenses of methods, synergies, and evaluations, aiming to address challenges in multi-modal generation and assess the practical effectiveness of these solutions.

In the methodologies section, we focus on image captioning, proposing the Discrete Mode Learning (DML) paradigm to combat mode collapse in generative models. DML significantly enhances linguistic variety, providing diverse and rich language patterns in generated captions. In the medical domain, our Single-for-Multiple (S4M) framework excels in automated radiology report generation, showcasing the versatility and precision of multi-modal models in healthcare scenarios.

Moving to the realm of cross-modal synergy, we introduce the Visual Voice Cloning (V2C) task, expanding traditional voice cloning to include emotional nuances. V2C incorporates vision, language, and voice modalities, with the goal of synthesizing voice from textual inputs while aligning it with emotions extracted from video data. This task finds applications in diverse fields, from entertainment, such as movie dubbing, to assistive technologies.

Finally, we propose the Likelihood-based tExt-to-Image evaluation with Credit Assignment (LEICA) metric for text-to-image synthesis. Overcoming the limitations of existing metrics, LEICA directly estimates the likelihood of generated images using a pre-trained model and incorporates a credit assignment strategy for both perceptual and semantic significance. This metric represents a significant advancement, offering a more accurate and efficient tool for evaluating multi-modal models and aligning closely with human perceptual judgments. Overall, this thesis contributes valuable insights and methodologies to advance the field of multi-modal content generation across various domains.

6.1 Limitations and Future Work

Based on the aforementioned research and findings for the methodologies, synergy, and evaluation of multi-modal generation, future work has the following possibilities.

Although the proposed DML can generate diverse captions based on different mode embeddings in the codebook, the more specific meanings for some of the modes may still be unclear. Thus, we seek to conduct detailed professional linguistic analyses to discern more specific meanings for the modes generated by the DML paradigm.

While our S4M exhibits superior performance compared to the baselines, it is important to acknowledge several limitations that could be addressed in future research. One of the limitations is the relatively small size of the dataset, even when data from multiple body parts are combined. This limited dataset size poses challenges for achieving optimal performance of the model. To overcome this limitation, in addition to a well-designed model architecture or learning method, it becomes necessary to explore data expansion methods such as data augmentation or synthesis. Besides, since we take this work as a promising start on multi-part report generation in radiology, we only focus on six specific body parts, which limits the coverage of anatomical regions, such as the brain. Additionally, for simplicity, we align the report lengths within a range of 30 to 60 tokens, disregarding reports that are too short or too long. For our future work, we plan to expand our scope by incorporating more body parts and including reports with more diverse lengths.

As for the V2C task, we may extend V2C-Net in two aspects in the future. First, to grasp the emotion from the video, we simply use the I3D model as the emotion encoder. However, it may not well disentangle emotion from character identity. To alleviate this issue, we may design an emotion-aware loss to capture more discriminative emotion features. Second, we integrate the multi-modal features (*i.e.*, text, audio, video) by a simple operation (*i.e.*, element-wise add), which may result in sub-optimal performance. Thus, a more promising model for feature fusion is necessary, *e.g.*, COOT [Ging et al., 2020] or VATT [Akbari et al., 2021].

For the evaluation metric, although LEICA is closely aligned with human judgments, sometimes it scores of generated images could be slightly higher than the ground-truth ones, *e.g.*, our score of ground-truth images is about 22 while samples from stable diffusion can reach 24. We hypothesize that the main reason is due to the mode collapse of the pre-trained generative model. It is infeasible to capture all occurred modes in real images, even with extremely large training data and model size. Therefore, the probability density of the generative model will inevitably concentrate on captured modes, making it possible that the likelihood of the captured mode is higher than that in the real images. To overcome this, as LEICA can be viewed as a general paradigm for text-to-image evaluation, we plan to update the pre-trained autoregressive text-to-image generation model with stronger models in the future.

6.2 Future Work in More Various Domains

The future of multi-modal research holds immense promise and potential for groundbreaking advancements across various domains. Here are some key aspects that could shape the future of multi-modal research:

- **Fine-grained Understanding:** Future research may focus on achieving a deeper and more fine-grained understanding of multi-modal data. This includes exploring techniques for capturing subtle nuances and relationships between different modalities, as well as developing models that can reason and make decisions based on complex multi-modal inputs.

- **Domain-specific Applications:** Multi-modal research will increasingly move towards domain-specific applications, such as healthcare, education, and robotics. Tailoring multi-modal systems to specific domains can lead to more impactful solutions that address real-world challenges and improve human-computer interaction.

- **Ethical and Social Implications:** As multi-modal systems become more pervasive in society, there will be a growing need to consider the ethical and social implications of their use. Future research will likely explore issues related to fairness, transparency, accountability, and privacy in multi-modal systems, as well as develop frameworks for responsible design and deployment.

Bibliography

https://github.com/kaustubh-sadekar/FunMirrors.

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):8717–8727, 2018.

Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*, 2021.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Eur. Conf. Comput. Vis.*, 2016.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6077–6086, 2018a.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3674–3683, 2018b.

Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *Int. Conf. Comput. Vis.*, 2019.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2425–2433, 2015.

Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *ICML*, pages 195–204, 2017.

Sercan Ö. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. In *Adv. Neural Inform. Process. Syst.*, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. Location-relative attention mechanisms for robust long-form speech synthesis. In *ICASSP*, pages 6194–6198, 2020.

Zhiwei Bi, Bing Cao, Wangmeng Zuo, and Qinghua Hu. Learning a prototype discriminator with rbf for multimodal image synthesis. *IEEE Trans. Image Process.*, pages 6664–6678, 2022.

Yanyao Bian, Changbin Chen, Yongguo Kang, and Zhenglin Pan. Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis. In *International Speech Communication Association*, 2019.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *Int. Conf. Learn. Represent.*, 2018.

Merlijn Blaauw, Jordi Bonada, and Ryunosuke Daido. Data efficient voice cloning for neural singing synthesis. In *ICASSP*, pages 6840–6844, 2019.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *Int. Conf. Learn. Represent.*, 2018.

Andrea Burns, Deniz Arsan, Sanjna Agrawal, Ranjitha Kumar, Kate Saenko, and Bryan A Plummer. A dataset for interactive vision-language navigation with unknown command feasibility. In *Eur. Conf. Comput. Vis.*, pages 312–328, 2022.

Jiezhang Cao, Yong Guo, Qingyao Wu, Chunhua Shen, and Mingkui Tan. Adversarial learning with local coordinate coding. *ICML*, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6299–6308, 2017.

Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baochang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, and Yan Wang. Variational structured semantic inference for diverse image captioning. *Adv. Neural Inform. Process. Syst.*, 2019a.

Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. Human-like controllable image captioning with verb-specific semantic roles. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021a.

Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Adaspeech: Adaptive text to speech for custom voice. *Int. Conf. Learn. Represent.*, 2021b.

Qi Chen, Qi Wu, Jian Chen, Qingyao Wu, Anton van den Hengel, and Mingkui Tan. Scripted video generation with a bottom-up generative adversarial network. *IEEE Trans. Image Process.*, pages 7454–7467, 2020a.

Qi Chen, Qi Wu, Rui Tang, Yuhan Wang, Shuai Wang, and Mingkui Tan. Intelligent home 3d: Automatic 3d-house design from linguistic descriptions only. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12625–12634, 2020b.

Qi Chen, Chaorui Deng, and Qi Wu. Learning distinct and representative modes for image captioning. *Adv. Neural Inform. Process. Syst.*, 2022a.

Qi Chen, Yutong Xie, Biao Wu, Minh-Son To, James Ang, and Qi Wu. S4m: Generating radiology reports by a single model for multiple body parts. *arXiv preprint arXiv:2305.16685*, 2023a.

Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9962–9971, 2020c.

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10638–10647, 2020d.

Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. "factual"or"emotional": Stylized image captioning with adaptive learning and attention. In *Eur. Conf. Comput. Vis.*, pages 519–535, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020e.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *Int. Conf. Learn. Represent.*, 2023b.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Eur. Conf. Comput. Vis.*, pages 104–120, 2020f.

Yutian Chen, Yannis M. Assael, Brendan Shillingford, D. Budden, Scott E. Reed, H. Zen, Q. Wang, Luis C. Cobo, Andrew Trask, Ben Laurie, Çaglar Gülçehre, Aäron van den Oord, Oriol Vinyals, and N. D. Freitas. Sample efficient adaptive text-to-speech. In *Int. Conf. Learn. Represent.*, 2019b.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *EMNLP*, 2020g.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. *ACL-IJCNLP*, 2022b.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3444–3453, 2017.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

Glen Coppersmith and Erin Kelly. Dynamic wordclouds and vennclouds for exploratory data analysis. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 22–29, 2014.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10578–10587, 2020.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Int. Conf. Comput. Vis.*, 2017.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, pages 304–310, 2016.

Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning. In *Eur. Conf. Comput. Vis.*, 2020a.

Chaorui Deng, Qi Wu, Qingyao Wu, Fan Lyu, Fuyuan Hu, and Mingkui Tan. Visual grounding via accumulated attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020b.

Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, and Qi Wu. Prompt switch: Efficient clip adaptation for text-video retrieval. In *Int. Conf. Comput. Vis.*, pages 15648–15658, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009.

Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Adv. Neural Inform. Process. Syst.*, pages 19822–19835, 2021.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Adv. Neural Inform. Process. Syst.*, 2022.

Brian Dolhansky and Cristian Canton Ferrer. Eye in-painting with exemplar generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *AAAI*, volume 32, 2018.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 4065–4080, 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12873–12883, 2021.

Andrew Gibiansky, Sercan Ö. Arik, G. Diamos, J. Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In *Adv. Neural Inform. Process. Syst.*, 2017.

Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *Adv. Neural Inform. Process. Syst.*, 2020.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15180–15190, 2023.

Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. In *ICML*, pages 7616–7633, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Adv. Neural Inform. Process. Syst.*, 2014.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124, 2013.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10696–10706, 2022.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Adv. Neural Inform. Process. Syst.*, 2017.

Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian He, Jie Jiang, and Hanqing Lu. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. *IJCAI*, 2020a.

Yong Guo, Qi Chen, Jian Chen, Junzhou Huang, Yanwu Xu, Jiezhang Cao, Peilin Zhao, and Mingkui Tan. Dual reconstruction nets for image super-resolution with gradient sensitive loss. *arXiv preprint arXiv:1809.07099*, 2018.

Yong Guo, Qi Chen, Jian Chen, Qingyao Wu, Qinfeng Shi, and Mingkui Tan. Auto-embedding generative adversarial networks for high resolution image synthesis. *IEEE Trans. Multimedia*, 2019.

Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5407–5416, 2020b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *EMNLP*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inform. Process. Syst.*, 2017.

Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

Keikichi Hirose and Jianhua Tao. *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*. Springer, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst.*, 33:6840–6851, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Wei-Ning Hsu, Y. Zhang, Ron J. Weiss, H. Zen, Y. Wu, Yuxuan Wang, Yuan Cao, Y. Jia, Z. Chen, Jonathan Shen, P. Nguyen, and Ruoming Pang. Hierarchical generative modeling for controllable speech synthesis. In *Int. Conf. Learn. Represent.*, 2019.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Int. Conf. Comput. Vis.*, pages 4634–4643, 2019.

Rui Huang, Shu Zhang, Tianyu Li, Ran He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *Int. Conf. Comput. Vis.*, 2017.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, pages 590–597, 2019.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, pages 203–211, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1125–1134, 2017.

Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

Ye Jia, Y. Zhang, Ron J. Weiss, Q. Wang, Jonathan Shen, Fei Ren, Z. Chen, P. Nguyen, Ruoming Pang, I. Lopez-Moreno, and Y. Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Adv. Neural Inform. Process. Syst.*, 2018.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *ACL*, 2018.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *ICML*, pages 2410–2419, 2018.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3128–3137, 2015.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *Int. Conf. Learn. Represent.*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and M. Welling. Auto-encoding variational bayes. *Int. Conf. Learn. Represent.*, 2014.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Adv. Neural Inform. Process. Syst.*, 2020.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. Conf. Comput. Vis.*, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.*, 25, 2012.

Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Adv. Neural Inform. Process. Syst.*, 2019.

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *Int. Conf. Learn. Represent.*, 2023.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, pages 436–444, 2015.

Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In *Int. Conf. Learn. Represent.*, 2023.

Dianqi Li, Qiuyuan Huang, Xiaodong He, Lei Zhang, and Ming-Ting Sun. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*, 2018a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9215–9223, 2018b.

Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. Cross-modal clinical graph transformer for ophthalmic report generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20656–20665, 2022b.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023b.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *AAAI*, pages 6706–6713, 2019.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Eur. Conf. Comput. Vis.*, 2020a.

Yang Li, Jiacong He, Xin Zhou, Yuan Zhang, and Jason Baldridge. Mapping natural language instructions to mobile ui action sequences. *ACL*, 2020b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13753–13762, 2021a.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. Auto-encoding knowledge graph for unsupervised medical report generation. *Adv. Neural Inform. Process. Syst.*, pages 16266–16279, 2021b.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Adv. Neural Inform. Process. Syst.*, 32, 2019.

Ruotian Luo. https://github.com/ruotianluo/self-critical.pytorch.

Ruotian Luo and Gregory Shakhnarovich. Analysis of diversity-accuracy tradeoff in image captioning. *arXiv preprint arXiv:2002.11848*, 2020.

Xuewei Ma, Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Yuexian Zou, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. *ACL-IJCNLP*, 2021.

Shweta Mahajan and Stefan Roth. Diverse image captioning with context-object split latent spaces. *Adv. Neural Inform. Process. Syst.*, 2020.

Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. *Int. Conf. Learn. Represent.*, 2020.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Int. Conf. Learn. Represent.*, 2015.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Int. Conf. Comput. Vis.*, pages 2794–2802, 2017.

Alexander Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *AAAI*, 2016.

Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502, 2017.

Sondess Missaouib, Shauna Concannonc, Liam Maloneyb, James Alfred Walker, et al. Interactive storytelling for children: A case-study of design and development considerations for ethical conversational ai. *arXiv preprint arXiv:2107.13076*, 2021.

Meinard Müller. Dynamic time warping. *Information Retrieval for Music and Motion*, pages 69–84, 2007.

Eliya Nachmani, Adam Polyak, Yaniv Taigman, and Lior Wolf. Fitting new speakers based on a short untranscribed sample. In *ICML*, pages 3683–3691, 2018.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5034. URL https://www.aclweb.org/anthology/W19-5034.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *ICML*, 2022.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10971–10980, 2020.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, 2015.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *AAAI*, volume 36, pages 2062–2070, 2022.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ö. Arik, A. Kannan, Sharan Narang, Jonathan Raiman, and J. Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *Int. Conf. Learn. Represent.*, 2018.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Int. Conf. Multimedia*, pages 484–492, 2020.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9982–9991, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *Int. Conf. Learn. Represent.*, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Adv. Neural Inform. Process. Syst.*, 32, 2019.

ITUT Rec. P. 800: Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva*, 1996.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019a.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inform. Process. Syst.*, 28, 2015.

Yi Ren, Yangjun Ruan, X. Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and T. Liu. Fastspeech: Fast, robust and controllable text to speech. In *Adv. Neural Inform. Process. Syst.*, 2019b.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *Int. Conf. Learn. Represent.*, 2021.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7008–7024, 2017.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3202–3212, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022.

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inform. Process. Syst.*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Adv. Neural Inform. Process. Syst.*, 2016.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2039–2048, 2019.

Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Int. Conf. Comput. Vis.*, 2017.

Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging image captioning via personality. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12516–12526, 2019.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *ICML*, pages 4693–4702, 2018.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6447–6456, 2017.

Daisy Stanton, Yuxuan Wang, and RJ Skerry-Ryan. Predicting expressive speaking style from text in end-to-end speech synthesis. In *Spoken Language Technology Workshop*, pages 595–602, 2018.

Liangtai Sun, Xingyu Chen, Lu Chen, Tianle Dai, Zichen Zhu, and Kai Yu. Meta-gui: Towards multi-modal conversational agents on mobile gui. *EMNLP*, 2022.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

Antti Santeri Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, Martti Vainio, et al. Wavelets for intonation modeling in hmm speech synthesis. In *ISCA Workshop on Speech Synthesis*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2818–2826, 2016.

Yaniv Taigman, Lior Wolf, A. Polyak, and Eliya Nachmani. Voiceloop: Voice fitting and synthesis via a phonological loop. In *Int. Conf. Learn. Represent.*, 2018.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *EMNLP*, 2019.

Hongchen Tan, Xiuping Liu, Meng Liu, Baocai Yin, and Xin Li. Kt-gan: Knowledge-transfer generative adversarial network for text-to-image synthesis. *IEEE Trans. Image Process.*, pages 1275–1290, 2020.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4631–4640, 2016.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *Int. Conf. Learn. Represent.*, 2016.

Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, page 10, 2006.

Rafael Valle, Jason Li, R. Prenger, and Bryan Catanzaro. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *ICASSP*, pages 6189–6193, 2020.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, pages 1747–1756, 2016.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Adv. Neural Inform. Process. Syst.*, 2017.

Sean Vasquez and Mike Lewis. Melnet: A generative model for audio in the frequency domain. *arXiv preprint arXiv:1906.01083*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4566–4575, 2015.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *AAAI*, 2018.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3156–3164, 2015.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *ICASSP*, pages 4879–4883, 2018.

Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. *Adv. Neural Inform. Process. Syst.*, 2017a.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340, 2022a.

Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

Qingzhong Wang, Jia Wan, and Antoni B Chan. On diversity in image captioning: Metrics and methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, 2018a.

Yuxuan Wang, R. Skerry-Ryan, Daisy Stanton, Y. Wu, Ron J. Weiss, Navdeep Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, R. Clark, and R. Saurous. Tacotron: Towards end-to-end speech synthesis. In *International Speech Communication Association*, 2017b.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *ICML*, pages 5180–5189, 2018b.

Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. A self-boosting framework for automated radiographic report generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2433–2442, 2021.

Zhanyu Wang, Hongwei Han, Lei Wang, Xiu Li, and Luping Zhou. Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Transactions on Medical Imaging*, pages 2803–2813, 2022b.

Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. A medical semantic-assisted transformer for radiographic report generation. In *MICCAI*, pages 655–664, 2022c.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022d.

M. Whitehill, S. Ma, Daniel McDuff, and Yale Song. Multi-reference neural tts stylization with adversarial cycle consistency. In *International Speech Communication Association*, 2020.

Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, pages 4–15, 2020.

Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. Discriminative multi-modality speech recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14433–14442, 2020.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1316–1324, 2018.

Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *MICCAI*, pages 457–466, 2018.

Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2019.

Sixing Yan, William K Cheung, Keith Chiu, Terence M Tong, Ka Chun Cheung, and Simon See. Attributed abnormality graph embedding for clinically accurate x-ray report generation. *IEEE Transactions on Medical Imaging*, 2023.

Yuzi Yan, Xu Tan, Bohan Li, Tao Qin, Sheng Zhao, Yuan Shen, and Tie-Yan Liu. Adaspeech 2: Adaptive text to speech with untranscribed data. In *ICASSP*, pages 6613–6617, 2021.

Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, page 102510, 2022.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

Yanhua Yang, Lei Wang, De Xie, Cheng Deng, and Dacheng Tao. Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis. *IEEE Trans. Image Process.*, pages 2798–2809, 2021.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Eur. Conf. Comput. Vis.*, 2018.

Senmao Ye, Fei Liu, and Minkui Tan. Recurrent affine transformation for text-to-image synthesis. *arXiv preprint arXiv:2204.10482*, 2022.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018a.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3165–3173, 2017.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Eur. Conf. Comput. Vis.*, pages 471–487, 2018b.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *MICCAI*, pages 721–729, 2019.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Adv. Neural Inform. Process. Syst.*, pages 27263–27277, 2021.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Int. Conf. Comput. Vis.*, pages 5907–5915, 2017.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 833–842, 2021a.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1195–1204, 2021b.

Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Wei Liu, and Hongdong Li. Adversarial spatio-temporal learning for video deblurring. *IEEE Trans. Image Process.*, 2019.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, pages 3836–3847, 2023.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *AAAI*, pages 12910–12917, 2020.

Fang Zhao, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE Trans. Image Process.*, 2018.

Chunting Zhou, Jiatao Gu, and Graham Neubig. Understanding knowledge distillation in non-autoregressive machine translation. In *Int. Conf. Learn. Represent.*, 2019.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049, 2020.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17907–17917, 2022.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5802–5810, 2019.