UTS UNIVERSITY OF TECHNOLOGY SYDNEY

# From Sound to Sign: Deep Learning for Audio Style Adaptation and Multi-modal Sign Language Analysis

**by Jian Ma**

Thesis submitted in fulfilment of the requirements for the degree of

**Doctor of Philosophy**

under the supervision of Yi Yang

University of Technology Sydney
Faculty of Engineering and Information Technology

August 2024

# CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Jian Ma*, declare that this thesis is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy*, in the *Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

DATE: 14th August, 2024

# ACKNOWLEDGMENTS

First and foremost, I extend my sincere gratitude to my principal supervisor, Prof. Yi Yang and Feng Zheng, whose expertise, understanding, and patience significantly enhanced my graduate experience. I am grateful for his generosity with his time and his insightful guidance throughout my research.

A special thanks to my co-supervisor, Prof. Wenguan Wang, whose enthusiasm and innovative thinking inspire me greatly. Without his knowledge and perspective, this thesis would not have been the same. His presence is a beacon of hope and strength, reminding me that perseverance shared is hardship halved.

I must acknowledge my peers in ReLER Group at UTS for their support, collaboration, and shared wisdom throughout my studies. The environment here has been both a catalyst for ideas and a continual source of friendship and motivation.

I am particularly indebted to my classmate, Jinliang Liu, whose encouragement and companionship were invaluable during my most challenging times. His unwavering support and thoughtful advice not only helped me overcome numerous obstacles but also made this journey significantly more pleasant and meaningful.

I am grateful for the financial support provided by the Southern University of Science and Technology, which enabled me to concentrate fully on my research.

Lastly, I wish to express my heartfelt gratitude to my family for their unconditional understanding and boundless love throughout the duration of my studies.

# ABSTRACT

The growing demand for accessible communication tools for the deaf community highlights the need to bridge auditory and visual languages. However, the inherent disparity between spoken and sign languages poses substantial challenges, making their conversion a quintessential problem in multimodal learning. Spoken language varies in patterns, accent, and intonation, requiring translation systems to adapt to diverse speech styles. Sign language similarly incorporates multimodal elements like gestures, facial expressions, and body movements, necessitating accurate distinction of actions and conveying of complex contexts. This study proposes an efficient multimodal learning framework to tackle bidirectional translation challenges between the two languages. In particular, we integrate audio style adaptation and multimodal sign language analysis within a unified system. For audio style adaptation, we leverage diffusion models and mutual learning mechanisms to convert raw audio into various target styles, minimizing the need for paired audio data. For multimodal sign language analysis, we implement sign language production and translation using sequence diffusion models and large language models, respectively. Our system is capable of adapting to various audio styles and harnessing the extensive prior knowledge of large models to preserve the original semantics in terms of both integrity and accuracy.

# LIST OF PUBLICATIONS

**RELATED TO THE THESIS :**

1. MS2SL: Multimodal Spoken Data-Driven Continuous Sign Language Production (ACL 2024)
2. Mutual Learning for Acoustic Matching and Dereverberation via Visual Scene-driven Diffusion (ECCV 2024)
3. Hybrid Model Collaboration for Sign Language Translation with VQ-VAE and RAG-enhanced LLMs (Under review)
4. Subband-based Generative Adversarial Network for Non-parallel Many-to-many Voice Conversion (Ready for submission)

# TABLE OF CONTENTS

# List of Figures

# List of Tables

CHAPTER 1

# INTRODUCTION

Globally, millions of hearing-impaired individuals encounter notable communication barriers when interacting with those who are hearing-abled, affecting areas such as education, employment, and equitable access to information. Sign language [16, 81, 97, 136, 222, 223], an important visual language, plays a crucial role in communication for the hearing-impaired due to its rich content and intuitive expression. Developing a technology that can achieve the conversion between spoken language and sign language is of great significance to improving the quality of life of the hearing-impaired and holds profound theoretical and practical value. Furthermore, the effective association of audio with sign language semantics can facilitate barrier-free communication between hearing-impaired and hearing communities.

Although advancements in deep learning have somewhat mitigated communication hindrances for the hearing-impaired with technologies like speech-to-text conversion [25, 113, 128, 130, 166, 190, 195, 208, 262], most of these technologies focus on converting sound to text, rarely addressing conversion to sign language—the more intrinsic and intuitive mode of communication for the hearing-impaired. In real-life applications, these technologies typically fail to fulfill their requirements for accurate interpretation of body language, particularly in conveying timbre style and space environment characteristics accurately. With the swift development of artificial intelligence, multimodal data processing has become a prominent research area. However, converting between speech and sign language still presents numerous technical challenges. First, the diverse styles of audio signals require the system to have high adaptability to recognize and emulate different speech styles. Second, the multi-

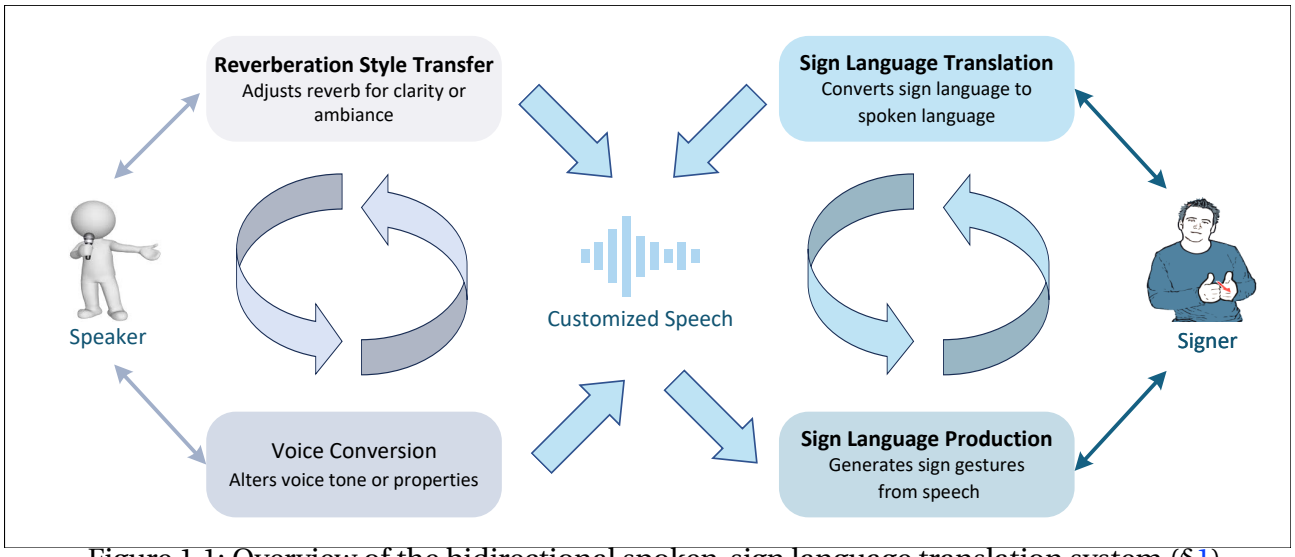


Figure 1.1: Overview of the bidirectional spoken-sign language translation system (§1).

modal nature of sign language, involving gestures, facial expressions, and body movements, demands that the system not only distinguish actions but also accurately convey complex contexts. This research is dedicated to leveraging deep learning to develop an efficient spoken-sign language translation system to provide technical support for communication for the hearing-impaired. In particular, this study explores the conversion between spoken language and sign language from two perspectives: audio style adaptation and multimodal sign language analysis. Audio style adaptation involves voice conversion and reverberation style transfer to provide customized speech timbres and styles. Multimodal sign language analysis includes generating and translating sign language to enhance communication between hearing-impaired and hearing individuals. Therefore, our investigation focuses on the following four aspects:

- Voice Conversion: Modifying certain characteristics of original voices, such as pitch, timbre, and gender, while maintaining clarity and accuracy of the language content.
- Reverberation Style Transfer: Altering the audio's auditory style by adding or removing reverberation, making audio processing more diverse, detailed, and comprehensible.
- Multimodal Sign Language Production: Translating spoken language into sign language animations or videos to improve accessibility for deaf individuals.
- Sign Language Translation: Converting complex sign language motions into corresponding spoken language content by recognizing and parsing sign gestures.

Our system integrates the above four key technologies to facilitate barrier-free communication with the hearing impaired (Fig. 1.1). These four technologies can function independently

or together as an integrated closed-loop sign language communication system. Firstly, voice conversion technology adjusts voice properties to meet diverse audience and recording requirements. Secondly, reverberation style transfer enhances auditory experiences by optimizing reverb styles for clarity or spatial enhancement. Thirdly, multimodal sign language production translates spoken content into sign language, making it more accessible to the deaf community. Lastly, the sign language translation feature instantly converts received sign language inputs into spoken language and text, ensuring smooth two-way communication.

To validate the effectiveness of the proposed method, extensive experiments are conducted across multiple datasets in targeted fields. The results demonstrate marked improvements in both the accuracy of sign language production and translation, and the naturalness of audio style transformation, over existing leading technologies.

Audio style adaptation and multimodal sign language analysis share some similar challenges and contribute complementary functionalities within the system. The ultimate goal of this study is to enable seamless communication between hearing and non-hearing individuals, ensuring both signers and non-signers can comprehend sentence meanings. Audio style adaptation can provide clear, highly comprehensible, and customizable audio for hearing users, while multimodal sign language analysis converts spoken language into sign language and vice versa for sign language users. This allows both groups to easily access information and personalize content according to their preferences. From a technical perspective, both fields encounter challenges including data scarcity, cross-modal feature alignment, and maintaining semantic consistency. Unifying them within a single system fosters collaboration and enhances synergistic solutions.

Through this study, we not only provide new technological avenues for the communication of hearing-impaired individuals but also demonstrate the potent application of deep learning in processing multimodal data. The development of this technology can greatly improve communication conditions for the hearing impaired, enhance their social participation. For instance, in smart education, this technology can be used to develop educational software for hearing-impaired children, teaching them through sign language to better understand and learn new information. In telemedicine, speech-to-sign language conversion enables doctors to communicate more effectively with hearing-impaired patients, providing more precise medical services. Future research will focus on optimizing deep learning models and enhancing system adaptability and accuracy. Additionally, considering regional differences in sign language, we plan to explore a customizable learning mechanism that adjusts to various regional sign languages. We hope to broaden the application of this technology to serve a larger group of individuals in need.

## 1.1 Audio Style Adaptation

**Voice Conversion** refers to the process of changing one voice style to another without altering semantic content [117, 249, 254, 256, 264]. In pursuit of high fidelity and accurate timbre targeting in voice conversion, we introduce a Subband-based Generative Adversarial Network (SGAN-VC). SGAN-VC exploits spatial features between different subbands and transforms the content of each subband independently, capturing subtle variations in sound (Chapter 3). Moreover, we utilize an attention mechanism to understand the complex relationships between the target voice style and the original audio content, thus improving the accuracy and naturalness of the produced voice. This technology is applicable in areas such as personalized voice adjustment and voice synthesis, making personalized voice services possible. Additionally, it supports multilingual conversion, enhancing convenience in international communications.

**Reverberation Style Transfer.** Reverb addition and dereverberation are crucial in audio style transfer, where reverb creates richer and more engaging soundscapes, ideal for music and performance art by emulating natural echoes of spaces [216, 225, 240, 258]. Conversely, dereverberation removes undesirable reflections and echoes through precise digital processing, enhancing clarity and precision in voice communications and professional recordings [79, 179, 181, 271]. These technologies significantly enhance the adaptability and professionalism of audio content. Existing approaches worsen the situation by independently tackling each task, neglecting the reciprocal nature of these tasks and limiting the use of vast unlabeled data. Additionally, these methods depend on hard-to-obtain paired training data, preventing the utilization of large volumes of unpaired data. This work presents MVSD (Chapter 4), a mutual learning framework based on diffusion models that symmetrically considers both tasks, leveraging their reciprocal relationship to enhance learning from each and overcome data scarcity. Additionally, we employ diffusion models as base condition converters to avoid the training instabilities and over-smoothing flaws of traditional GAN architectures. The effective integration of the two tasks allows for more flexible audio style adaptation across a range of different scenarios.

## 1.2 Multimodal Sign Language Analysis

**Multimodal Sign Language Production** integrates computer vision, machine learning, and image processing to automate the conversion from speech or text to sign language [96, 227, 228]. This enables the hearing impaired to more easily access information and better

integrate into social and work environments. Despite significant advances in sign language understanding, there is still no viable solution for directly generating gesture sequences from complete spoken content, such as text or speech. In this study, we introduce a unified continuous sign language generation framework MS2SL to simplify communication between sign and non-sign language users (Chapter 5). Specifically, sequence diffusion models utilize embeddings extracted from text or speech to progressively generate sign predictions. Furthermore, by establishing a joint embedding space for text, audio, and gestures, we link these modalities and harness their semantic consistency to inform model training. This embedding consistency learning strategy minimizes dependence on gesture triples and ensures continuous model optimization even in the absence of the audio modality. Sign language production not only aids the deaf community in better understanding and communicating but also allows non-sign language users to learn and understand sign language through visualization. Applications include education, social media, and customer service, providing the hearing impaired with more platforms for information access and communication, greatly enhancing content accessibility and interactivity.

**Sign Language Translation** utilizes computer vision and deep learning models to capture and interpret sign language gestures, converting them into spoken text [282, 298]. Beyond mere gestures, sign language also conveys information through expressive facial features and body language. This work proposes an innovative framework, VRG-SLT, to translate sign language into spoken language (Chapter 6). VRG-SLT employs a hierarchical VQ-VAE to transform continuous gesture sequences into discrete representations (termed 'sign codes'), which are then aligned with text using a fine-tuned pretrained language model. Furthermore, we employ RAG to improve and polish the initial translation of language model, producing semantically richer and more precise text.

## 1.3 Thesis Organization

This dissertation presents a comprehensive system that integrates voice conversion, audio style transfer, sign language production, and sign language translation to facilitate barrier-free communication between spoken and sign languages. This thesis is organized as follows:

**Chapter 1: Introduction**

This chapter provides an in-depth look at the research background, motivation, significance, and objectives of the study. It comprehensively outlines the key research questions, main contributions, and the detailed structure of the thesis.

**Chapter 2: Literature Survey**

An exhaustive review of research in areas such as condition-guided generation, consistency learning, audio style adaptation, sign language translation and production. It establishes a theoretical and technical foundation for the methodologies and experimental designs that are discussed in subsequent chapters, and examines the integration of these technologies within the context of this research.

**Chapter 3: Subband-based GAN for Voice Conversion**

This chapter delves into the proposed voice conversion framework, SGAN-VC, covering the model structure, experimental design, datasets used, and evaluation metrics to thoroughly assess the study's effectiveness and potential applications.

**Chapter 4: Mutual Learning for Acoustic Matching and Dereverberation**

In this chapter, we introduce MVSD to explore the tasks of visual-acoustic matching and dereverberation, utilizing the duality between these tasks to reduce the dependency on paired data. The focus is on developing techniques that leverage visual cues to enhance acoustic clarity, thereby facilitating more effective audio style transfer. Through detailed experiments and analysis, this chapter demonstrates how the interplay between visual and acoustic elements can be optimized to achieve superior audio style transformation.

**Chapter 5: Multimodal Spoken Data-Driven Sign Language Production**

This chapter concentrates on closing communication gaps by discussing the use of diffusion models to develop MS2SL, a multimodal method for generating sign language. MS2SL explores the transformation of spoken language (both audio and text) into sign language and provides an in-depth look at the model's capabilities in interpreting and generating intricate sign language sequences.

**Chapter 6: Hybrid Model Collaboration for Sign Language Translation**

This chapter presents the application of large models in sign language translation. We employ a hierarchical VQ-VAE to transform sign language movements into discrete sign codes and aligning them with text tokens in training. In this chapter, we explain how the proposed VLT-RAG method translates sign fragments into precise, coherent text descriptions.

**Chapter 7: Conclusion and Future Work**

This chapter summarizes the key findings and contributions of the thesis, reflecting on the integration and effectiveness of the developed technologies. Moreover, it outlines potential areas for future research that could further refine and expand the technologies discussed, suggesting specific improvements and new areas of application.

CHAPTER 2

# LITERATURE SURVEY

This survey explores the technologies, development trajectories, and unique aspects of our system in the fields of condition-guided generation, consistency learning, audio style adaptation, sign language translation and production. The convergence of these technologies presents a golden opportunity to bridge the communication gap between diverse auditory and visual language modalities, catering to both hearing individuals and those with hearing impairments.

## 2.1 Condition-guided Generation

Condition-guided generation is an AI technique that generates data—such as text, images, and audio—under specific conditions or parameters that dictate certain characteristics or properties of the output. This technology encompasses methods like Conditional Generative Adversarial Networks (GANs) [75, 109, 114, 162] and Variational Autoencoders (VAEs) [115, 215, 259], which integrate additional condition information to guide the generation process, as well as conditional sequence models such as LSTMs [90] or Transformers [260] that use conditions as part of their input for tasks like text generation and speech synthesis. Recently, diffusion models [89, 244] have also gained prominence, utilizing a reversed diffusion process conditioned on parameters to create high-fidelity images and audio. Initially used for simpler tasks such as text templates and image captioning, these technologies have evolved with the rise of deep learning, enhancing their capability to handle more complex data and detailed conditions.

### 2.1.1 Vector Quantized Variational AutoEncoder

Vector Quantized Variational AutoEncoder (VQ-VAE), a deep learning tool, excels in compressing and reconstructing high-fidelity images, videos, and audio by mapping them into a lower-dimensional latent space [259]. Its recent applications extend to realistic image and video generation with GANs, diverse vocal representations in speech processing, and feature extraction in unsupervised learning [26, 59, 127, 297]. Research advancements focus on enhancing latent space expressiveness and reconstruction quality, alongside expansion into multimodal learning and natural language processing, particularly in text generation and semantic analysis. In text-to-motion generation [100, 288], VQ-VAE delivers compelling outcomes in semantic coherence and motion precision. However, VQ-VAE attains comparable accuracy with discretized latent variables to those with continuous ones, but also suffers from the typical AutoEncoder issue of blurry images. The following VQ-VAE-2 [215] achieves higher-quality image generation through a multi-level hierarchical structure, while maintaining diversity and preventing mode collapse. In Chapter 6, inspired by VQ-VAE-2, we carefully craft a Sign-tokenizer that employs top and bottom level vectorizers to model global human body and hand features, thereby capturing more detailed and comprehensive motion trajectories. To our knowledge, this is the pioneering effort to utilize VQVAE specifically for the realm of sign language translation.

### 2.1.2 Generative Adversarial Networks

GANs are the representative methods for generation, which are widely employed in many areas, *e.g.*, computer vision [28, 85, 142], natural language processing [60, 61, 141, 284], *etc.* GANs are pioneered in the field of image generation [75], which advances by manipulating the input noise to achieve the desired result. The classic GAN framework consists of two distinct models: a generative model (the generator) that captures the data distribution, and a discriminative model (the discriminator) that estimates the probability that a sample came from the training data rather than the generator. The generator's goal is to produce data that are indistinguishable from genuine data, attempting to fool the discriminator. On the other hand, the discriminator aims to identify whether a given data instance is real or counterfeit. Through their adversarial process, both networks incrementally improve their methods: the generator learns to produce more accurate imitations of real data, while the discriminator becomes better at detecting fakes. Huang *et al.* [95] point out that the adaptive instance normalization design adequately addresses the demands of style transfer through the exchange of mean and variance in the normalization layer between the source

and reference samples. For voice style transfer, recent works [241, 242, 264] find that GANs based on the melspectrogram also achieve impressive results. Although diffusion models are increasingly favored for their high fidelity and training stability, their inherent speed limitations make them less suitable for tasks requiring real-time performance. In Chapter 3, considering the time-sensitive requirements of voice conversion, we adopt GAN as our foundational architecture in SGAN-VC.

### 2.1.3 Diffusion Model

In recent years, there have been significant advancements in the field of conditional generation [94, 207, 239, 268]. Diffusion models have demonstrated impressive results in various generative tasks due to their superior visual quality and training stability [12, 40, 53, 89, 121, 144, 152, 169, 182, 219]. The diffusion probability model [244] is based on a Markov chain, proceeding through finite steps in two opposing directions: one transition moves from the data distribution to noise, and the other transitions back from noise to the data distribution. Ho *et al.* [89] introduce the variational lower bound objective, which is subsequently improved in [182] to obtain higher log-likelihood scores. By combining a language model with a conditional diffusion model, [219] is able to generate high-quality images using textual features. Beyond image generation, diffusion models also perform well in generating sequence data [274, 285]. In recent years, some work has begun to apply diffusion models to SLP. By iteratively updating information, diffusion models can gradually infer the distribution of subsequent data, thereby providing more accurate and coherent results. Ham2Pose [8] leverages diffusion to animate HamNoSys, a lexicon of sign symbols, into sign keypoint sequences. Though impressive, Ham2Pose can only produce videos with a single sign symbol, falling short in conveying sentences with complete semantics. DiffWave [121], a flexible diffusion model, specializes in synthesizing high-quality audios for various tasks. In Chapter 4, we regard audio spectrograms as images and elegantly employ two diffusion-based generators for controllable reverberation style transfer. Given the superiority of diffusion models in generating sequential data, Chapter 5 employs a sequential diffusion model as a sign language action generator in MS2SL framework.

### 2.1.4 Large Language Models

In the field of artificial intelligence, large models, particularly Large Language Models (LLMs), have transformed natural language processing by enabling the generation of human-like text [17, 52, 54, 125, 145, 197, 279, 291]. LLMs, such as OpenAI's GPT series, Google's BERT,

and T5, have demonstrated remarkable capabilities in addressing complex language understanding and generation tasks, achieved through pre-training on extensive datasets. These models are particularly valued for their ability to produce semantically coherent and contextually appropriate outputs [45, 131, 173, 250].

The term "LLM" generally refers to models with a large number of parameters, often in the range of tens of billions. However, the threshold for classifying a model as an LLM is not strictly defined and may vary depending on the context. For example, while models like BERT and T5 are often categorized as pre-trained language models due to their training methodology, their substantial parameter sizes and significant impact on large-scale NLP tasks also qualify them as LLMs. This dual classification highlights the intersection of pre-training methodologies and model scale in defining LLMs.

A distinctive feature of LLMs is their ability to generate outputs conditioned on input text, such as prompts or questions. These inputs guide the model to produce relevant and contextually aligned responses, making LLMs versatile tools for a wide range of natural language processing applications.

These LLMs typically utilize the Transformer architecture [260], enhancing their ability to learn efficiently from large data volumes. For example, GPT-3, trained on billions of words, demonstrates the capability to produce high-quality textual output through prompt tuning alone, without specific task-oriented training. Their applications extend across various fields including natural language understanding, text generation, sentiment analysis, machine translation, and automated summarization. Current research also investigates their use in multimodal tasks like the combined generation and understanding of text and images. The T5 model [212] adopts a unified text-to-text approach, enabling it to manage diverse NLP tasks by transforming them all into text generation problems. Trained on multiple language corpora, T5 possesses rich prior knowledge and powerful cross-lingual learning skills. FLAN-T5 [44] boosts multi-task capabilities by training with natural language and refining its response to commands. For the first time, VRG-SLT integrates extensive prior knowledge from FLAN-T5 and collaborates with the Sign-tokenizer, offering new prospects for cross-lingual alignment in whole-sentence sign-to-text translation (Chapter 6).

### 2.1.5 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a language enhancement technique that merges information retrieval with generation models [132]. RAG is developed to address the limitations of traditional language models that generate responses solely based on a fixed, pre-trained dataset. It introduces a dynamic component to generation process by retrieving

external, relevant documents or data that can provide additional context or factual details. In particular, RAG operates by first pulling relevant information from an extensive knowledge base and merge it with a generation model to produce precise and comprehensive text output [11, 140, 159, 175, 237]. Compared with traditional generation models, RAG offers several benefits. RAG substantially improves the fidelity of generated responses, especially in scenarios requiring accurate answers to factual queries. Additionally, it diminishes the frequency of 'hallucinations'-incorrect or invented outputs-by the generation model. The capability to dynamically refresh its knowledge database enables the model to remain updated with the latest developments, making it well-suited for deployment in enhanced text generation. We incorporate RAG into VRG-SLT, improving the translation with a knowledge base after converting sign language into text (Chapter 6).

## 2.2 Consistency Learning

Consistency learning is a robust approach in the field of machine learning, particularly in unsupervised and semi-supervised learning settings. It focuses on improving the reliability and accuracy of models by enforcing consistent predictions on perturbed versions of the same data. The fundamental principle of consistency learning is that a model should output similar predictions for inputs that are essentially the same but have been slightly altered or augmented in non-essential ways. By training a model to be invariant to these perturbations, it can learn more generalizable and robust features that are not overly reliant on the idiosyncrasies of the training data.

**Mutual Learning,** originating from the field of language translation, aims to reduce dependence on data annotation [83]. This mechanism allows alternating between the two sides and enables the language model to train solely from one-sided data. The core idea of mutual learning involves establishing a dual-learning game between two agents, each agent is assigned an individual task. In the primal task, mutual learning maps $x$ from primal domain to dual domain $y$, and then restore the original $x$ through the reverse mapping in dual task [296]. Hence, mutual learning can produce two feedback signals without requiring parallel data: a style evaluation score indicating the likelihood that the synthesized audio matches the target style, and a reconstruction loss measuring the difference between the reconstructed audio and the original audio. This mechanism alternates between agents, allowing the generator to train from only one-way data [150, 236, 276, 281, 290, 296]. Mutual learning is similar to cycle-consistency learning in [301] and forward-backward object tracking [151, 267]. Mutual learning explores correlations between different tasks, while

consistent learning aims to improve self-reconstruction within the same model. In Chapter 4,we are the first to investigate the duality of VAM and dereverberation. These two tasks are trained together in a mutual learning framework and provide mutual reinforcement signals based on the structural symmetry, even for unpaired samples.

**Cross-modal Consistency Learning.** Deep learning often requires ample labeled data to work properly. However, the cost of collecting sign data is prohibitive and audio data is often lacking. Recent methods enhance model training by applying consistency training to massive unlabeled data [13, 46, 170, 220]. The principle of consistency learning, employing the cyclical duality between different tasks or data as feedback signals to regularize training [83], has its roots in the domain of language translation [150, 281, 296]. As mentioned above, consistency learning primarily encompasses inter-task (mutual-learning) and intra-task (cycle-consistency learning) varieties. Mutual-learning simultaneously trains bidirectional mapping functions between tasks, creating a primal-dual pair where one function's output approximates the input of the inverse function [236, 266, 269, 276, 281, 290, 296]. Cycle-consistency learning is designed to enhance the self-reconstruction capabilities of samples produced intrinsically by the same model [6, 163, 213, 301]. However, these methods frequently emphasize the duality between two tasks or modalities, overlooking the interplay and mutual influence among multimodal data within the same task. To lessen the demands for extensive data, we employ an Embedding Consistency Learning (ECL) strategy that leverages the reciprocal relationships between modalities to facilitate training. The MS2SL (Chapter 5), utilizing ECL, generates various feedback signals by evaluating reconstruction losses through signs generated from reconstructed audio embeddings, even in the absence of tri-modal co-occurrence data.

## 2.3 Audio Style Adaptation

Audio Style Adaptation refers to the process of modifying an audio signal to match the style of another, such as changing a voice recording to resemble a particular speaker's tone or adjusting a piece of music to mimic a different genre or artist's sound. The adaptation can involve aspects like pitch, timbre and reverberation style, making it valuable for applications in film dubbing, personalized voice assistants, virtual reality (VR), augmented reality (AR)and more. This field is rapidly evolving, driven by advancements in artificial intelligence that enable more nuanced and accurate audio transformations. The objective of this research is to facilitate communication between hearing-impaired and hearing individuals, emphasizing the transfer of audio reverberation styles and speech timbres.

### 2.3.1 Melspectrogram

Melspectrograms [47, 246] are widely used representations in speech processing and audio analysis, derived from the Mel scale, which models the frequency perception of the human auditory system. Unlike linear frequency scales, the Mel scale provides higher resolution at lower frequencies and lower resolution at higher frequencies, reflecting the human ear's heightened sensitivity to low frequencies and reduced sensitivity to high frequencies.

The process of generating a Melspectrogram typically involves several key steps [49, 88, 176]. First, a Short-Time Fourier Transform (STFT) is applied to the audio signal to extract its time-frequency representation, resulting in an amplitude spectrum. This spectrum is then filtered through a Mel filter bank, which consists of filters distributed according to the Mel scale, capturing the energy across different frequency bands. Finally, a logarithmic transformation is applied to the filtered energy values to better match the logarithmic nature of human auditory perception.

Melspectrograms are especially effective in a range of applications, including speech recognition [51], music analysis [147], and sound classification [261]. By providing a compact yet expressive representation of the audio signal, they closely mimic the way the human auditory system processes sound. This makes them a valuable feature for machine learning and deep learning models, which leverage Melspectrograms for tasks such as speech recognition, sound event detection, and audio classification.

### 2.3.2 Audio Reverberation Style Transfer

Our efforts in audio reverberation style transfer efforts are centered around two tasks: acoustic matching for converting between different reverberation styles and dereverberation. Acoustic matching involves adjusting the acoustic properties of an audio signal to align with those of another environment. This process typically includes altering the reverberation characteristics to mimic a specific sound space or effect, such as transforming a home studio recording to sound as if it were made in a cathedral or concert hall. Acoustic matching is widely used in music production, post-production for films, and virtual reality to enhance auditory experiences and maintain consistent sound styles. Dereverberation refers to the process of removing unwanted echoes and reflections from audio signals, aimed at enhancing speech intelligibility. Reverberation can obscure and distance sounds in recordings, adversely affecting understanding and automatic speech recognition, especially in settings like telecommunication or speech recognition systems. Dereverberation restores the direct sound and diminishes background noise, resulting in a clearer and more natural audio.

**Acoustic Matching.** Acoustic matching involves modifying audio to simulate the sound in a given environment. Schroeder *et al.* [233] first propose the concept of reverberation and apply a series of percolators and delay lines to mimic environmental space characteristics. There are two main methods for acquiring RIRs in the audio community [70, 158, 178]. **(1)** Simulation techniques can be employed to produce RIRs when the geometry and material properties of the spatial environment are available [15, 24, 68]. **(2)** If detailed information is inaccessible, RIRs can be blindly estimated from audio captured in the room [178, 248]. RIRs are then employed to synthesize an auralized audio signal. Both methods have weaknesses. The former requires exhaustive measurements of space that may be infeasible, while the latter may introduce some disturbances due to limited acoustic information. Some recent works [119, 240] attempt to approximate RIRs from an environmental image, necessitating paired image and impulse response training data. Regrettably, these methods also require estimating the acoustic parameters from the recorded audio, which severely limits the application scopes. Chen *et al.* [30] introduce VAM and utilize visual observation to simulate the target environment for generating reverberant audio. However, VAM focuses on acoustic matching, neglecting the correlation and inherent consistency with the reverse dereverberation task. By adopting a de-biaser as the acoustic residue metric, [245] improves performance while also raising the overall complexity and training difficulty of the framework. In Chapter 4, we strategically harness the RGB image of a specified environment for precise acoustic matching, and utilize the inherent reciprocity with dereverberation techniques to improve the accuracy of our reverberation simulations.

**Dereverberation.** Due to the challenge of collecting both anechoic and reverberant audio simultaneously, acoustic dereverberation can enhance training data quality by minimizing reverberation disturbance [116, 294]. The main stream dereverberation technologies utilize devices like microphone arrays to remove reverberation [171]. Deep learning techniques have also made great strides in reverberation removal [79, 271, 295]. Tan *et al.* [253] exploit the movement of the upper lip region to isolate interfering sounds, yet it does not intentionally eliminate reverberation based on visual scene understanding. These methods either disregard or only partially take into account visual information. Chen *et al.* [32] propose learning all the acoustics characteristics associated with indoor dereverberation. Like acoustic matching, these unidirectional approaches neglect the reciprocal relationship between the two tasks, leading to an incomplete utilization of naturally recorded audio. In contrast, the MVSD model demonstrates significantly stronger dereverberation capabilities through the effective assistance of symmetric tasks. Furthermore, MVSD can substantially alleviate the data collection burden by utilizing ample naturally recorded audio resources (Chapter 5).

### 2.3.3 Voice Conversion

Voice Conversion (VC) is a speech technology that allows for the conversion of voice characteristics from one person to another while preserving the original semantic content. It is widely used in various fields such as personalized voice assistants, the entertainment industry, and speech synthesis, enriching and diversifying voice communication.

Recent years, deep learning methods have dominated voice conversion. To obtain high naturalness and intelligibility, previous works utilize some text annotation information or auxiliary networks [7, 62, 106, 275], such as automatic speech recognition (ASR), fundamental frequency (F0) networks, *etc.*Liu *et al.* [143] employ an ASR network trained with text annotations to recognize phoneme sequences from speech signals. StarGANv2-VC [135] supplemented by F0 and ASR network, significantly improves the naturalness and intelligibility of the converted speech. However, methods [135, 221] rooted in StarGANv2 [41] can only transform the style of the 'heard' data, *i.e.*, all speakers appear in the training set. After removing the auxiliary networks, the performance of them drops a lot. Some works try non-GANs methods [3, 148, 235] such as VAM, normalizing flow, *etc.* Akuzawa *et al.* [3] utilize a deep hierarchical VAE to achieve high model expressivity and fast conversion speed. Blow [235] proposes a normalizing flow generation model to maps the voices of different speakers to a latent space. To reduce data collection and annotation dependence, researchers are exploring unsupervised methods [82, 106]. Qian *et al.*[205] propose a holistic rhythm transition model without text transcription. AUTOVC [206] and AdaIN-VC [42] conduct zero-shot attempts, but exhibit flaws in the conservation of the source speech content.

There is also a voice cloning task that is similar in function to voice conversion and can be easily confused with it. Generally speaking, voice conversion refers to altering the vocal timbre of the source speech, whereas voice cloning [7, 29, 154, 277, 287] involves replicating the voice of a specific person, including not just timbre but also characteristics like intonation, rhythm and accent. For input, voice conversion typically uses source and target speaker audio, while popular voice cloning methods synthesize audio through target vocal traits and content text, which may not convey all intended information like emotions.

Regrettably, prior research has often overlooked the disparities in vocal ranges across different speakers. Simply converting the entire melspectrogram on a global scale cannot fully decouple timbre style and speech content. Our method leverages the unique spatial characteristics of individual subbands for more precise conversion. By tailoring each subband to align with the specific vocal range between the source and target speakers, SGAN-VC shows impressive performance in both timbre similarity and content retention (Chapter 3).

## 2.4 Sign Language Translation and Production

### 2.4.1 Sign Language Translation

Sign Language Translation aims to precisely recognize and explain sign language components such as the shapes, positions, and movements, translating them into equivalent verbal language. Similar to spoken language, sign language follows specific linguistic rules [16, 198, 223, 224]. Existing researches are primarily dedicated to sign language recognition (SLR) and translation (SLT). SLR [2, 234] means interpreting and classifying of body movements in videos, covering isolated [98] and continuous signs [21, 22, 48]. SLT typically involves translating sign language into spoken language [21, 22, 50, 138].

Isolated SLR and the more challenging continuous SLR are two fundamental tasks for understanding sign language. One aims to identify single annotated word labels in short video clips [4, 134], while the other seeks to convert continuous sign videos into gloss [1] sequences using only weak sentence-level annotations [48, 118, 204, 300]. While some previous studies [188] equate SLR with SLT, the former merely classifies signs without considering their grammatical and morphological structures into spoken language. Jiang *et al.* [101] propose a fresh multimodal framework featuring a global integrated model for skeleton-aware multimodal learning in discrete SLR. Generally, SLR serves as an intermediate step in the translation process, annotating sign language videos before converting them into spoken language through a sequence-to-sequence method [21, 22, 56, 133, 298, 300]. For instance, Camgöz *et al.* [22] integrate the training of SLT to regularize the translation encoder, and Zhou *et al.* [298] introduce a data augmentation approach that uses annotations as pivots to back-translate text into visual features. Cico [38] models the relationship between signs and text from a cross-lingual retrieval perspective. Chen *et al.* [36] develop a unified framework for SLT, dividing it into visual and linguistic modules bridged through a visual-linguistic mapper for training. Influenced by action recognition [10, 99, 257], some studies [20, 37, 80, 167, 184] explore directly modeling RGB videos to understand sign language. However, these methods still struggle with translating entire sentences. In Chapter 6, we utilizes Sign-tokenizer to treat raw sign motions as equivalent to textual 'words', co-training with spoken text to transcend the linguistic barriers.

[1] Glosses are the word-for-word transcription of sign language, where each gloss is a unique identifier for a sign. For example, the American Sign Language (ASL) sign for 'dog' involves patting your hip, and is glossed as DOG in all caps, linking it directly to its English counterpart. Another example is 'WHAT'S-UP' in ASL, signed with 'W' hands moving upwards, glossed as a single term, reflecting its unique cultural significance.

### 2.4.2 Sign Language Production

Sign language production (SLP) is the process of creating sign sequences from spoken text, and can be seen as the reverse process of SLT [8]. These existing studies on SLT and SLP primarily focus on converting between sign videos and discrete glosses, either directly or indirectly. A few of Text2Sign works [226–228] are grounded in datasets with relatively homogeneous scenario [21] and discrete spoken transcriptions. However, limited studies focus on directly generating sign language sequences from entire spoken sentences. To our best knowledge, we are the pioneers in effecting this conversion. Chapter 5 harnesses sequential diffusion models to incrementally generate noise predictions, enabling cross-modal sign language generation. With the help of ECL, our SLP framework (MS2SL) can generate various feedback signals even in the absence of co-occurring ternary data: assessing the reconstruction loss with the signs generated from the reconstructed audio embeddings.

CHAPTER 3

# SUBBAND-BASED GAN FOR VOICE CONVERSION

This chapter delves into the domain of non-parallel many-to-many voice conversion (NPVC), an area that stands at the intersection of signal processing and machine learning. NPVC offers the capability to transform vocal timbres without the necessity for paired source and target speaker audio or text transcriptions, thereby providing a more adaptable solution for real-world scenarios. This capability not only enhances personalized communication but also serves critical roles in entertainment, therapeutic applications, and privacy protection by adjusting the timbre properties of speech.. Within the overall system, NPVC plays a key role in tailoring the generated speech to suit different audience and recording needs, improving the clarity and personalization of spoken content before it is translated into sign language. However, conventional methods for holistic spectrogram transformation commonly neglect the disparities and unique traits inherent to different frequency bands. This oversight in acknowledging the distinct roles played by low and high-frequency components in information expression can compromise both the conversion quality and the accuracy of voice timbre.

In this work, we introduce a subband-based generative adversarial network (SGAN-VC)[1], which explicitly leverages the spatial characteristics among different subbands to individually convert the content of each subband. Additionally, we utilize attention mechanisms to capture the complex dependencies between the target vocal style and the source speech content, thereby enhancing the timbre accuracy and naturalness of the generated speech.

[1]This chapter is based on collaborative research [157] by Jian Ma, Zhedong Zheng, Hao Fei, Feng Zheng, Tat-seng Chua and Yi Yang, and is prepared for peer review submission.

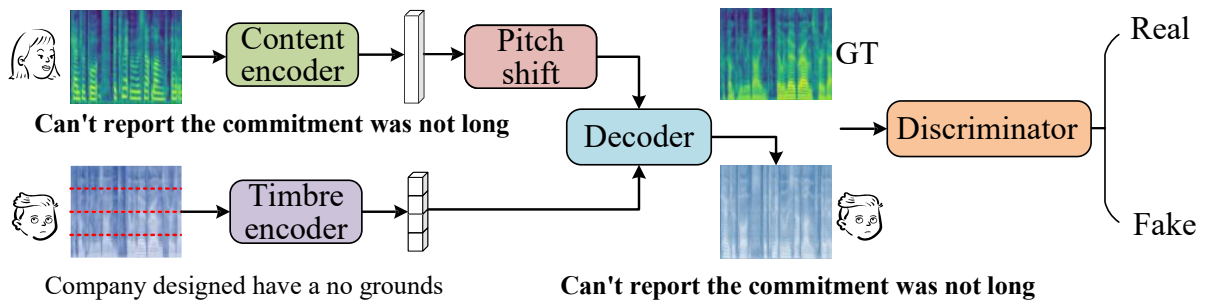


Figure 3.1: SGAN-VC trains a GAN network based on the strategy of decoupling content and style and can transfer the speech of the source speaker into the style of an arbitrary target speaker without parallel data. Performing voice conversion independently for each subband enhances the ability to capture frequency-domain differences within the context (§3.1).

Experimental setups, datasets used, and evaluation metrics are meticulously described to provide clarity on how the voice conversion models are trained, validated, and tested. This includes a discussion on objective measures such as pMOS and subjective listening tests to assess voice quality and speaker similarity. Comprehensive experimental evaluations reveal that our proposed method sets a new benchmark for performance on both the VCTK Corpus and AISHELL-3 datasets. Notably, SGAN-VC surpasses the content intelligibility of StarGANv2-VC, even when the latter is augmented with an ASR network.

Finally, the chapter concludes with insights into the potential future directions of voice conversion research, emphasizing the need for more robust, adaptable, and efficient models to handle diverse real-world scenarios.

## 3.1 Introduction

Voice conversion (VC) seeks to produce a new speech that merge source content with target vocal style [1, 39]. The generated speech is expected to preserve the source linguistic content, while transferring to the vocal characteristics of the target speaker [174, 242]. VC can be applied to various real-world applications, such as robot voice personalization [230, 293], voice de-identification [102, 201], video dubbing [272, 286], speech intelligibility enhancement [196, 273], *etc.* Traditional VC models [14, 41, 42] typically require paired audio from both the source and target speakers for training. These models are limited to generating the voice styles of target speakers that are present in the training set, often referred to as 'heard' speakers. Despite producing competitive speeches, they still fail to meet practical demands, which usually requires generalization to arbitrary voices. Hence, some research efforts [105, 107, 135, 206] have begun to explore NPVC. Compared to traditional voice con-

version, NPVC can generate vocal timbres for speakers who are not present in the training dataset, thus broadening its range of application.

VC demands the simultaneous attainment of high timbre similarity and intelligibility in the generated audio, whereas NPVC presents ongoing challenges in achieving speech naturalness, content clarity, and stylistic resemblance [292]. Prevalent methods that convert entire melspectrograms often overlook subtle differences present in local details. These nuances may manifest as distinct vocal characteristics in timbre. Therefore, when endeavoring to accurately replicate a target speaker, the absence of these key details may result in noticeable discrepancies in the vocal tone.

In this chapter, based on the intrinsic structure of the melspectrogram[2] (where the vertical axis represents audio frequency and the horizontal axis represents time), we introduce SGAN-VC, a subband-based NPVC framework (Fig. 3.1). We vertically splits the melspectrogram into four distinct subbands, each of which independently undergoes voice conversion. In particular, SGAN-VC follows a GAN architecture with a generator and a discriminator, employing a common strategy for decoupling content and style. The generator includes a timbre style encoder, a content encoder, and a decoder. The timbre encoder is designed to learn the local style codes of different subbands for the target speaker, and the content encoder captures the content information of the source speech. Finally, after replacing the source voice's timbre with the target style, the decoder produces the specific subband content. To ensure overall style consistency, each local feature is combined with global features through concatenation. In harmony with this division, our decoder consists of four vertically aligned modules with a shared architecture but independent parameters. Each module focuses on rendering content to its designated frequency band. SGAN-VC stands out for its ability to interchange local and global timbre information, making it to produce voices closer to the target speaker. To further bridge the gap with target style, we employ a unique pitch transformation module to fine-tune the pitch of the source speaker, improving the precision of the pitch conversion.

SGAN-VC shines in the realm of voice conversion for several compelling reasons. Firstly, it uniquely excels at simultaneously focusing on both the overall spectrogram and intricate details simultaneously. Secondly, when faced with source and target speakers of differing vocal ranges, its decoder astutely determines the content generated for each corresponding subband, ensuring the capture of subtle stylistic differences between the two speakers.

[2]The melspectrogram is an audio feature representation derived from the Mel scale. It transforms the frequency spectrum of an audio signal into a form that better matches human auditory perception, using short-time Fourier transform and Mel filter banks. Its key advantage is the ability to simulate human frequency perception while providing a compact and informative audio representation.

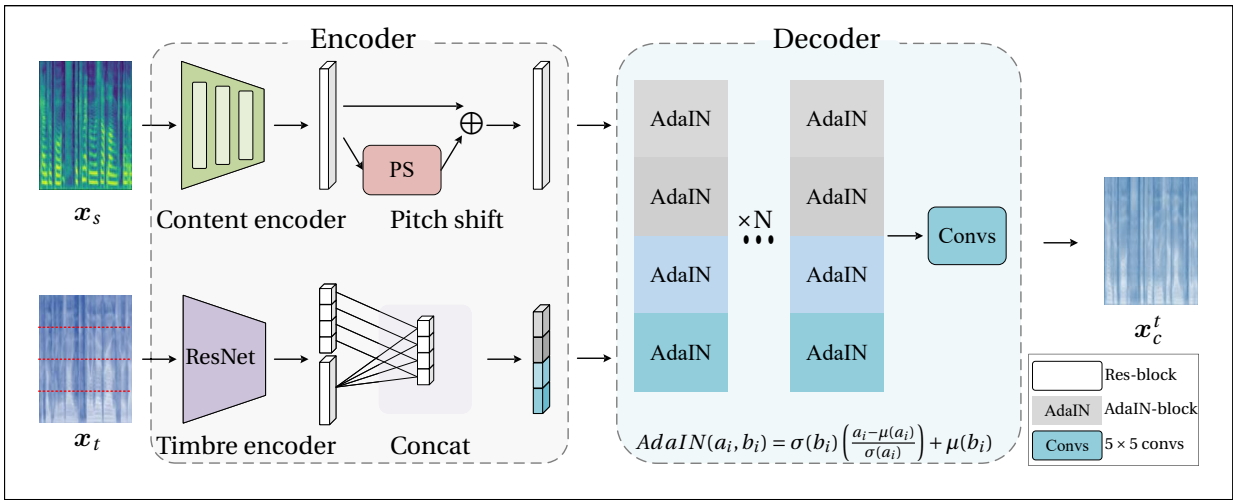


Figure 3.2: The structure of SGAN-VC. A Subband-Block consists of four AdaIN-Blocks. Each part of the style feature from the target speech is fed into the corresponding AdaIN-Block. The Subband-Block exports the respective feature after the style information is exchanged (§3.2).

We conduct experiments on both English and Mandarin datasets. Extensive experiments substantiate that our approach delivers highly competitive performance. We observe the generalizability and effectiveness of SGAN-VC, demonstrating its proficiency in handling both same-gender and cross-gender voice conversion.

To encapsulate, the principal contributions of this chapter can be articulated as follows:

- We present SGAN-VC, a highly efficient and robust NPVC architecture. This innovative framework is designed to access data within individual subbands, significantly elevating the quality of voice conversion. Moreover, SGAN-VC eliminates the need for textual annotations and thereby streamlines the model deployment process.

- The pitch-shift module is adept at forecasting frame-level pitch variations, enhancing both the robustness and explainability of the timbre conversion.

- We verify our method on both the English dataset, *i.e.*, VCTK Corpus [278] and the Mandarin dataset,*i.e.*, AISHELL3 [238], achieving top-tier performance in retaining content fidelity and style consistency.

This chapter is systematically organized as follows: Section 3.2 describes the proposed SGAN-VC in detail. Section 3.3 then comparatively discusses the experimental results, culminating with the conclusion in Section 3.4.

## 3.2 Method

As illustrated in Fig. 3.2, SGAN-VC tightly integrates a generative network and a discriminative module for NVPC. The generator transforms the source sample to produce speech in the desired target style. Meanwhile, the discriminator assesses if the input melspectrogram originates from an authentic speech sample. The generator is composed of a content encoder, a timbre style encoder, and a decoder to decouple and perform style transfer on different speeches. To address voice variations, we uniquely incorporate the Subband-block to amalgamate local features within the decoder. In particular, our style features are divided into four distinct segments, each segment corresponding to a specific subband in the speech. We next provide a formal explanation of the problem and its associated symbols.

### 3.2.1 Problem Formulation

We denote the real melspectrogram and speaker labels as $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$ and $Y = \{\boldsymbol{y}_i\}_{i=1}^{N}$, where $N$ is the total number of melspectrograms in the dataset. In addition, we suppose that the source speaker $S$ and the target speaker $T$ are two variables randomly selected from the speaker pool, respectively. As shown in Fig. 3.2, the Generator consists of a content encoder $E_c : \boldsymbol{x}_s \rightarrow \boldsymbol{c}_s$, a timbre encoder $E_s : \boldsymbol{x}_t \rightarrow \boldsymbol{f}_t$ and a decoder $G : (\boldsymbol{x}_s, \boldsymbol{f}_t) \rightarrow \boldsymbol{x}_s^t$. Our generative module generates a new spectrogram $\boldsymbol{x}_s^t$ after exchanging source and target style information (Eq. 3.1). The role of the discriminator is to distinguish whether the input data comes from the real data distribution or is generated by the generator.

$$\boldsymbol{x}_s^t = G(\boldsymbol{c}_s, \boldsymbol{f}_t) = G(E_c(\boldsymbol{x}_s), E_s(\boldsymbol{x}_t)), \tag{3.1}$$

where $\boldsymbol{c}_s$ and $\boldsymbol{f}_t$ represent content and style features, respectively.

### 3.2.2 The Generative Network

**Subband Block.** Dividing the melspectrogram into subbands can offer several advantages [27, 139, 199]. Firstly, this segmentation allows for finer processing within each frequency band, enhancing the accuracy of signal analysis and conversion. Additionally, the independent adjustment and optimization of each sub-band provide greater flexibility in audio quality control, particularly useful in applications like sound transformation and speech synthesis, where it enables more precise control over pitch, loudness, and timbre, resulting in a more natural and high-quality output. Lastly, this division helps reduce interference between frequency bands, improving overall sound quality, especially in multitasking

environments where it prevents issues in one band from affecting others. These benefits collectively enhance the overall performance and output quality of audio processing.

To integrate the style feature, we harness the AdaIN-Block module [95] for style information interchange. AdaIN adjusts the feature map of source spectrogram to match the style of target spectrogram. Firstly, AdaIN applies instance normalization (IN) to the feature map of the content spectrogram, conducting this process independently for each sample.

$$IN(a_i) = \frac{a_i - \mu(a_i)}{\sigma(a_i)}. \tag{3.2}$$

Subsequently, the normalized content feature map is adjusted using the mean and standard deviation of the style feature map.

$$AdaIN(a_i, b_i) = \sigma(b_i)\left(\frac{a_i - \mu(a_i)}{\sigma(a_i)}\right) + \mu(b_i), \tag{3.3}$$

wherein, $a_i$ is a channel of the content feature map, $b_i$ is the corresponding channel of the style feature map, $\mu(b_i)$ and $\sigma(b_i)$ are respectively the mean and standard deviation of that channel in the style feature map. AdaIN can adeptly modify feature maps of content spectrograms to match feature distributions of the target timbre, simultaneously preserving the integrity of the original content.

Motivated by the inherent structure of melspectrogram, where the vertical axis represents Mel frequencies, we design a subband-focused module named Subband-Block, comprised of four discrete AdaIN-Blocks. In particular, as shown in Fig. 3.2, we partition the style feature into four distinct subbands. Each AdaIN-Block is fed with a single subband feature. This arrangement allows the subband-Block to construct the converted speech sequentially from higher to lower frequencies. By executing style transfer separately for each frequency band, the subband-Block is better equipped to discern pitch variations among speakers. Conclusively, we transform the amalgamated features from the four subbands into the final melspectrogram using a pair of $3 \times 3$ convolutions.

**Pitch-shift Module.** The pitch-shift module (PS) is utilized to precisely adjust the pitch of the source speaker. As illustrated in Fig. 3.2, this module is harnessed to make vertical alterations to the content feature, thereby modifying the frequency of the source spectrogram. The structure of the pitch-shift module encompasses a sequence of $5 \times 5$ convolutions followed by a $1 \times 1$ convolution. Ultimately, PS produces an offset vector that aligns with the time dimension of the content feature. Subsequently, the Tanh activation function is invoked to standardize the vector within the $(-1, 1)$ range. Consequently, this offset vector serves as a representation of the displacement for each frame in the mel-spectrogram. In order to preserve the content integrity of the source spectrogram effectively, shifts are exclusively made in the vertical direction for every frame.

### 3.2.3 Optimization

Our goal is to learn the mapping of $x_s$ to $x_s^t$ from the source domain to the target domain without parallel data. Referring to and StarGANv2-VC [135], we adopt the following loss function as the objective optimization.

**Adversarial loss.** The generator takes a source content feature $\boldsymbol{c}_s$ and a style feature $\boldsymbol{f}_t$ and learns how to generate a new spectrogram $\boldsymbol{x}_s^t$. The objective of the generator is to cheat the discriminator by leveraging the adversarial loss.

$$\begin{aligned}\mathscr{L}_{adv} = \mathbb{E}[log(Dis(\boldsymbol{x}_s, \boldsymbol{y}_s))] \\ + \mathbb{E}[log(1 - Dis(G(\boldsymbol{c}_s, \boldsymbol{f}_t), \boldsymbol{y}_t))],\end{aligned} \tag{3.4}$$

where $Dis(\cdot; \boldsymbol{y})$ denotes the output of discriminator for the speaker class $\boldsymbol{y} \in Y$.

**ID loss.** To enhance the timbral style discrimination capability of SGAN-VC, we incorporated an identity classification loss into the target melspectrogram.

$$\mathscr{L}_{id} = -\mathbb{E}[log(p(\boldsymbol{y}_t | G(\boldsymbol{c}_s, \boldsymbol{f}_t)))], \tag{3.5}$$

where $p(\boldsymbol{y}_t | x_\cdot)$ denotes the predicted probability of $\boldsymbol{x}_\cdot$ belonging to the class $\boldsymbol{y}_t$.

**Style Consistency Loss.** To maintain a consistent timbre style with the target speaker, we use a style consistency loss in our generation model. This loss helps to minimize deviations in vocal timbre, ensuring that the synthesized speech sounds more like the target speaker across different utterances.

$$\mathscr{L}_{style} = \mathbb{E}[||\boldsymbol{f}_t - E_s(G(\boldsymbol{c}_s, \boldsymbol{f}_t))||_1]. \tag{3.6}$$

**Content Consistency Loss.** Voice conversion alters the stylistic attributes of speech while retaining its linguistic content. The content consistency loss guarantees alignment between the generated speech and the original source content.

$$\mathscr{L}_{content} = \mathbb{E}[||\boldsymbol{c}_s - E_c(G(\boldsymbol{c}_s, \boldsymbol{f}_t))||_1]. \tag{3.7}$$

**Norm consistency loss.** The absolute column-sum norm of a mel-spectrogram significantly signifies the overall sound energy level, which can subsequently be utilized to discern between active speech and silence states. Like StarGANv2-VC [135], we utilize norm consistency loss $\mathscr{L}_{norm}$ to retain the speech/silence status of the source speech. Define $m$ is the index of the $m^{th}$ frame in $x$. The norm consistency loss is given by:

$$\mathscr{L}_{norm} = \mathbb{E}[||||\boldsymbol{x}_s[m]||_1 - ||G(\boldsymbol{c}_s, \boldsymbol{f}_t)[m]||_1||_1]. \tag{3.8}$$

where $||\boldsymbol{x}[m]||_1$ and $||G(\boldsymbol{c}_s, \boldsymbol{f}_t)[m]||_1$ represent the absolute column-sum norm of the $m^{th}$ frame of source and converted melspectrograms, respectively.

**Reconstruction loss.** Based on the spirit of auto-encoder, we propose that source speech can be reconstructed by its content and style features.

$$\mathscr{L}_{rec} = \mathbb{E}[||\boldsymbol{x}_s - G(\boldsymbol{c}_s, \boldsymbol{f}_s)||_1]. \tag{3.9}$$

**Full generator objective.** It is designed to minimize the discrepancy between the generated and real data, incorporating both adversarial and feature matching losses. Our full generator objective function can be summarized as follows:

$$\begin{aligned}\mathscr{L}_{total}(E_c, E_s, G) = \lambda_{adv}\mathscr{L}_{adv} + \lambda_{id}\mathscr{L}_{id} + \lambda_{style}\mathscr{L}_{style} \\ +\lambda_{content}\mathscr{L}_{content} + \lambda_{norm}\mathscr{L}_{norm} + \lambda_{rec}\mathscr{L}_{rec}.\end{aligned} \tag{3.10}$$

where $\lambda_{adv}$, $\lambda_{id}$, $\lambda_{style}$, $\lambda_{content}$, $\lambda_{norm}$ and $\lambda_{rec}$ are hyperparameters for each term. Besides, the discriminator is update by $-\lambda_{adv}\mathscr{L}_{adv}$.

## 3.3 EXPERIMENT

### 3.3.1 Datasets

Our evaluation of SGAN-VC mainly focuses on two distinct datasets: VCTK Corpus [278] and AISHELL3 [238], also including comprehensive evaluation of heard and unheard data. We strongly encourage readers to listen to the audio samples[3].

**VCTK Corpus** [278] contains 47 male speakers and 62 female speakers, with a relatively balanced gender ratio. For a fair comparison, we first utilize the same 20 speakers reported in [43, 135] for the heard data experiment, called VCTK20. The discrepancy from [43, 135] is that all our audio fragments are randomly sampled from the original VCTK Corpus. Therefore, the pMOS of ground truth is lower than reported in [135]. Regarding the test set, we select 5 males and 5 females from VCTK20. Each speaker contains 50 samples that do not present in the training data. For the unheard data experiment, our training set applies all the speakers of the original VCTK Corpus except the 10 speakers in the test set.

**AISHELL3** [238] is an extensive Mandarin corpus featuring multiple speakers. It encompasses recordings from 218 Chinese speakers, comprising 176 females and 42 males. We use all male speakers and a randomly chosen subset of 42 female speakers to constitute our evaluation dataset, which we designate as **AISHELL3-84**. Likewise, 5 male and 5 female

[3] https://hechang25.github.io/SGAN-VC

speakers are randomly selected in AISHELL3-84 as the final test set. We remove samples shorter than 2.5 seconds from AISHELL3-84. Eventually, each speaker has 50 and 48 audio samples in the training and test sets, respectively.

### 3.3.2 Evaluations

We employ the predicted mean opinion score (pMOS) from MOSNet [146], classification accuracy (CLS), and character error rate (CER) for quantitative evaluation. Similar to [202] and StarGANv2-VC [135], we adopt ResNet as the classifier to determine timbre similarity. A higher classification accuracy indicates a more accurate timbre conversion. Given that the style characteristics of certain speakers are similar, training solely on the chosen 10 test speakers will result in some inaccurate conversion results being misclassified, leading to falsely high accuracy. Therefore, we train ResNet on all speaker data for the VCTK Corpus and AISHELL3-84. To evaluate intelligibility, we assess CER metric through utilizing the open-source ASR toolkit WeNet [280]. For the English corpus VCTk, we apply the pre-trained model on LibriSpeech [191] dataset. Meanwhile, for the Mandarin dataset AISHELL3-84, we employ the pre-trained model on multiple fusion data sets [18, 55, 203, 265]. Furthermore, distinguishing subtle differences between two conversion results for the same speaker can be challenging for the human ear, especially when the differences are minimal. To overcome this, we propose using the average fundamental frequency difference, denoted as $mF0_{diff}$, between the conversion sample and the target speaker. $mF0_{diff}$ can provide a more objective measure and enhances the effectiveness of evaluating style similarity.

$$mF0_{diff} = \mathbb{E}||F0_{x_s^t} - F0_{x^t}||_1. \tag{3.11}$$

### 3.3.3 Network Architectures

) **Content encoder** is designed to extract the content information of the source speech. In particular, the content encoder extracts language-related features from the spectrogram and removes speaker-related style information. As shown in Fig. 3.3, we adopt Res-Block as the basic module like in StarGANv2 [41]. Generally speaking, a larger feature map enhances the capability of the encoder to perceive fine-grained information. Thus, to ensure a richer detail capture, we downsample only twice vertically and once horizontally.

**Timbre encoder.** The primary objective of the timbre encoder is to filter out content information, leaving only the style embedding of the reference speaker. Our style encoder refers to the ResNet50 [84], which has been proven to be a robust image classification model. Given that the mel-spectrogram features a single channel, we adjust the input dimension

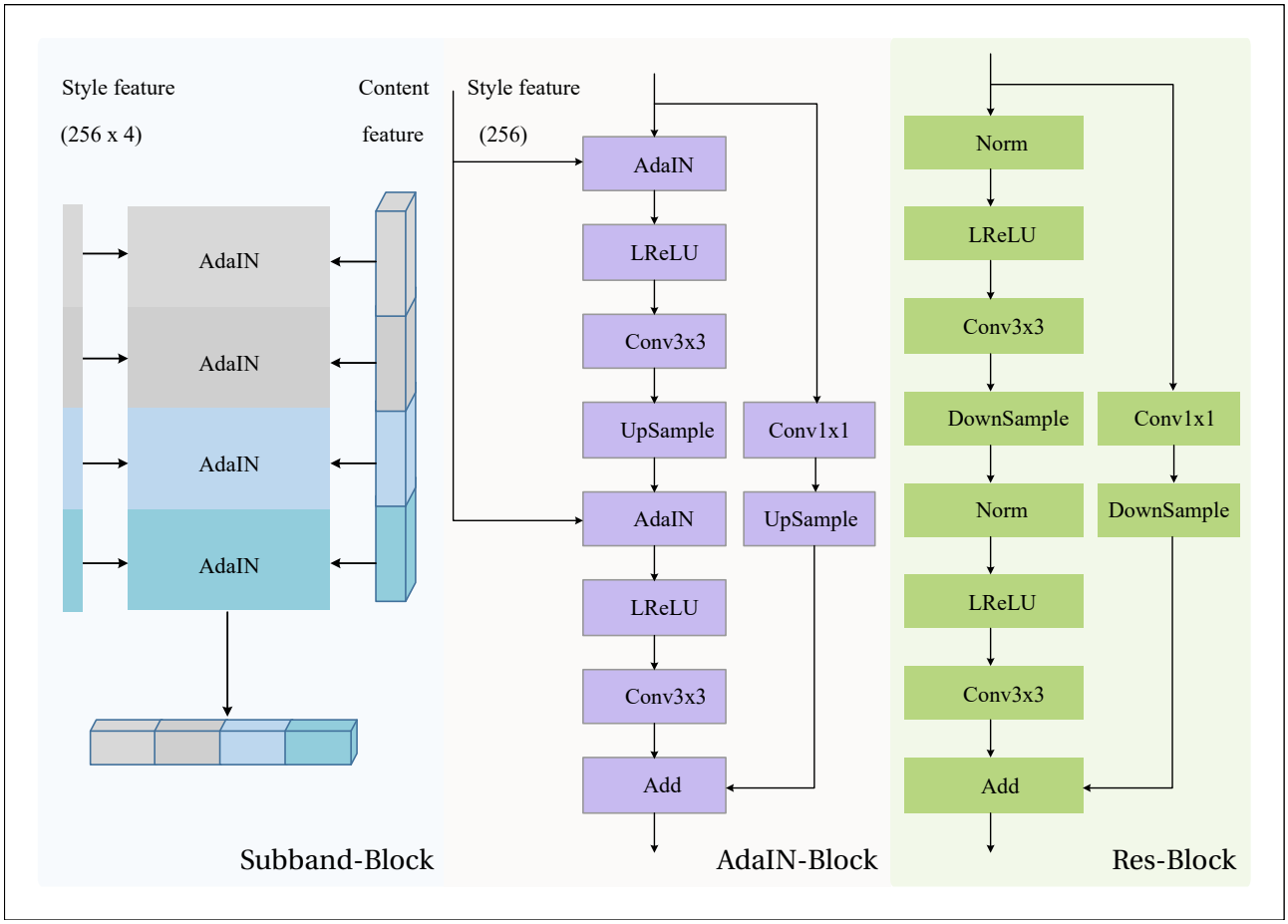


Figure 3.3: Diagram of network components in SGAN-VC (§3.3.3).

of ResNet50 to 1. We are particularly attentive to the local traits of each subband and its detailed information, leading us to eliminate the final downsampling stage present in the original ResNet50. Then, as illustrated in Fig. 3.3, the feature map is vertically divided into four parts by adaptive average pooling, each part can represent the spatial characteristic of a subband. Furthermore, to preserve the overarching style consistency, we harness a global feature from the entire feature map through average pooling. This overarching global feature is subsequently concatenated with each localized feature to individually represent the content of every subband. Ultimately, we implement two linear layers to effectively reduce the dimensions of the style features.

**Decoder.** The decoder is mainly composed of six subband-blocks and two $3 \times 3$ convolutions (Fig. 3.2). As can be seen in Fig. 3.3, a subband-block contains 4 AdaIN-blocks. Each AdaIN-Block completes style transfer, according to the content feature from the source speaker and the style feature from the target speaker. Each AdaIN-Block focuses on the conversion of the corresponding frequency band content. To make the generated mel-spectrogram and the source spectrogram have the same size, AdaIN-Block upsamples once in the horizontal

direction. The last subband-block generates four feature maps of size $(64, 20, 224)$. We concatenate the feature maps together to form features of $(64, 80, 224)$. Eventually, we utilize two $3 \times 3$ convolutions for feature fusion and generate converted melspectrograms.

**Discriminator.** During training, there is an adversarial relationship between the discriminator and generator. The discriminator processes features from input samples, evaluating if they are real or generated by the generator. The dynamic adversarial interaction fosters improvement in both networks, ultimately leading the generator to produce high-quality, hard-to-distinguish fake samples. To obtain the real/fake prediction, we deploy one $3 \times 3$ convolution, four Res-Blocks, one $5 \times 5$ convolution, and one $1 \times 1$ convolution.

### 3.3.4 Training Details

**Data Processing.** We begin by resampling all audio clips to a rate of 22.050 kHz, subsequently transforming the original speech waveform into a mel-spectrogram. The FFT (Fast Fourier Transform) size is set at 1024, with a hop size of 256. This mel-spectrogram is scaled to an 80-bin configuration. For consistent input dimensions, we adjust the width of each mel-spectrogram to 224, padding with zeroes where necessary. As a result, the final dimension of each processed mel-spectrogram stands at $(1, 80, 224)$.

**Training.** We initiate the process by transforming the sound signal into a mel-spectrogram. Subsequently, the source and target spectrograms are channeled into the content and timbre encoders, respectively. The decoder then outputs a mel-spectrogram, merging the content of the source with the style of the target. Concurrently, the discriminator evaluates the proximity of the generated spectrogram to authentic data. We train our model for 100 epochs with a batch size of 16, about 2.6 second long audio segments. We employ AdamW [149] optimizer with a learning rate of 0.0001. For data augmentation, we mainly use time warping and frequency masking proposed in [194]. The timbre encoder is first pre-trained on the same training set. Drawing on StarGANv2-VC, we empirically set $\lambda_{adv} = 2$, $\lambda_{id} = 0.5$, $\lambda_{style} = 5$, $\lambda_{content} = 10$, $\lambda_{norm} = 1$, $\lambda_{ds} = 1$, $\lambda_{rec} = 5$.

**Inference.** Generally, there are two common pipelines for voice conversion: 1) directly modifying the audio signal to obtain the target audio. 2) Initially, the audio signal is transformed from the time domain to the frequency domain to generate a spectrogram, after which the voice conversion is completed on the spectrogram. Finally, a vocoder is used to revert the converted spectrogram back into an audio signal. Here, we adopt the second approach, where a vocoder HiFi-GAN [120] is employed to transition the mel-spectrogram, produced by the generator, back into an auditory waveform.

| Methods | Heard | VCTK | | | | AISHELL3-84 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | pMOS↑ | CLS↑ | CER↓ | $mF0_{diff}$ ↓ | pMOS↑ | CLS↑ | CER↓ | $mF0_{diff}$ ↓ |
| Ground truth | - | 3.484 | 96.60 % | 5.27 % | - | 3.122 | 99.79 % | 2.52 % | - |
| AUTOVC | × | 3.031 | 2.21 % | 74.18 % | 16.87 | 3.191 | 1.78 % | 87.15 % | 14.45 |
| AdaIN-VC | × | 3.573 | 76.48 % | 60.58 % | 3.27 | 3.138 | 89.68 % | 64.59 % | 5.46 |
| SGAN-VC | × | 3.595 | 27.70 % | 25.42 % | 5.13 | 3.130 | 57.08 % | 25.75 % | 6.66 |
| AUTOVC | ✓ | 3.027 | 86.49 % | 73.36 % | 17.22 | 3.055 | 94.25 % | 101.03 % | 10.62 |
| AdaIN-VC | ✓ | 3.616 | 73.05 % | 68.95 % | 2.90 | 3.090 | 99.06 % | 91.19 % | 4.96 |
| StarGANv2-VC | ✓ | **3.665** | 94.10 % | 35.22 % | 12.12 | 3.155 | 92.92 % | 69.05 % | 5.52 |
| SGAN-VC | ✓ | 3.479 | **97.60 %** | **20.78 %** | **1.88** | **3.206** | **99.90 %** | **35.61 %** | **3.67** |

Table 3.1: Quantitative results on VCTK Corpus and AISHELL3-84 test set (§3.3.5). The term "Heard" specifies if the speakers are included in the training set. The test samples in VCTK and AISHELL3-84 consist of 1,000 and 960 utterances, respectively.

### 3.3.5 Experimental Results

**Quantitative Results.** We conduct comprehensive comparative experiments in both 'heard' and 'unheard' settings on the datasets VCTK and AISHELL3-84.

SGAN-VC demonstrates competitive performance in both settings in Table 3.1. In the unheard results, we can observe that SGAN-VC achieved a significantly lower CER compared to other methods. Most notably, AdaIN-VC stands out with the highest scores in the CLS metric for the unheard test. However, the primary goal of VC is to craft utterances that merge the content of source speech with the timbre of the target voice. The CER of AdaIN-VC soars to a staggering 64.59%, gravely compromising the clarity and intelligibility of the source speech. In a similar vein, AUTOVC falls short in preserving speech intelligibility. Such a high CER could result in listeners not comprehending the original speech content, indicating that AUTOVC and AdaIN-VC are not capable of effectively completing the VC task.

SGAN-VC also achieves exceptional performance in the 'heard' results and alleviates the demand for impeccable audio quality. Instead of solely pursuing high-fidelity speech, our prime concern gravitates towards timbral similarity and content comprehensibility. SGAN-VC simultaneously attains the lowest CER (20.78% and 35.61%) and the highest CLS. In terms of CLS, we witness enhancements of +2.80% and +6.98% over StarGANv2-VC on VCTK Corpus and AISHELL3-84, respectively. This hints at our generated samples

| ID Number | Gender | StarGANv2-VC | SGAN-VC | |
|---|---|---|---|---|
| | | Heard | Heard | Unheard |
| 1 | F | 13.15 | 2.97 | 3.99 |
| 2 | F | 16.30 | 0.85 | 7.91 |
| 3 | F | 16.49 | 0.88 | 4.37 |
| 4 | F | 11.34 | 0.44 | 1.50 |
| 5 | F | 15.03 | 2.66 | 9.75 |
| 6 | M | 18.64 | 3.79 | 4.51 |
| 7 | M | 17.90 | 0.90 | 2.20 |
| 8 | M | 1.05 | 5.39 | 6.98 |
| 9 | M | 5.11 | 0.96 | 3.19 |
| 10 | M | 6.25 | 0.00 | 6.91 |
| $\mathbf{m}F0_{diff}$ | - | 12.12 | **1.88** | **5.13** |

Table 3.2: Comparison results of $F0_{diff}$ on the VCTK Corpus test set (§3.3.5). We replace the ID information of the speakers with numbers 1 – 10. All values in the table are in Hz.

bearing a heightened resonance with the target timbre. Considering the pMOS metric, SGAN-VC marginally lags behind StarGANv2-VC on the VCTK Corpus, yet pulls ahead slightly on the AISHELL3-84 dataset. As reflected in Table 3.1, the pMOS for VCTK Corpus significantly surpasses that of AISHELL3-84. StarGANv2-VC suffers a notable decline across distinct datasets. Conversely, the performance of SGAN-VC remains consistently robust, even outperforming the ground truth by approximately 0.08. Besides, SGAN-VC consistently outperforms other models, with the $mF0_{diff}$ for several speakers dropping below 1Hz, which underscores the efficacy of SGAN-VC. Due to the inferior content retention ability of AUTOVC and AdaIN-VC, our subsequent analysis mainly focuses on contrasting the performance of SGAN-VC and StarGANv2-VC.

**User Study.** We conduct subjective perceptual experiments with human listeners on VCTK. For ease of comparison, 'heard' and 'unheard' utilize identical test data and the source and target samples are chosen at random from the pool of ten test speakers. Traditionally, prior works relied on subjects to assign scores to audio clips, with each sample rated on a scale of 1 to 5 across various evaluation metrics. The final comparative metric is determined by the mean opinion score (MOS) of these audio samples. However, the MOS scoring process can be intricate and potentially swayed by past data. As a solution, we employ a straightforward

| Method | Heard | Quality↑ | Similarity↑ |
|---|---|---|---|
| Ground truth | - | 4.800 | - |
| StarGANv2-VC | ✓ | 1.619 | 2.128 |
| SGAN-VC | × | 2.849 | 1.963 |
| SGAN-VC | ✓ | **3.248** | **3.173** |

Table 3.3: Qualitative evaluation on the VCTK Corpus (§3.3.5).

and efficient qualitative evaluation method. Specially, we curate a questionnaire where each conversion set comprises the original speech, the target speech, the results from the SOTA models, our conversion outputs, and the textual content of the source speech. Participants are asked to rank these based on two primary criteria: the **quality** of speech audio and the style **similarity** to the target speech. The assessment of quality focuses on three facets: the level of noise, clarity of content, and the naturalness of the speech. In terms of quality evaluation, considering the inclusion of the source speech, scores can range from a high of 5 to a low of 1. For the similarity assessment, the score spectrum lies between 1 and 4. Higher scores on both of these metrics naturally suggest more favorable results. To finalize our assessment, we average the scores of all samples.

Table 3.3 showcases the rankings of the converted samples as perceived by listeners. SGAN-VC yields a total of 100 samples and a group of 18 volunteers is enlisted. Table 3.3 presents the average scores across these metrics. A higher score indicates a superior rank. For the heard data, SGAN-VC excels over StarGANv2-VC in both quality and similarity. Given that the dataset volume in the unheard experiment considerably surpasses the heard data, the speech quality still trumps StarGANv2-VC. Notably, in terms of similarity, SGAN-VC based on unheard data is nearly on par with StarGANv2-VC, which signifies that SGAN-VC retains impressive generalization abilities even for unfamiliar speakers.

**Results of Different Conversion Types.** Table 3.4 reports the results for various transformation types within the VCTK Corpus test set. Due to the limited samples for each conversion type, $mF0_{diff}$ is not computed in this context. We primarily focus on the similarity of conversion style and content comprehensibility. Thus, we emphasize the CLS and CER metrics. Here, $F2F$ and $M2M$ represent same-gender conversions: $F2F$ indicates conversions between female speakers, while $M2M$ denotes those between male speakers. Likewise, $F2M$ and $M2F$ are used for cross-gender conversions. A consistent trend observed across all methods is that the CLS for $F2M$ exceeds that of $M2F$ in the heard results. As illustrated

| Method | Type | CLS↑ | CER↓ |
|---|---|---|---|
| Ground truth | F | 93.60 % | 5.69% |
| | M | 99.60 % | 4.85% |
| StarGANv2-VC-Heard | F2F | 95.36 % | 30.69 % |
| | F2M | 94.44 % | 39.85 % |
| | M2F | 92.74 % | 34.89 % |
| | M2M | 93.65 % | 36.25 % |
| SGAN-VC-Heard | F2F | **95.64** % | **19.60** % |
| | F2M | **99.20** % | **22.85** % |
| | M2F | **95.58** % | **21.43** % |
| | M2M | **100.00** % | **19.29** % |
| SGAN-VC-Unheard | F2F | 30.95 % | 22.75 % |
| | F2M | 24.50 % | 30.92 % |
| | M2F | 28.11 % | 26.34 % |
| | M2M | 27.20 % | 21.70 % |

Table 3.4: Results of four conversion types (§3.3.6). $M$ and $F$ represent male and female, respectively. $F2M$ denotes a female source and a male target speaker, similar for others.

in Table 3.4, our SGAN-VC-Heard significantly surpasses StarGANv2-VC across all voice conversion types. Particularly, our model shows a remarkable advantage of roughly 5% over StarGANv2-VC in the $F2M$ conversion.

**Visualization.** Fig. 3.4 presents the visual representation of the converted melspectrograms from the VCTK test set. The topmost row highlights the self-reconstruction outcome, where the mel-spectrogram is reconstructed leveraging both its content and style features. By comparing the spectrograms of different types, it is evident that SGAN-VC has a high timbre similarity and effectively preserves the linguistic information.

### 3.3.6 Ablation Studies

**Number of Subbands.** As displayed in Table 3.5, the value of $n$ denotes the number of subbands into which the mel-spectrogram is segmented. At $n = 1$, SGAN-VC employs the entire mel-spectrogram for the conversion process. When the mel-spectrogram remains undivided,

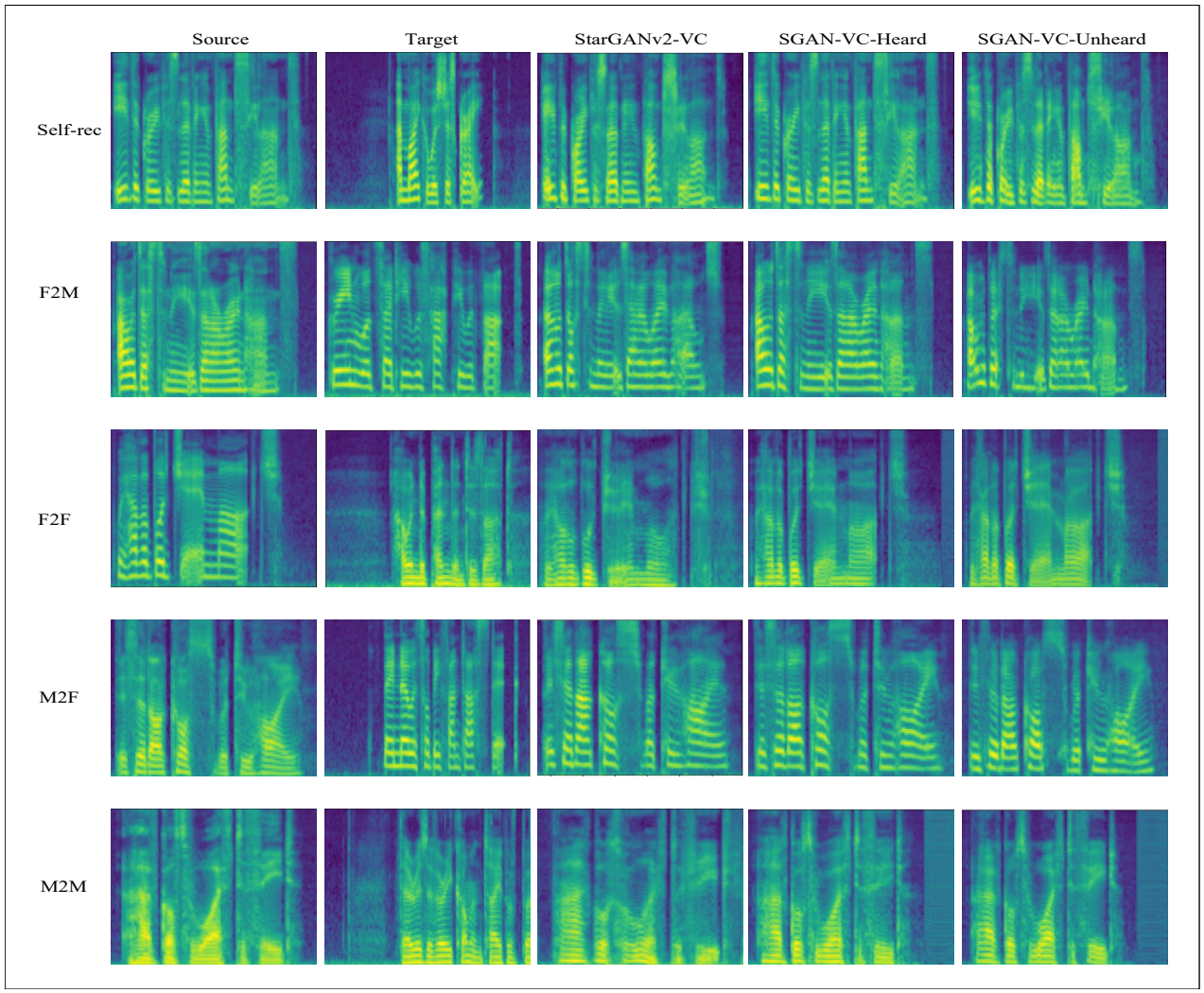


Figure 3.4: Visualization of different voice conversion types in SGAN-VC (§3.3.5).

SGAN-VC focuses on the comprehensive style data, with no localized details interfering with the resultant style. Although achieving a decent level of style similarity 97.2%, other metrics, however, fall into the lower tier. Only the overarching style undergoes conversion, certain phonemes might not transition naturally, leading to diminished intelligibility. The outcomes for CER and $mF0_{diff}$ under this setting are not commendable. For $num = 2$, where the spectrogram is split merely into two subbands, the localized information is not adequately segmented, culminating in the least favorable outcomes across the three metrics. With an uptick in $n$, both the CER and $mF0_{diff}$ metrics display enhancement, attributed to the modeling of more granulated details. Intuitively, a higher $n$ equips the model to discern finer features. These intricate details lend greater authenticity and precision to the converted style. Nevertheless, a minor decline across all four metrics is noticeable when $n = 5$. Given these observations, we have chosen $n = 4$ as the default setting. This decision balances the

| Num | pMOS↑ | CLS↑ | CER↓ | $mF0_{diff}$ ↓ |
|---|---|---|---|---|
| Ground truth | 3.484 | 96.6 % | 5.27 % | - |
| 1 | 3.432 | 97.2 % | 22.70 % | 3.31 |
| 2 | 3.453 | 96.7 % | 39.33 % | 3.91 |
| 3 | 3.462 | 97.1 % | **18.62** % | 3.57 |
| 4 | **3.479** | **97.6** % | 20.78 % | **1.88** |
| 5 | 3.464 | 96.1 % | 21.30 % | 2.69 |

Table 3.5: Ablation study of the number of subbands on VCTK (§3.3.6). "num" indicates how many subbands we divide the spectrograms.

extraction of detailed contextual information with an appropriate receptive field size.

Dividing the spectrogram into four subbands for timbre conversion inevitably increases computational complexity, as each sub-band requires separate processing. However, the actual increase in computation is relatively modest. This is because the additional operations, such as segmenting the spectrogram and independently processing each band, are straightforward and efficiently parallelizable. Importantly, the performance gains in terms of improved timbre quality and enhanced detail preservation across frequency bands outweigh the slight increase in computation.

**Pitch-shift Module.** As illustrated in Table 3.6, the inclusion of the pitch-shift module enhances the performance across all metrics for experiments on heard data. Likewise, when evaluating on unheard data, the integration of the pitch-shift module brings about improvements in the three metrics: pMOS, CLS, and CER. This suggests that the pitch-shift module can refine the pitch during the generation. However, it is notable that the $mF0_{diff}$ metric does not show improvement in the experiments with unheard data. This can be attributed to the fact that in the unheard data scenario, the speakers in the test set have never been encountered during training. Thus, the parameters that are learned by the pitch-shift module may not align perfectly with the characteristics of the test data, potentially leading to a minor decline in performance accuracy.

**Audio Duration.** As illustrated in Table 3.7, the quality of the converted sound diminishes when the audio clip is either too short or excessively long. For brief audio samples, the model struggles to produce appropriate reference conversions, largely because of the insufficient linguistic and stylistic data present. Conversely, when audio clips are excessively lengthy, the model is burdened with too much content to process effectively. This overloading can lead to suboptimal integration of style and content information.

| Pitch-shift | Heard | pMOS↑ | CLS↑ | CER↓ | $mF0_{diff}$ ↓ |
|---|---|---|---|---|---|
| Ground truth | - | 3.484 | 96.60 % | 5.27% | - |
| × | × | 3.531 | 26.10 % | 17.74 % | 4.18 |
| × | ✓ | 3.468 | 95.00 % | 24.28 % | 2.81 |
| ✓ | × | 3.595 | 27.70 % | 25.42 % | 5.13 |
| ✓ | ✓ | 3.479 | **97.60** % | 20.78 % | **1.88** |

Table 3.6: Ablation of pitch-shift module on VCTK dataset (§3.3.6).

| Method | pMOS↑ | CLS↑ | CER↓ | $mF0_{diff}$ ↓ |
|---|---|---|---|---|
| Ground truth | 3.484 | 96.60 % | 5.27% | - |
| 2s | 3.308 | 91.80 % | 67.70 % | 7.52 |
| 2.6s | **3.479** | **97.60** % | **20.78** % | **1.88** |
| 3s | 3.390 | 97.30 % | 79.61 % | 7.47 |

Table 3.7: Ablation of audio duration on VCTK dataset (§3.3.6).

## 3.4 Conclusion

We introduce SGAN-VC, a subband-focused generative adversarial network tailored for non-parallel many-to-many voice conversion, which sets new benchmarks in naturalness, content clarity, and style resemblance. Unlike traditional methods, SGAN-VC independently transfers content across each frequency band from the source to the target style. With subbands generated independently, our generative approach accentuates the unique timbre differences amongst speakers. Complementing this, our pitch-shift module meticulously further adjusts the pitch of speakers. Our comprehensive tests underscore the adaptability and broad applicability of SGAN-VC. Remarkably, on both Mandarin and English datasets, SGAN-VC has achieved competitive performance. Notably, in terms of style similarity, SGAN-VC trained on unheard data nearly mirrors the performance of StarGANv2-VC trained on heard data. Looking ahead, our endeavors will be geared towards enhancing the style similarity of unheard speakers even when constrained by limited training datasets.

CHAPTER 4

# MUTUAL LEARNING FOR ACOUSTIC MATCHING AND DEREVERBERATION

In the ever-evolving field of audio processing, the technique of audio style transfer represents a significant stride towards creative digital audio manipulation. This chapter delves into the transformative process of audio reverberation style transfer, a method that enables the modification of a target audio signal to reflect the style of another, while preserving its linguistic content. Drawing inspiration from the success of neural style transfer in images, this approach applies similar principles to audio features, facilitating new possibilities in music production, speech synthesis, and beyond.

Reverberation characteristics in audio styles are one of the core factors influencing auditory perception. In related tasks, *Visual acoustic matching (VAM)* involves combining visual cues, such as images or videos, with audio data, like reverberation and spectrum, to infer or synthesize audio properties consistent with the visual scene [30]. It is pivotal for enhancing the immersive experience, and the task of *dereverberation* is effective in improving audio intelligibility. However, existing methods treat each task independently, overlooking the inherent reciprocity between them. Moreover, these methods depend on paired training data, which is challenging to acquire, impeding the utilization of extensive unpaired data. In this chapter, we introduce MVSD[1], a mutual learning framework based on diffusion models. MVSD considers the two tasks symmetrically, exploiting the reciprocal relationship to facilitate learning from inverse tasks and overcome data scarcity. Furthermore, we employ

[1]This chapter is based on collaborative research [156] with Jian Ma, Wenguan Wang, Yi Yang and Feng Zheng, presented primarily as it appears in the ECCV 2024 proceedings.

the diffusion model as foundational conditional converters to circumvent the training instability and over-smoothing drawbacks of conventional GAN architectures. Specifically, MVSD employs two converters: one for VAM called reverberator and one for dereverberation called dereverberator. The dereverberator judges whether the reverberation audio generated by reverberator sounds like being in the conditional visual scenario, and vice versa. By forming a closed loop, these two converters can generate informative feedback signals to optimize the inverse tasks, even with easily acquired one-way unpaired data. Extensive experiments on two standard benchmarks, *i.e.*, SoundSpaces-Speech and Acoustic AVSpeech, exhibit that our framework can improve the performance of the reverberator and dereverberator and better match specified visual scenarios. Remarkably, the performance of the models can be further enhanced by adding more unpaired data.

## 4.1 Introduction

Sound interacts with its environment, giving listeners a sense of objects and spatial imprints [258]. Reverberation is the persistence of sound in a spatial environment, caused by the reflectivity of the surfaces of the objects and the slow propagation of sound in the air [69, 124]. The 'room divergence effect', stemming from disparities in acoustic reverberation properties between the virtual space and the real environment, can disrupt the user's immersion [5, 232]. Thus, reverberant sound, faithfully replicating real-world acoustics, is vital for realistic and immersive experiences in applications like augmented and virtual reality [112, 160]. Although reverberation can bestow a realistic sense of space, it may make speech content less intelligible [122, 216]. In line with human perception, automatic speech recognition systems also suffer from lower accuracy when processing reverberant speeches [58, 79, 270]. Therefore, dereverberation techniques [179] can benefit applications such as teleconferencing, hearing aids, voice assistants, *etc.* Existing works train VAM and dereverberation separately [30, 32, 66, 71, 240, 248]. The traditional methods of acoustic matching primarily involve unraveling the spatial characteristics of sound through the examination of Room Impulse Responses (RIRs), which assess the propagation and variation of sound within a specific environment [15, 24, 68, 178, 231, 248]. Rather than estimating RIRs, VAM [30] directly achieves specified reverberation by employing images of the target environment and original audio clips. For dereverberation, classical methodologies often encompass the application of signal processing and statistical techniques [180, 181], recent advances highlight neural network-based approaches that learn transfer functions from reverberation to anechoic spectrograms [58, 67, 242, 271]. Nonetheless, optimizing each

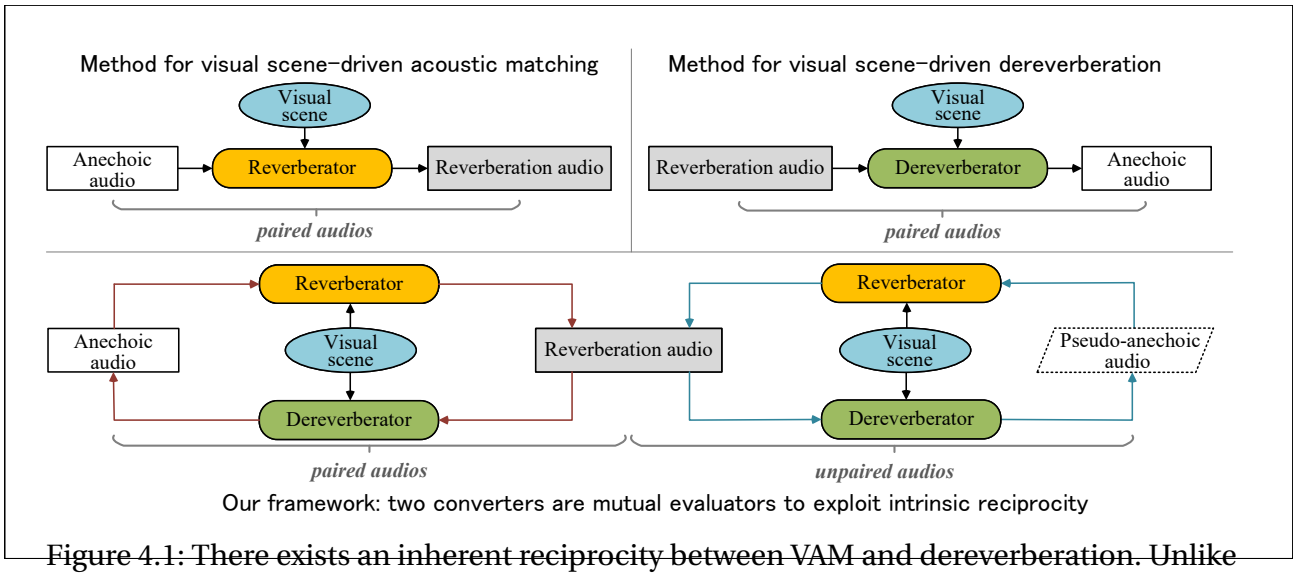


Figure 4.1: There exists an inherent reciprocity between VAM and dereverberation. Unlike previous approaches that treat these two tasks independently, our framework simultaneously handles the both tasks. Forming a closed loop between the two converters can generate informative feedback signals to optimize the inverse tasks, even with easily acquired one-sided unpaired data (§4.1).

task individually fails to leverage the inherent reciprocity between the two tasks (Fig. 4.1). Moreover, training these methods commonly requires extensive paired data. Yet capturing large volumes of aligned anechoic and reverberant audio pairs in real-world scenarios, whether in dereverberation or VAM, is not feasible. For VAM, the shortage of paired audio usually leads to average-style reverberation. When it comes to dereverberation, the model struggles to produce highly 'clean' audio in response to complex scenarios. Natural audio recordings often include various levels of reverberation. This variability poses a significant challenge for existing methods, which commonly struggle to exploit the large amount of easily available one-sided unpaired audio data.

In this chapter, we consider dereverberation as the inverse task of VAM, serving as an evaluator to provide feedback signals for VAM training, and vice versa. Specifically, given a visual environment $\boldsymbol{v}$, an anechoic audio $\boldsymbol{a}_c$, and a reverberant audio $\boldsymbol{a}_r$, VAM reverberator $f_\theta(\boldsymbol{v}, \boldsymbol{a}_c) \rightarrow \hat{\boldsymbol{a}}_r$ maps the visual observation and anechoic audio into reverberant audio, while the dereverberator $g_\phi(\boldsymbol{v}, \boldsymbol{a}_r) \rightarrow \hat{\boldsymbol{a}}_c$ restores reverberant audio to anechoic audio conditioned on visual characteristics. There exists a solid reciprocal relationship between the input and output spaces of $f_\theta$ and $g_\phi$. In this study, we delve into exploiting their intrinsic reciprocity to overwhelm the scarcity of parallel data. We propose a Mutual learning mechanism based on Visual Scene-driven Diffusion (MVSD) (Fig. 4.1). In MVSD, two converters, namely reverberator and dereverberator, are employed and capable of learning from the symmetric tasks. Taking VAM as an example, the reverberator, conditioned on the visual scene $\boldsymbol{v}$, simulates

environmental acoustic effects and converts anechoic audio $\boldsymbol{a}_c$ to reverberant audio $\hat{\boldsymbol{a}}_r$. Since the output of one converter can be used as the input for another, the reverberator and dereverberator can act as mutual evaluators. Concretely, in the primal task VAM, the reverberator generates reverberated audio $\hat{\boldsymbol{a}}_r$ conditioned on the visual scene $\boldsymbol{v}$ and anechoic audio $\boldsymbol{a}_c$. Then the reverse converter $g_\phi$ takes $\hat{\boldsymbol{a}}_r$ as input and reconstructs the anechoic audio $\tilde{\boldsymbol{a}}_c$ within the symmetric dereverberation task. Finally, the errors between $\tilde{\boldsymbol{a}}_c$ and $\boldsymbol{a}_c$ are used as feedback signals to optimize reverberator $f_\theta$, and vice versa. The training process of reverberator $f_\theta$ and dereverberator $g_\phi$ can form a closed loop, providing feedback for inverse tasks to enhance data efficiency. When the dereverberator encounters a unpaired natural audio $\boldsymbol{a}'_r$ with reverberation, it first eliminates the reverberation factors and creates a pseudo-anechoic audio $\hat{\boldsymbol{a}}'_c$. Likewise, the reverberator regenerates $\tilde{\boldsymbol{a}}'_r$ based on $\hat{\boldsymbol{a}}'_c$ and visual observations $\boldsymbol{v}'$. Hence, MVSD allows these two converters to benefit from each other's training instances and can be extended to easily acquired unpaired audio samples. For conditional generation, the architecture built on GANs is presently the prevailing choice [34, 75, 108, 109, 168, 209]. However, the training of GAN may introduce potential risks of instability and over-smoothing. Diffusion model [12, 40, 53, 89, 144, 152, 165, 219] recently show remarkable milestones in image generation, enabling the creation of high-quality images based on conditioning cues. Some works introduce diffusion into audio generation, such as converting spectrograms into sound signals [121], generating symbolic music [169], *etc.*However, diffusion generation of specified reverberation styles under visual guidance remains underexplored. To bridge this gap, we meticulously devise a visual scene-driven diffusion model to mitigate the computational overhead. Specifically, the diffusion model for each task includes a visual scene encoder for extracting features to control reverberation style, and a controllable Unet that serves as the generator for producing the desired audio. Additionally, we adopt cross-modal attention in selective blocks to establish correlations between visual cues and audio, reducing computational demands.

The reverberators and dereverberators used in visual acoustic matching rely on visual information to provide critical spatial context, such as room geometry, size, material properties, and object placement. These factors are essential for accurately modeling acoustic properties like reverberation and reflection. Without visual inputs, these tools face ambiguity in estimating the environment, leading to less precise sound predictions and mismatched audio outputs. Visual data also ensure consistency between audio and visual modalities, particularly in VR, AR, and media production, where synchronized audio-visual experiences are vital. Without visual guidance, these systems require more extensive datasets and become less generalizable, compromising their effectiveness in diverse acoustic scenarios.

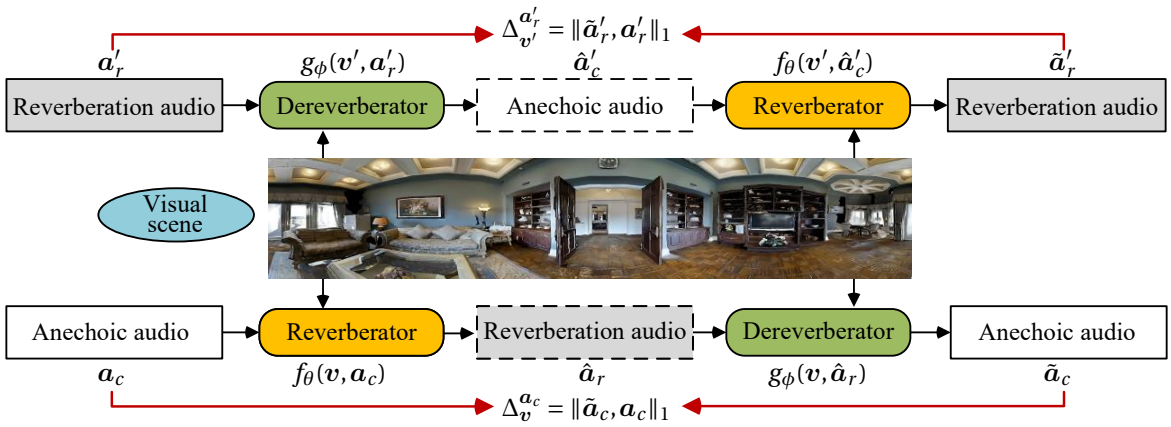


Figure 4.2: The overview of MVSD. The output of a converter can serve as pseudo-input for the reverse task, providing an intermediate transition. Concretely, the reverberator $f_\theta$ and dereverberator $g_\phi$ can generate feedback signals $\mathcal{L}_m$ (Eq. 4.4) for mutual optimization of training, even with one-way unpaired data $(a'_r, v')$ (§4.2.1).

We spotlight the notable strengths of MVSD in visual-audio cross-modal style transfer. MVSD effectively enhances the performances and consistently reports promising results on both tasks. We achieve a remarkable reduction of 0.157 in STFT-distance on the 'Seen' test set of SoundSpaces-Speech [30] (23.6% relative performance). Moreover, MVSD successfully reduces WER by 1.4% to an impressive 11.63% in the dereverberation task, showcasing an outstanding relative improvement of 8% over SOTA methods. Notably, the utilization of unpaired audios can further boost the relative performance by 9.1% in RTE for VAM.
To summarize, the main contributions of this chapter are as follows:

- We initially propose an end-to-end approach that leverages the reciprocity between VAM and dereverberation tasks to reduce reliance on paired data.
- We introduce a novel and elegant mutual learning framework, MVSD, incorporating diffusion models and utilizing symmetrical tasks as evaluators to provide feedback signals to facilitate model training.
- We conduct a comprehensive evaluation of MVSD, demonstrating its superiority and confirming the potential of unpaired data in real-world applications.

## 4.2 Method

We propose a mutual learning framework MVSD to leverage feedback signals from symmetrical tasks. It can promote model training and better exploit unpaired data. MVSD involves

two tasks: a primal task VAM [30] that employs the reverberator $f_\theta$ to convert an anechoic audio $\boldsymbol{a}_c$ into a reverberated audio $\hat{\boldsymbol{a}}_r$, which is aurally recorded in the specified environment. In the dual task, dereverberator $g_\phi$ removes the reverberant characteristics in $\boldsymbol{a}_r$. Here, $f_\theta$ and $g_\phi$ are jointly trained in an end-to-end mutual learning framework MVSD (§4.2.1). Furthermore, we employ visual scene-driven diffusion models as foundational conditional converters $f_\theta$ and $g_\phi$ to achieve stable and accurate reverberation style transfer (§4.2.2).

**Reverberator.** Consider paired data distributions: $\boldsymbol{A}_c = \{\boldsymbol{a}_c^{(1)}, \boldsymbol{a}_c^{(2)}, \boldsymbol{a}_c^{(3)}, ..., \boldsymbol{a}_c^{(n)}\}$ and $\boldsymbol{A}_r = \{\boldsymbol{a}_r^{(1)}, \boldsymbol{a}_r^{(2)}, \boldsymbol{a}_r^{(3)}, ..., \boldsymbol{a}_r^{(n)}\}$, representing anechoic and reverberant audio, respectively. The set of visual scenes $\boldsymbol{V} = \{\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, ..., \boldsymbol{v}^{(n)}\}$ corresponds to the audio set of $\boldsymbol{A}_r$. The goal of VAM is to convert the anechoic audio $\boldsymbol{a}_c$ with condition $\boldsymbol{v}$ to its reverberant counterpart $\boldsymbol{a}_r$, *i.e.*, to estimate the conditional distribution $f_\theta(\boldsymbol{a}_r|\boldsymbol{a}_c; \boldsymbol{v})$. Based on diffusion models, we encode $\boldsymbol{a}_c$ into content features and switch the reverberation style to the visual environment $\boldsymbol{v}$.

**Dereverberator.** Contrary to VAM, we can regard dereverberation as a process of audio denoising. The goal of the dereverberation task is to eliminate reverberation factors and enhance the intelligibility of audio content. Correspondingly, the dereverberator $g_\phi$ based on VSD calculates the anechoic distribution $g_\phi(\boldsymbol{a}_c|\boldsymbol{a}_r; \boldsymbol{v})$ under a given scene $\boldsymbol{v}$.

### 4.2.1 Mutual Learning

We jointly learn the VAM and dereverberation tasks (Fig. 4.2): the reverberator $f_\theta$ and dereverberator $g_\phi$ can mutually benefit from each other. Suppose we have two (vanilla) converters that can map anechoic audio to a specified reverberation style and vice versa. Our goal is to simultaneously improve the style accuracy of the VAM task and the content intelligibility of the dereverberation task by employing paired and unidirectional non-paired data. To achieve this, we leverage the reciprocity between these two tasks, wherein the input-output spaces of VAM and dereverberation exhibit a strong correlation and can interchangeably act as the input and output for each other. Starting from either task, we first convert it forward to another audio, then transfer it backward to the original audio. By evaluating the results of this two-hop transfer process, we can gauge the quality of both converters and optimize them accordingly. Namely, dereverberator $g_\phi$ is employed to evaluate the quality of $\hat{\boldsymbol{a}}_r$ generated by $f_\theta$ and sends back an error singal $\triangle(\tilde{\boldsymbol{a}}_c, \boldsymbol{a}_c)$ to $f_\theta$, and vice versa. This process can be iterated many rounds until both converters converge. Please note that in MVSD, $\boldsymbol{a}_r$ and $\boldsymbol{a}_c$ are not necessarily aligned and may even not have a typical relationship.

We denote a labeled collection as $\mathscr{D} = \{(\boldsymbol{v}^n, \boldsymbol{a}_c^n, \boldsymbol{a}_r^n)\}_{n=1}^N$, which consists of $N$ aligned tuples of anechoic and reverberant audio. Given a triplet $\langle \boldsymbol{v}, \boldsymbol{a}_c, \boldsymbol{a}_r \rangle$, where $\boldsymbol{v}$, $\boldsymbol{a}_c$, $\boldsymbol{a}_r$ are sets of *environmental spaces*, *anechoic* and *target audios*. Our goal is to uncover the bi-

directional relationship between the $\boldsymbol{a}_c$ and $\boldsymbol{a}_r$. For the primal process starting from VAM, denote $\hat{\boldsymbol{a}}_r$ as the mid-transition output. Firstly, we obtain a reverberated audio $\hat{\boldsymbol{a}}_r$ through the reverberator $f_\theta(\boldsymbol{v}, \boldsymbol{a}_c)$. Then, the dereverberator $g_\phi$ translates $\hat{\boldsymbol{a}}_r$ to $\tilde{\boldsymbol{a}}_c$ by mapping $g_\phi(\boldsymbol{v}, \hat{\boldsymbol{a}}_r)$. The $\tilde{\boldsymbol{a}}_c$ is expected to be consistent with $\boldsymbol{a}_c$ in audio clarity, *i.e.*, achieving a small cycle-consistency error $\Delta_{\boldsymbol{v}}^{\boldsymbol{a}_c}$. Similarly, for dereverberator $g_\phi$, we have $\tilde{\boldsymbol{a}}_r = f_\theta(\boldsymbol{v}, \hat{\boldsymbol{a}}_c)$ and $\tilde{\boldsymbol{a}}_r$ should have a reverberation effect akin to $\boldsymbol{a}_r$ in auditory perception. Likewise, $\Delta_{\boldsymbol{v}}^{\boldsymbol{a}_r}$ can be employed to evaluate the discrepancies between $\boldsymbol{a}_r$ and $\tilde{\boldsymbol{a}}_r$. Finally, the errors $\Delta_{\boldsymbol{v}}^{\boldsymbol{a}_c}$ and $\Delta_{\boldsymbol{v}}^{\boldsymbol{a}_r}$ can be specified as two reconstruction losses, which are minimized for the model training. Prior researches [83, 126] on conditional image synthesis suggest that $L1$ distance, unlike $L2$, can reduce blurriness. Hence, we employ $L1$ distance to assess the feedback errors:

$$\begin{aligned} \Delta_{\boldsymbol{v}}^{\boldsymbol{a}_r} &= \|\tilde{\boldsymbol{a}}_r, \boldsymbol{a}_r\|_1 = \|f_\theta(\boldsymbol{v}, g_\phi(\boldsymbol{v}, \boldsymbol{a}_r)) - \boldsymbol{a}_r\|_1; \\ \Delta_{\boldsymbol{v}}^{\boldsymbol{a}_c} &= \|\tilde{\boldsymbol{a}}_c, \boldsymbol{a}_c\|_1 = \|g_\phi(\boldsymbol{v}, f_\theta(\boldsymbol{v}, \boldsymbol{a}_c)) - \boldsymbol{a}_c\|_1. \end{aligned} \tag{4.1}$$

In real-world scenarios, the challenge of capturing parallel data arises from the difficulty of simultaneously recording sound at the source and listener locations. This obstacle is mitigated in our approach, as it does not necessitate aligned anechoic and reverberant pairs $(\boldsymbol{a}_c, \boldsymbol{a}_r)$ for the errors $\Delta_{\boldsymbol{v}}^{\boldsymbol{a}_r}$ and $\Delta_{\boldsymbol{v}}^{\boldsymbol{a}_c}$, As a result, Eq. 4.1 can be effectively applied to one-way unpaired audios. As in common practice [65, 110, 252], we build two unlabeled collections: $\mathscr{U} = \{(\boldsymbol{v}^{m\prime}, \boldsymbol{a}_r^{m\prime})\}_{m=1}^{M}$ for audios with natural reverberation, and $\mathscr{C} = \{\boldsymbol{a}_c^{k\prime\prime}\}_{k=1}^{K}$ with only anechoic audio. We obtain $\mathscr{U}$ by sampling natural audios $\boldsymbol{a}_r'$ from existing environments $\boldsymbol{v}'$, which lack corresponding anechoic audios. Similarly, we create collection $\mathscr{C}$ by filtering anechoic audios $\boldsymbol{a}_c''$ from an open-source dataset [192], which do not have matching reverberated audios and visual images. For unpaired natural audios $\mathscr{U}$, we first generate intermediate output $\hat{\boldsymbol{a}}_c'$ using the dereverberator $f_\theta(\boldsymbol{a}_r', \boldsymbol{v}')$, followed by reconstructing $\tilde{\boldsymbol{a}}_r'$ based on $\hat{\boldsymbol{a}}_c'$ and scene $\boldsymbol{v}'$, *i.e.*, $g_\phi(\boldsymbol{v}', \boldsymbol{a}_r')$, and computing error $\Delta_{\boldsymbol{v}'}^{\boldsymbol{a}_r'}$ against the original input $\boldsymbol{a}_r'$. We can derive the formula:

$$\Delta_{\boldsymbol{v}'}^{\boldsymbol{a}_r'} = \|\tilde{\boldsymbol{a}}_r', \boldsymbol{a}_r'\|_1 = \|f_\theta(\boldsymbol{v}', g_\phi(\boldsymbol{v}', \boldsymbol{a}_r') - \boldsymbol{a}_r'\|_1. \tag{4.2}$$

As for unpaired anechoic audios $\mathscr{C}$, since there are rarely accompanying visual scene images when recording audio, we randomly sample an image $\boldsymbol{v}''$ from $\mathscr{U}$ to simulate the specified environment, and the formulate is as:

$$\Delta_{\boldsymbol{v}''}^{\boldsymbol{a}_c''} = \|\tilde{\boldsymbol{a}}_c'', \boldsymbol{a}_c''\|_1 = \|g_\phi(\boldsymbol{v}'', f_\theta(\boldsymbol{v}'', \boldsymbol{a}_c'')) - \boldsymbol{a}_c''\|_1. \tag{4.3}$$

Our training process utilizes paired data, complemented by unpaired natural and anechoic audios. Consequently, our mutual learning loss is defined as:

$$\mathscr{L}_m = \frac{1}{N}\sum\nolimits_{(\boldsymbol{v},\boldsymbol{a}_r,\boldsymbol{a}_c)\in\mathscr{D}}(\Delta_{\boldsymbol{v}}^{\boldsymbol{a}_c} + \Delta_{\boldsymbol{v}}^{\boldsymbol{a}_r}) + \frac{1}{M}\sum\nolimits_{(\boldsymbol{v}',\boldsymbol{a}_r')\in\mathscr{U}}\Delta_{\boldsymbol{v}'}^{\boldsymbol{a}_r'} + \frac{1}{K}\sum\nolimits_{(\boldsymbol{v}'',\boldsymbol{a}_c'')\in\mathscr{C}}\Delta_{\boldsymbol{v}''}^{\boldsymbol{a}_c''}. \tag{4.4}$$

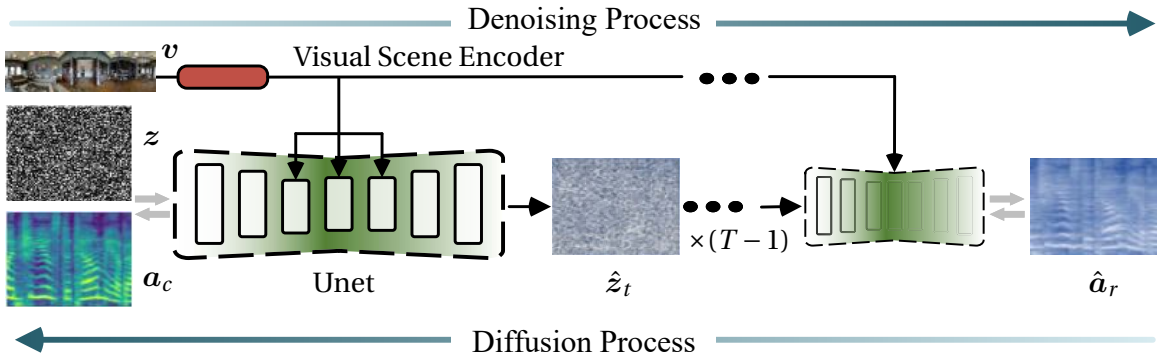


Figure 4.3: The diffusion and denoising processes of VSD. Taking VAM as an example, MVSD converts anechoic audio $\boldsymbol{a}_c$ into reverberant audio $\hat{\boldsymbol{a}}_r$ that aligns with the acoustics of the visual scene $\boldsymbol{v}$ (§4.2.2).

During training, $\mathscr{L}_m$ is applied only for predictions and backpropagation at time step $t$ of the diffusion model. Hence, MVSD does not significantly increase the training time compared to training the two tasks separately.

**Remark.** MVSD consists of two main concepts: First, an ideal reverberator should be able to adapt audio to any visual environment, and a dereverberator is also effective at removing disturbances that affect speech intelligibility. Therefore, we investigate VAM and dereverberation in a unified learning framework, allowing the converters to better exploit the cross-modal and cross-task correlations. Second, we utilize paired and unpaired data in training (Eq. 4.4), and they can provide complementary benefits. The addition of unpaired data can boost model performance, and paired data guides the reverberator and dereverberator converge to the target distribution, preventing extreme domain deviation from the unpaired data. The two converters are optimised jointly and collaboratively, with mutual and close intervention between them. Furthermore, all parts within the loop are differentiable, thus allowing the end-to-end training of MVSD.

### 4.2.2 Visual Scene-driven Diffusion

Within the MVSD framework, both the reverberator and the dereverberator utilize a similar model structure. We introduce visual scene-driven diffusion (VSD) with the reverberator $f_\theta$ as an example. The diffusion model employs a $T$-step iterative denoising process to transform Gaussian noise into the desired data distribution [89, 219, 244]. By introducing prompt conditions such as class labels and text [53, 183], the generated content can be controlled precisely. In MVSD, visual scene embeddings are employed as control conditions to guide the generation of reverberator $f_\theta$ and dereverberator $g_\phi$. In particular, the diffusion process fol-

lows a Markov chain, progressively adding noise to the input spectrogram $\boldsymbol{x}_0$ (sampled from the real distribution $q(\boldsymbol{x})$ until it evolves into white Gaussian noise $\mathcal{N}(0,1)$. At each step $t$, the spectrogram $\boldsymbol{x}_t$, following the distribution $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$, is derived by the pre-defined variance $\beta_t$ scaled with $\sqrt{1-\beta_t}$:

$$q(\boldsymbol{x_t}|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{z}_t); \quad \boldsymbol{z}_t \sim \mathcal{N}(\sqrt{1-\beta_t}\,\boldsymbol{z}_{t-1}, \beta_t \mathbf{I}). \tag{4.5}$$

The denoising process attempts to restore the original spectrogram $\boldsymbol{x}_0$ from the noisy data $\boldsymbol{x}_T$ by removing the noise introduced in the forward diffusion process. The prediction $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ at step $t-1$ is approximated by a parameterized model $p$ (*e.g.*, a neural network), involving the estimation of $\mu(\boldsymbol{z}_t, t)$ and $\sigma(\boldsymbol{z}_t, t)$ from a Gaussian distribution. By employing the reverse process across all time steps, we are able to make a transition. Starting from $\boldsymbol{x}_T$, we move backwards to reach the initial spectrogram $\boldsymbol{x}_0$:

$$\begin{aligned} p(\boldsymbol{x}_{0:T}) &= p(\boldsymbol{x}_T) \textstyle\prod_{t=1}^{T} p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \\ &= p(\boldsymbol{x}_T) \textstyle\prod_{t=1}^{T} \mathcal{N}(\boldsymbol{x}_{t-1}; \mu(\boldsymbol{x}_t, t), \sigma(\boldsymbol{x}_t, t)). \end{aligned} \tag{4.6}$$

**Visual Scene Encoder.** We apply an embedding with 256 dimensions to represent visual scenes, extracted by a pre-trained ResNet-18 [84] encoder. Then, the embedding serves as the condition to guide the generation of diffusion models.

**Controllable Unet.** We meticulously design an controllable Unet for predicting $\boldsymbol{x}_t$ of diffusion (Fig. 4.3). Controllable Unet is composed of multiple stages with attention blocks [219], *i.e.*, self-attention and cross-attention. Self-attention allows a model to weigh the importance of different parts within the same element. Cross-attention, similar to self-attention, targets relationships across different components. We employ a classic encoder-decoder with a symmetric design, where each part incorporating 3 attention blocks. The encoder progressively reduces the resolution of the feature map, and then the decoder gradually increases it to align with the size of the original spectrogram. In the self-attention block, we utilize the downsampling method in [251] with a stride of 4 to rapidly decrease the size of feature maps. The downsampling utilizes dilated convolutions and attention to increase the receptive field without reducing spatial dimensions. Cross-modal attention is selectively employed to the third encoder block and the first decoder block, mitigating computational overhead. Both VAM and dereverberation need to preserve the linguistic information in the audios. Therefore, we concatenate source spectrogram with the noise $\boldsymbol{z}_0$ as the content input for the controllable Unet.

While melspectrograms can be viewed as images, their characteristics differ significantly from natural images in terms of structure and frequency content. Using pretrained diffusion

models like stable diffusion [218] would require substantial adaptation and training to align with the unique properties of audio data. Due to computational constraints, we do not pursue this option. However, investigating pretrained diffusion models is an interesting direction for future work and may enhance audio generation performance.

### 4.2.3 Training Objective

For training the diffusion model, we employ the simplified objective [89]:

$$\mathscr{L}_d = \mathbb{E}_{\boldsymbol{x}_0,t,\boldsymbol{z}}[\|\boldsymbol{z} - \hat{\boldsymbol{z}}(\sqrt{\overline{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{z}, t)\|_2], \tag{4.7}$$

where $\alpha_t$ in diffusion models is a scaling factor that modulates the noise level at each time step $t$. VSD can predict the noise $\hat{\boldsymbol{z}}_t$ and use it to iteratively refine the denoising process. With the reparameterization trick, a method for differentiable sampling [115], we can represent the estimation of $\hat{\boldsymbol{x}}_0$:

$$\hat{\boldsymbol{x}}_0 = \frac{1}{\sqrt{\overline{\alpha}_t}}(\boldsymbol{x}_t - \sqrt{1-\overline{\alpha}_t}\hat{\boldsymbol{z}}_t). \tag{4.8}$$

Moreover, we introduce a style loss $\mathscr{L}_{sty}$ (Eq. 4.9) to make the generated audios with the environmental characteristics. Taking VAM task as an example, during training, the Unet predicts the noise $\hat{\boldsymbol{z}}_t$ at time step $t$. Then, $\hat{\boldsymbol{z}}_t$ can be used to gradually derive the predicted original spectrogram $\hat{\boldsymbol{x}}_r$ at step 0 (Eq. 4.8). Here, we do not explicitly extract the stylistic features of the $\boldsymbol{a}_r$ and $\hat{\boldsymbol{a}}_r$; instead, we directly employ $\mathscr{L}_1$ loss to regularize style consistency:

$$\mathscr{L}_{sty} = \|\hat{\boldsymbol{a}}_r - \boldsymbol{a}_r\|_1 + \|\hat{\boldsymbol{a}}_c - \boldsymbol{a}_c\|_1. \tag{4.9}$$

We learn models $f_\theta$ and $g_\phi$ by minimizing the combination of the diffusion loss, the style loss and the mutual learning regularization term. Moreover, we empirically set the weight of each loss term to 1. In summary, the overall training objective is given as:

$$\mathscr{L}_{total} = \mathscr{L}_d + \mathscr{L}_m + \mathscr{L}_{sty}. \tag{4.10}$$

### 4.2.4 Implementation Details

**Training.** In MVSD, converters and visual scene encoder are trained separately. We adopt the loss function in [111] to train the visual scene encoder. The mutual learning is integrated into each mini-batch update, spanning the entire training process for the two tasks. Training starts with supervised data, with unsupervised data progressively merged for optimization. This stepwise strategy can preserve model stability. At each iteration, we compute the predictions of both converters and update their parameters based on the feedback from the

symmetrical models. In practice, we first perform supervised training and conduct the loop of mutual learning (Alg. 1). Besides minimizing the cycle-consistent loss $\mathscr{L}_m$ (Eq. 4.4), our MVSD framework is learnt with the diffusion objectives for VAM and dereverberation, over the labeled data $\mathscr{D}$. After the model converges, we get a vanilla reverberator and dereverberator. Then, we build on the vanilla model and introduce unlabeled data for further training. Finally, we receive a prepared model when MVSD converges on all training data.

**Algorithm 1:** Mutual learning with visual scene-driven diffusion.

**Input:** Labeled set $\mathscr{D}$, unpaired sets $\mathscr{U}$ and $\mathscr{C}$, reverberator $f_\theta$ and dereverberator $g_\phi$.

***Repeat:***
- Sample a mini-batch of paired tuples $\langle \boldsymbol{a}_c, \boldsymbol{v}, \boldsymbol{a}_r \rangle$;
- Generate random Gaussian noise $\boldsymbol{z}_c$ and $\boldsymbol{z}_r$ for the converters;
- Execute the diffusion processes of reverberator $f_\theta$ and dereverberator $g_\phi$;
- Calculate the training objective $\mathscr{L}_{total}$ (Eq. 4.10);
- Update the parameters of $\theta$ and $\phi$: $\theta \leftarrow \theta - \gamma \nabla_\theta \mathscr{L}(\theta)$, $\phi \leftarrow \phi - \gamma \nabla_\phi \mathscr{L}(\phi)$;
- Introduce unpaired data and continue training when the epoch exceeds 100;

***Until:*** Convergence

**Inference.** The inference of each task follows the sampling process of the diffusion model. Take VAM as an example: First, a noise spectrogram is randomly generated and concatenated with a anechoic test spectrogram. Next, at each step $t$ of the denoising process, the controllable Unet synthesizes the intermediate spectrogram conditioned on visual features. The completion of the diffusion sampling process marks the generation of the desired spectrogram. Finally, we apply a Vocoder to convert the spectrogram to an audio signal.

**Reproductibility.** Our model is implemented in PyTorch and trained using two NVIDIA Tesla V100 GPUs[2]. Training MVSD from scratch takes approximately 144 hours. The average inference time is 1.09 seconds. We set FFT size, hop size and mel scale for audio processing to 1024, 256 and 128, respectively. We then truncate the mel-spectrogram to a width of 128, resulting in a size of $128 \times 128$. We utilize a pre-trained BigVGAN [129] as the vocoder.

## 4.3 Experiments

**Dataset.** We conduct experiments on two datasets [30]: *SoundSpaces-Speech Dataset* and *Acoustic AVSpeech*. The former employs a simulated environment [31] to generate reverberation audio, is perfectly aligned paired audio and accurate ground truth. Regardless, there has a realism gap. Considering the difficulty of obtaining depth information in the real

[2] https://hechang25.github.io/MVSD

| Method | *SoundSpaces-Speech* | | | | | | *Acoustic AVSpeech* | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Seen* | | | *Unseen* | | | *Seen* | | *Unseen* | |
| | STFT ↓ | RTE (s) ↓ | MOSE ↓ | STFT ↓ | RTE (s) ↓ | MOSE ↓ | RTE (s) ↓ | MOSE ↓ | RTE (s) ↓ | MOSE ↓ |
| Input audio | 1.192 | 0.331 | 0.617 | 1.206 | 0.356 | 0.611 | 0.387 | 0.658 | 0.392 | 0.634 |
| AEE [248] | 2.746 | 0.319 | 0.571 | - | - | - | - | - | - | - |
| Image2Reverb [240] | 2.538 | 0.293 | 0.508 | 2.318 | 0.317 | 0.518 | - | - | - | - |
| AV U-Net [71] | 0.638 | 0.095 | 0.353 | 0.658 | 0.118 | 0.367 | 0.156 | 0.570 | 0.188 | 0.540 |
| AViTAR [30] | 0.665 | 0.034 | 0.161 | 0.822 | 0.062 | 0.195 | 0.144 | 0.481 | 0.183 | 0.453 |
| MVSD *w/o* visual scene | 0.691 | 0.188 | 0.156 | 0.803 | 0.155 | 0.194 | 0.137 | 0.526 | 0.171 | 0.474 |
| MVSD *w/o* unpaired data | 0.573 | 0.033 | 0.148 | 0.736 | 0.055 | 0.184 | 0.131 | 0.427 | 0.159 | 0.394 |
| **MVSD** | **0.508** | **0.030** | **0.142** | **0.637** | **0.051** | **0.178** | **0.112** | **0.392** | **0.148** | **0.379** |

Table 4.1: Quantitative results on *SoundSpaces-Speech* and *Acoustic AVSpeech* [30] (§4.3.1).

world (which usually requires professional video equipment), we only apply RGB images to represent the visual environment. Finally, there are 28,853/1,441/1,489 samples for the *train/val/test* splits. Acoustic AVSpeech is a subset of AVSpeech dataset [57]. It offers more realism but poses evaluation challenges due to lacking corresponding reverberant audios. Acoustic AVSpeech contains $113k/3k/3k$ video clips for the training, validation, and test splits, respectively. For the unpaired data, we randomly sample 5,000 natural audios with video in AVSpeech dataset [57] and $5k$ anechoic audio in Librispeech [192], with no overlap with the training sets. We apply 'Seen' and 'Unseen' to denote whether visual scenes are encountered during training.

**Evaluation Metrics.** For VAM task, following [30], we employ STFT-distance to measure deviation from the ground truth, Reverberation Time 60 error (RTE) for room acoustics, and Mean Opinion Score Error (MOSE) for speech quality evaluation. For dereverberation task, as in [32], we adopt Perceptual Evaluation of Speech Quality (PESQ) [217], Word Error Rate (WER) and Equal Error Rate (EER) to assess the aspects of audio quality, content precision, and speaker verification error, respectively. The evaluation is conducted on the anechoic version of LibriSpeech test set [32, 192].

### 4.3.1 Performance on VAM

As shown in Table 4.1, our method demonstrates noteworthy improvements across all three metrics on both datasets. Specifically, MVSD achieves a notable absolute boost of 0.157 STFT-distance (23.6% relative improvement), 0.004 RTE (11.8% relative improvement), and

| | Speech Enhancement PESQ ↑ | Speech Recognition WER(%) ↓ | Speaker Verification EER(%) ↓ |
|---|---|---|---|
| Anechoic (Ceiling) | 4.64 | 2.50 | 1.89 |
| Reverberant | 1.54 | 8.86 | 5.23 |
| MetricGAN+ [66] | 2.33 | 7.49 | 5.16 |
| VIDA [32] | 2.37 | 4.44 | 4.58 |
| **MVSD** | **2.53** | **4.27** | **4.46** |

Table 4.2: Quantitative dereverberation results on SoundSpaces-Speech [30] (§4.3.2).

| | SoundSpaces | AVSpeech |
|---|---|---|
| Input Speech | 39.3% / **60.7%** | 38.2% / **61.8%** |
| Image2Reverb [240] | 20.8% / **79.2%** | - / - |
| AV U-Net [71] | 23.4% / **76.6%** | 21.9% / **78.1%** |
| AViTAR [30] | 34.7% / **65.3%** | 44.1% / **55.9%** |

Table 4.3: User study results. X%/Y% means that X% of participants prefer this method while Y% prefer MVSD (§4.3.3).

0.019 MOSE (11.8% relative improvement) in the 'Seen' split of SoundSpaces-Speech dataset compared with SOTA method. There is also a similar improvement in Acoustic AVSpeech dataset. We can see that MVSD exhibits outstanding strengths in all three assessed aspects: preserving source audio content better, getting in more precise signal attenuation and more consistent quality with the target audio. MVSD has the capability to infer and extract relevant factors that influence reverberation from target images, even in never-before-seen scenes. Our results also indicate that panoramic images of the visual environment can amplely represent acoustic properties. It should be noted that blind reverberator [248], a traditional acoustic method, needs reference audio, making it unsuitable for scenarios 'Unseen' of SoundSpaces (no reference audio) and AVSpeech, as reported in [30]. Fig. 4.4 showcases the visual comparisons of different methods for the VAM task on the SoundSpaces-Speech and AVSpeech datasets, respectively, highlighting the superiority of MVSD.

### 4.3.2 Performance on Derverberation

Table 4.2 presents the dereverberation performance of MVSD on SoundSpaces-Speech [30] dataset. We observe that MVSD also demonstrates superior performance across all three met-

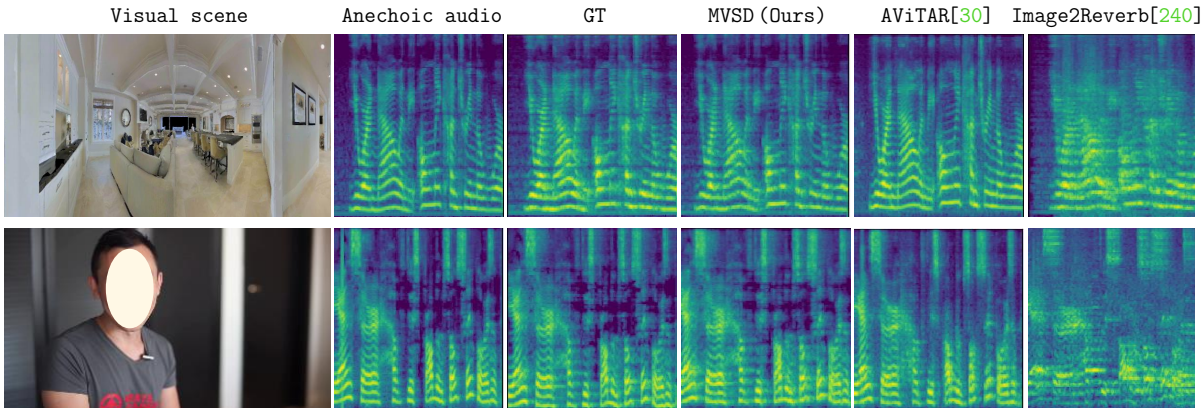


Figure 4.4: Visualization results for VAM task on the SoundSpaces-Speech (top) and AVSpeech datasets (bottom) [30] (§4.3.1).

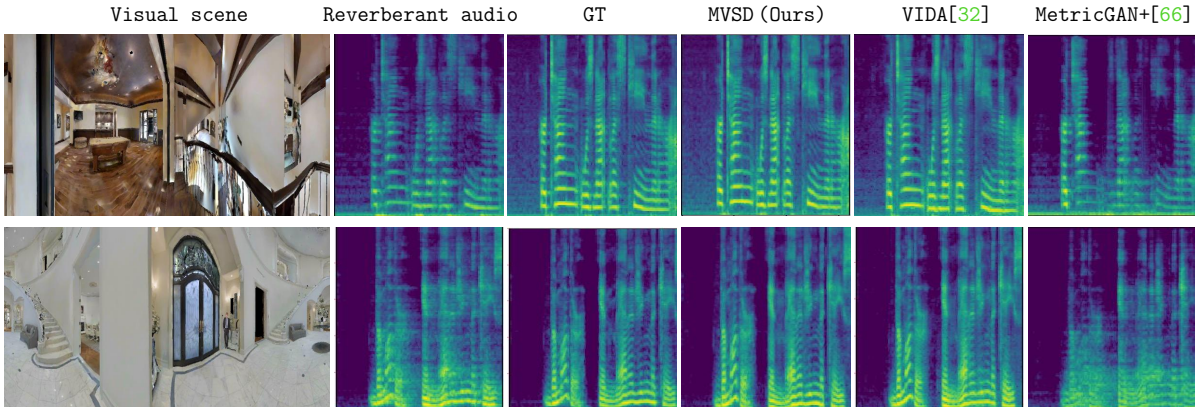


Figure 4.5: Visualization of the dereverberation task on SoundSpaces-Speech [30] (§4.3.2).

rics in the dereverberation task. Particularly in terms of WER, MVSD exhibits a remarkable error reduction of 0.17 compared to VIDA, achieving a value of 4.27%. This highlights the robust dereverberation capability of MVSD. Additionally, MVSD achieves an EER of 4.46%, demonstrating its ability to mitigate reverberation while preserving the timbre information. Fig. 4.5 depicts the spectrograms for the dereverberation task on SoundSpaces-Speech, with AVSpeech dataset omitted due to the absence of groundtruth anechoic audio. From the spectrogram, it is evident that MVSD demonstrate superior performance in both clarity and noise reduction for the dereverberation task.

### 4.3.3 User Study

The human ear is the most accurate tool for evaluating acoustic experiences. Therefore, we conduct a user study as a complement to quantitative indicators. We invited 15 volunteers to participate in the evaluation. Following the configuration in [30], we show participants

some images of the target environment, real audio clips, and samples generated by all test methods. Participants is asked to choose the audio sample that exhibit the highest consistency with the groundtruth reverb style and 30 samples are selected for each dataset. Table 4.3 reports the final preference scores. As expected, MVSD consistently exceeds the results of other methods, achieving a high preference ratio against AViTAR [30] (65.3% in SoundSpaces and 55.9% in AVSpeech). A certain percentage (39.3% and 38.2%) of subjects prefer 'clean' audio devoid of reverberation, which can be seen from the first row of Table 4.3. This tendency can be attributed to the general preference for 'clean' sounding audio among those participants without professional acoustical knowledge, which is also reported in [30].

### 4.3.4 Ablation Study

To assess the effectiveness of MVSD's key components, we conduct diagnostic studies and report VAM results on SoundSpaces-Speech [30] dataset and employing WER and PESQ metrics for dereverberation.

**Mutual Learning.** We conduct three diagnostic experiments: i) VSD – training two tasks separately with the structure akin to mutual learning; ii) MVSD *w/o* unpaired data – training exclusively with labeled data; and iii) MVSD – augmenting the second experiment with unpaired data. Table 4.4 reveals that our baseline model VSD can achieve performance matching SOTA on metrics STFT 0.657 and MOSE 0.159. The introduction of MVSD results in a slight edge over SOTA, and incorporating unpaired data notably surpasses SOTA in both tasks (achieved 0.508 STFT, 0.030 RTE, 0.142 MOSE in VAM, and 4.27% WER, 2.53 PESQ in dereverberation, respectively). These findings emphasize the synergy between dual tasks, which can greatly enhance learning capabilities. This cooperation helps in efficiently processing unpaired data, showcasing the value of a wealth of natural data.

**Model Design.** To validate the superiority of diffusion model, we conduct comparative experiments with two different generator architectures: (1) a conditional generative network based on GAN, influenced by our single-task model, and (2) The controllable Unet in MVSD, designed to showcase the diffusion process. Table 4.5 shows the controllable Unet outperforms the GAN-based model significantly in all evaluated metrics, STFT reduced by 0.078 to 0.753. WER decreased by 1.57% to 6.74%. The diffusion process significantly further enhances the performance to achieve 0.657 STFT and 4.27% WER. Compared to GANs, diffusion models have shown superior performance in terms of stability and sample quality. This advancement makes the generation process more controllable and precise, allowing for more consistent outputs.

| Method | *VAM* | | | *Derverberation* | |
|---|---|---|---|---|---|
| | STFT ↓ | RTE(s) ↓ | MOSE ↓ | WER ↓ | PESQ ↑ |
| VSD | 0.657 | 0.037 | 0.159 | 4.39 | 2.41 |
| MVSD *w/o* unpaired data | 0.573 | 0.033 | 0.148 | 4.32 | 2.47 |
| **MVSD** | **0.508** | **0.030** | **0.142** | **4.27** | **2.53** |

Table 4.4: Ablation study of mutual learning on SoundSpaces-Speech dataset (§4.3.4).

| Method | *VAM* | | | *Derverberation* | |
|---|---|---|---|---|---|
| | STFT ↓ | RTE(s) ↓ | MOSE ↓ | WER ↓ | PESQ ↑ |
| CNN-GAN | 0.831 | 0.076 | 0.237 | 8.31 | 1.93 |
| Unet *w/o* diffusion | 0.753 | 0.067 | 0.194 | 6.74 | 2.19 |
| **Diffusion** | **0.657** | **0.037** | **0.159** | **4.27** | **2.53** |

Table 4.5: Ablation study of diffusion model on SoundSpaces-Speech dataset (§4.3.4).

| | *VAM* | | | *Derverberation* | | Timeliness |
|---|---|---|---|---|---|---|
| Steps | STFT ↓ | RTE(s) ↓ | MOSE ↓ | WER ↓ | PESQ ↑ | RTF ↑ |
| 150 | 1.452 | 0.242 | 0.376 | 8.39 | 1.48 | **0.253** |
| 250 | 0.508 | 0.030 | 0.142 | 4.27 | 2.53 | 0.426 |
| 350 | 0.493 | 0.035 | **0.139** | **4.26** | 2.47 | 0.619 |
| 500 | 0.492 | **0.029** | 0.144 | 4.28 | **2.55** | 0.898 |
| 1000 | **0.487** | 0.033 | 0.141 | 4.35 | 2.49 | 1.809 |

Table 4.6: Ablation study of denoising steps on SoundSpaces-Speech dataset (§4.3.4).

**Denoising Steps.** We conduct experiments varying the number of denoising steps, as detailed in Table 4.6 and apply the Real-Time Factor (RTF) to measure the speed of audio generation relative to the actual duration of the audio. Results indicate suboptimal generation for steps under 250. At 250 steps, MVSD matches SOTA performance. However, more steps require longer time but yield minimal improvement, increasing the runtime by at least 30%. Consequently, we establish 250 as the optimal number of diffusion steps.

**Unpaired Data Size.** We diagnose the impact of unlabeled data by investigating the correlation between the quantity of unpaired data and performance. As shown in Table 4.7, increasing the amount of unpaired data consistently boosts the performance. Incorporating

| Num | *VAM* | | | *Derverberation* | |
|---|---|---|---|---|---|
| | STFT ↓ | RTE(s) ↓ | MOSE ↓ | WER ↓ | PESQ ↑ |
| 0 $k$ | 0.573 | 0.033 | 0.148 | 4.32 | 2.47 |
| 1 $k$ | 0.547 (+4.5%) | 0.033 (+0%) | 0.147 (+0.7%) | 4.28 | 2.50 |
| 3 $k$ | 0.521 (+9.1%) | 0.031 (+6.1%) | 0.143 (+3.4%) | 4.29 | 2.52 |
| **5 $k$** | **0.508 (+11.3%)** | **0.030 (+9.1%)** | **0.142 (+4.1%)** | **4.27** | **2.53** |

Table 4.7: Ablation study of unpaired data size on SoundSpaces-Speech dataset (§4.3.4).

unpaired data equivalent to 17.3% of the supervised samples, STFT-distance shows a notable improvement of 11.3%. Similar conclusions can also be observed in the dereverberation results. More unpaired data enables the model to learn from a broader data distribution, improving its predictive accuracy and stability. Thus, MVSD can offer a powerful potentiality for the usage of easily acquired unlabeled data.

## 4.4 Conclusion

We introduce MVSD, a mutual learning framework based on visual scene-driven diffusion model, designed for VAM and dereverberation tasks. In early exploration, we combine diffusion model with mutual learning, a strategy that leverages the complementary aspects between tasks to improve both the performance and the generalization capabilities. Consequently, MVSD achieves outstanding performance in both tasks. We empirically demonstrate that by utilizing a symmetric diffusion model architecture, MVSD can effectively extract and utilize cross-task knowledge across both tasks. Furthermore, by integrating an additional 17.3% of unpaired data into the training set, we have observed a 9.1% relative improvement in RTE for VAM. This strategy allows MVSD to access easily acquired unpaired data, thereby reducing the reliance on annotation.

CHAPTER 5

# MULTIMODAL SPOKEN DATA-DRIVEN SIGN LANGUAGE PRODUCTION

The advent of multimodal sign language generation marks a pivotal advancement in the intersection of linguistics, computer science, and accessibility. This chapter explores the innovative field of generating sign language using multimodal inputs, integrating visual, textual, and auditory data to produce accurate and expressive sign language animations or avatars. Our focus is on bridging communication gaps for the deaf and hard-of-hearing communities by enhancing the automatic generation of sign language from spoken descriptions of multiple sources.

However, there is still no viable solution for generating sign sequences directly from entire spoken content, *e.g.*, text or speech. In this chapter, we propose a unified framework MS2SL[1] for continuous sign language production, easing communication between sign and non-sign language users. MS2SL can capably convert multimodal spoken data (speech or text) into continuous sign keypoint sequences. In particular, a sequence diffusion model, utilizing embeddings extracted from text or speech, is crafted to generate sign predictions step by step. Moreover, by creating a joint embedding space for text, audio, and sign, we bind these modalities and leverage the semantic consistency among them to provide informative feedback for the model training. This embedding-consistency learning strategy minimizes the reliance on sign triplets and ensures continuous model refinement, even with a missing audio modality. Experiments on How2Sign and PHOENIX-2014T datasets demonstrate that

[1]This chapter is based on collaborative research [155] with Jian Ma, Wenguan Wang, Yi Yang and Feng Zheng, presented primarily as it appears in the ACL 2024 proceedings.

our model achieves competitive performance in sign language production. These results highlight the transformative potential of multimodal sign language generation in fostering inclusive communication across diverse settings.

## 5.1 Introduction

Signlanguage, a visual language, combines both manual (hand gestures) and non-manual cues for communication. It is specifically designed for the deaf and hearing-impaired community [9, 23, 87, 299]. According to the World Federation of the Deaf, there are 70 million deaf people and more than 200 kinds of sign languages in the world [63, 185]. Improvements in sign language production (SLP) can bridge the communication gap between the deaf and hearing [81, 104, 153, 164, 214, 255].

The challenges primarily arise from **phonological difference** and **data scarcity**. Phonological difference: signs are composed of various manual and non-manual features [161], such as hand gestures, facial expressions and limb movements [103, 136, 222]. The differences in phonological structure and means of expression create challenges in modeling the two languages. Data scarcity: multimodal high-quality sign language datasets are relatively scarce, and some datasets tend to be specific to a particular language or domain, *e.g.*, American sign [56], German weather [21, 64]. Furthermore, hearing impairments hinder pronunciation [172, 283], making it strenuous to collect sign video with aligned audio and usually resulting in the lack of auditory information. Previous researches [20, 92, 93, 282, 289] primarily focused on sign language recognition, which identifies sign fragments as the corresponding sign language lexicons (*e.g.*, gloss). Several work [96, 226, 227, 229, 263] manage the transition from gloss to sign sequences, yet the grammar of gloss can be perplexing for those without sign language training. [226, 228] can transcribe discrete words or phrases into continuous sign language sequences. However, directly producing continuous signs from entire spoken sentences still remains more exploration and efforts.

To promote barrier-free communication between signers and speakers, we introduce a Multimodal Spoken Data-Driven Continuous Sign Language Production (MS2SL) framework (Fig. 5.1). MS2SL can animate sign keypoint sequences from either speech audio or text. In addition, to alleviate data demands, we adopt an embedding-consistency learning (ECL) strategy, which is inherently based on the reciprocity among modalities, to bolster the model training. Specifically, MS2SL initially employs pre-training models like CLIP (text) [207] and HuBERT (audio) [91] to extract features from input. Subsequently, we utilize these features, serving as control conditions for the diffusion, to generate sign sequences. The attention

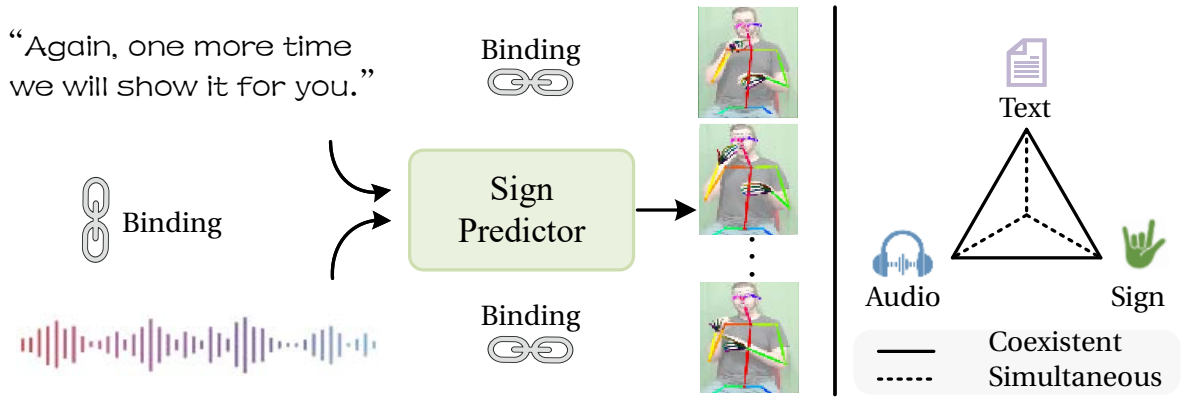


Figure 5.1: **Illustration of our sign language producer.** 1) We propose a unified, multimodal spoken data-driven framework for SLP that can directly produce sign sequences from spoken text or speech audio. 2) To overcome data scarcity, we train a joint embedding space through the spontaneous alignment of multimodal data. Within this space, we establish a consistency learning strategy to provide feedback signals that boost training.

mechanism [260] is employed to model the relationships among conditions, denoising steps, and sign movements. Besides that, ECL does not require the three modalities to coexist in the dataset. By learning a joint embedding space, inspired by ImageBind [73], ECL tightly binds the properties of different modalities and generates feedback signals to boost the training process. First, we utilize contrastive learning to bind audio and text in the embedding space. Then, we leverage the semantic consistency between co-occurring data to infer and reconstruct the embedding of missing modalitiy. The reconstruction error between the generated signs and groundtruth can be used to iteratively update MS2SL until convergence. ECL can foster cross-learning between different generation streams, allowing training even in the absence of certain modality. Furthermore, the inclusion of audio data not only enriches sample diversity and enhances multimodal comprehension but also assists in accurately capturing the expression and semantic content of sign language. We validate the effectiveness of our method across two prevalent datasets How2Sign [56] and PHOENIX-2014T [21]. Experimental results demonstrate that MS2SL achieves superior performance, both in terms of semantic consistency and sign accuracy.
In conclusion, our primary contributions are outlined as follows:

- We propose MS2SL, a unified diffusion framework for efficient multimodal spoken to sign language production. MS2SL is able to directly convert entire speech or text sentences into corresponding sign keypoints sequences.
- We present an ECL strategy that takes advantage of the intrinsic relations to enhance data

utilization. It maximizes the data potential by tapping into the underlying connections.

- We show that joint embedding is suitable for tasks that are prone to modality missing.

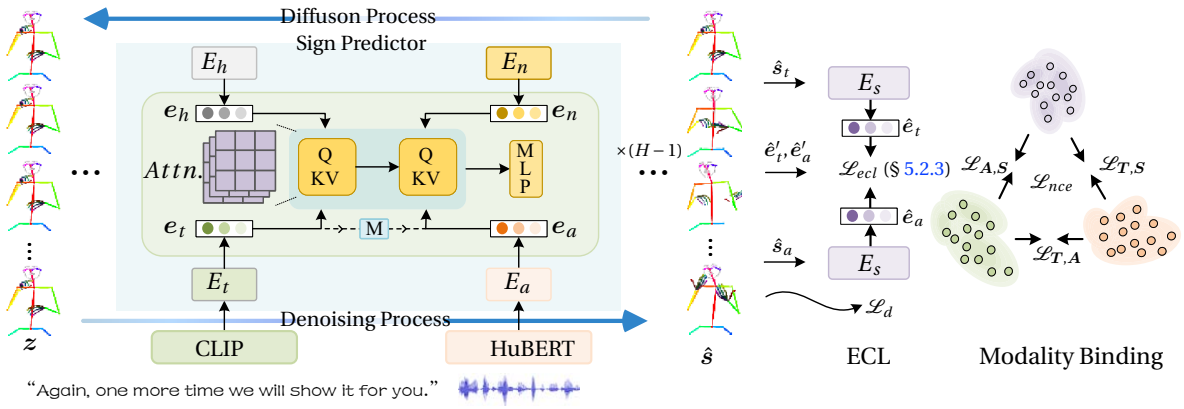


Figure 5.2: **Overview of MS2SL network architecture.** It includes three key components: sign predictor (§5.2.1), modality binding (§5.2.2) and ECL strategy (§5.2.3). MS2SL directly unifies spoken content from different modalities into a common sign language production framework. The introduction of the joint embedding space and ECL reduces the reliance on co-occurring (text, audio, sign) triplet.

## 5.2 Method

Assuming the triplets $\langle \boldsymbol{a}, \boldsymbol{t}, \boldsymbol{s} \rangle$ represent the audio, text, and sign space respectively. Our goal is to learn the mapping from text or audio to sign within a unified framework (Fig. 5.2). Given a training dataset $\mathscr{D} = \{(\boldsymbol{a}, \boldsymbol{t}, \boldsymbol{s}) \in \mathscr{A} \times \mathscr{T} \times \mathscr{S}\}$, MS2SL can realize text-to-sign $\mathscr{T} \mapsto \mathscr{S} : \boldsymbol{s} = G(\boldsymbol{t})$ and audio-to-sign $\mathscr{A} \mapsto \mathscr{S} : \boldsymbol{s} = G(\boldsymbol{a})$, where $G$ is the sign sequence diffusion generator. We initially employ pretrained models CLIP [207] and HuBERT [91] to extract features from text $\boldsymbol{t}$ and audio $\boldsymbol{a}$. Next, we employ three encoders $E_a, E_t, E_s$ to encode these features, acquiring their embeddings $\boldsymbol{e}_a$, $\boldsymbol{e}_t$, and $\boldsymbol{e}_s$. Subsequently, drawing on the operating mechanism of diffusion models, we employ a diffusion step encoder $E_h$ and a sign noise encoder $E_n$ to encode step $h$ and noise $\boldsymbol{n}$ to $\boldsymbol{e}_h$ and $\boldsymbol{e}_n$, respectively. Finally, we utilize the generator $G$ to produce the sign sequences: $\hat{\boldsymbol{s}}_t = G(\boldsymbol{e}_t, \boldsymbol{e}_h, \boldsymbol{e}_n)$ and $\hat{\boldsymbol{s}}_a = G(\boldsymbol{e}_a, \boldsymbol{e}_h, \boldsymbol{e}_n)$.

The paucity of co-occurring triplet data renders the direct training of MS2SL a formidable task. To overcome this challenge, we develop a joint embedding space that facilitates the natural alignment of multimodal data. In addition, the ECL strategy is employed to leverage the reciprocal relationships among modalities within the embedding space. It furnishes valuable feedback signals to boost the model training.

### 5.2.1 Sign Predictor

**Cross-linguistic Modeling.** MS2SL aims to solve the problem of generating variable-length sequences across modalities. It necessitates phonological modeling between spoken and sign language, associating text and audio to the same target sign sequence. The causal attention mechanism can serve as a potent remedy for this challenging issue. Taking text-to-sign as an example, we first concatenate the embeddings of text $\boldsymbol{e}_t$, denoising step $\boldsymbol{e}_h$ and noise $\boldsymbol{e}_n$. Next, we apply the causal self-attention [210] to model the relationship among them. The *mask* in causal attention ensures that the model only processes past and present information, maintaining temporal and logical coherence in the output. As such, the output is computed as: CausalAtt$[\boldsymbol{e}_t;\boldsymbol{e}_h;\boldsymbol{e}_n]$. During inference, we initiate from the text embedding and produce indices autoregressively, ceasing generation when the model predicts the sequences. Likewise, the concatenated entity of the audio $\boldsymbol{e}_a$, step $\boldsymbol{e}_h$, and noise $\boldsymbol{e}_n$ can also undergo the causal attention to capture the relationship between audio and sign. In causal attention, we adopt the common practice of positional encoding, which can model keypoints and inter-frame context while capturing cross-modal relations. Thus, to simplify the model structure, we does not explicitly design a temporal module. Finally, we employ two fully connected layers to output the sign prediction $\hat{\boldsymbol{s}}_h$ for step $h$.

**Sign Language Production.** We apply a diffusion model as the sign generator. Similarly, taking text-to-sign as an example, the diffusion generator $G$ is responsible for the gradually producing a continuous sign sequence $\hat{\boldsymbol{s}}$. Diffusion generator $G$ simulates data distribution through a gradual forward and reversible process [89], training by maximizing the evidence lower bound to approximate target distributions. Diffusion model aims to reconstruct the input from a latent variable. The forward process gradually transforms the input into noise by adding Gaussian noise. The reverse process starts from random noise and progressively removes the noise to recover the original data.

Common training for diffusion models involves independent noise prediction at each forward step $h$, potentially reducing sequence coherence and consistency. Following [8], we adopt the holistic training method. We apply a schedule function $\delta_h = 1/\log(h+1)$ $(\delta \in [0,1])$ and a step size $\alpha_h = \delta_h - \delta_{h+1}$. The predicted signs $\hat{s}_h$ at step $h$, as:

$$\hat{s}_h = \alpha_h p_h + (1-\alpha_h)\hat{s}_{h-1}, \tag{5.1}$$

where the predicted signs $p_h$ at step $h$ are given as $G(\boldsymbol{t})$. This method utilizes the output from the previous iteration as the input for the subsequent step, gradually reducing the step size as the process continues. Each step combines previous outcomes with current predictions, reducing reliance on the initial noise. We also enhance training robustness by introducing a

random noise to $\hat{s}_h$ at each step. Finally, the predicted initial sign $\hat{s}_0$ is outputted. The loss of the diffusion is defined as:

$$\mathscr{L}_d = \alpha_h s_0 + (1-\alpha_h) s_{h+1}. \tag{5.2}$$

### 5.2.2 Modality Binding

MS2SL operates in an aligned embedding space, typically dependent on audio, text, and sign data for tri-modal alignment. However, the difficulty for people with hearing impairments to perceive sound variations poses a challenge in recording these co-occurring triplets. Fortunately, ImageBind [73] reveals that a model can learn to align modalities in a joint embedding space by employing contrastive learning [78]. Training with (Image, Modality1) and (Image, Modality2) pairs can lead to a spontaneous alignment of Modality1 and Modality2 in embedding space. This alignment allows the model to excel in various tasks without requiring direct training on specific pairs of (Modality1, Modality2).

We extend the findings of ImageBind and construct a joint embedding space for the triplet dataset $(\mathscr{A},\mathscr{T},\mathscr{S})$, where MS2SL employs (text, sign) pairs as anchors to establish a cohesive space linking audio, text, and sign. Let's explore a pair of modalities $(\mathscr{T},\mathscr{S})$ with aligned observations. Given a sign sequence $\boldsymbol{s}$ and its corresponding caption $\boldsymbol{t}$. We first employ pretrained models CLIP [207] to extract textual features and encode them into normalized embeddings: $\boldsymbol{e}_t$ and $\boldsymbol{e}_s$. Then, we leverages the paired modalities $(\mathscr{T},\mathscr{S})$ to align the text with sign. The corresponding encoders are optimized by InfoNCE [186] loss $\mathscr{L}_{\mathscr{T},\mathscr{S}}$:

$$\mathscr{L}_{\mathscr{T},\mathscr{S}} = -\log\frac{\exp(\mathrm{sim}(\boldsymbol{e}_t,\boldsymbol{e}_s)/\tau)}{\sum_{m=1}^{M}\exp(\mathrm{sim}(\boldsymbol{e}_t,\boldsymbol{e}_{s_m})/\tau)}. \tag{5.3}$$

Within the mini-batch, we consider each instance, whose index is not equal to $m$, as a negative example. This approach aims to draw different embedding pairs closer within their joint embedding space. Similarly, we can also obtain $\mathscr{L}_{\mathscr{A},\mathscr{S}}$ and $\mathscr{L}_{\mathscr{T},\mathscr{A}}$ for the pairs $(\mathscr{A},\mathscr{S})$ and $(\mathscr{T},\mathscr{A})$. Interestingly, we also observe the emergent alignment between modal pairs $(\mathscr{T},\mathscr{A})$ in our embedding space. This phenomenon can occur when the training is solely based on pairs $(\mathscr{T},\mathscr{S})$ and $(\mathscr{A},\mathscr{S})$, a trend that mirrors the findings reported in [73]. Accordingly, MS2SL is designed to mainly leverage modal pairs $(\mathscr{T},\mathscr{S})$ and $(\mathscr{T},\mathscr{A})$, circumventing the need for triplet data. In practice, this is achieved by employing a triadic loss:

$$\mathscr{L}_{nce} = \mathscr{L}_{\mathscr{T},\mathscr{S}} + \mathscr{L}_{\mathscr{T},\mathscr{A}} + \mathscr{L}_{\mathscr{A},\mathscr{S}}. \tag{5.4}$$

As such, the embedding space can not only spontaneously align unseen triples but also be used in reconstructing unobserved modalities in ECL.

### 5.2.3 Embedding-consistency Learning

Given a tuple $(\mathscr{A}, \mathscr{T}, \mathscr{S})$, we employ a cyclic approach with the bound joint embedding to generate feedback signals for bidirectional cross-learning, fostering model training. When triplet data is available, the encoders first extract features from their respective modalities. Then, audio and text independently generate predicted sign language sequences $\hat{\boldsymbol{s}}_a$ and $\hat{\boldsymbol{s}}_t$. To fully utilize real data, we calculate ECL loss after 500 epochs of model training. The vanilla model, built on authentic data, guarantees minimal distribution differences between generated pseudo-embeddings and the original dataset. Semantic consistency is calculated using the embeddings $\hat{\boldsymbol{e}}_t$ and $\hat{\boldsymbol{e}}_a$ from encoder $E_s$, which encodes the two predicted sequences. We can obtain the text-to-sign error $\Delta(\hat{\boldsymbol{e}}_t,\ \boldsymbol{e}_s)$ and the audio-to-sign loss $\Delta(\hat{\boldsymbol{e}}_a,\ \boldsymbol{e}_s)$:

$$\begin{aligned} \Delta(\hat{\boldsymbol{e}}_t,\ \boldsymbol{e}_s) &= \|\hat{\boldsymbol{e}}_t, \boldsymbol{e}_s\|_2, \\ \Delta(\hat{\boldsymbol{e}}_a,\ \boldsymbol{e}_s) &= \|\hat{\boldsymbol{e}}_a, \boldsymbol{e}_s\|_2. \end{aligned} \tag{5.5}$$

Evaluation scores are derived from comparing the two embeddings $\hat{\boldsymbol{e}}_t$ and $\hat{\boldsymbol{e}}_a$. Both audio and text can receive feedback signals from the generative streams of each other. To compensate for the missing audio modality and ensure smooth processing, we use a mapping network $M$ and text embeddings to generate pseudo audio features. The operation is conducted in the embedding space, thus minimally affecting inference speed. For unpaired natural audios $\mathscr{U}$, we can get the formula:

$$\begin{aligned} \mathscr{L}_{(\mathscr{T},\ \mathscr{A},\ \mathscr{S})} &= \|E_s(G(\boldsymbol{e}_a)) - E_s(G(\boldsymbol{e}_t))\|_2, \\ \mathscr{L}_{(\mathscr{T}',\ \mathscr{S}')} &= \|E_s(M(G(\boldsymbol{e}'_t))) - E_s(G(\boldsymbol{e}'_t))\|_2. \end{aligned} \tag{5.6}$$

Then our ECL loss is defined as:

$$\mathscr{L}_{ecl} = \mathscr{L}_{(\mathscr{T},\ \mathscr{A},\ \mathscr{S}) \in \mathscr{D}} + \mathscr{L}_{(\mathscr{T}',\ \mathscr{S}') \in \mathscr{U}}. \tag{5.7}$$

MS2SL translates entire spoken sentences into continuous sign language sequences. Overall, our total loss comprises three components, *i.e.*, the diffusion model loss, ECL loss, and joint embedding loss:

$$\mathscr{L} = \lambda_1 \mathscr{L}_d + \lambda_2 \mathscr{L}_{ecl} + \lambda_3 \mathscr{L}_{nce}, \tag{5.8}$$

where the cofficients are empirically set as $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

### 5.2.4 Implementation Details

**Training.** MS2SL takes speech audio or text as inputs. We utilize pre-trained models for encoding both speech and text, HuBERT [91] for speech and CLIP [207] for text. We first

extract embeddings $e_t, e_a, e_s, e_h, e_n$ through five encoders. We employ keypoints to represent signs, like the 137 human keypoints in How2Sign [56], which are normalized and standardized before being input into the model. $e_t$, $e_a$ and $e_s$ participate in learning the joint embedding space. Concurrently, $e_t, e_a, e_h$ and $e_n$ serve as conditions to control the generation of text-to-sign and audio-to-sign, respectively. Here, we adopt the common practice [8, 227, 228] of using the first sign pose as initial noise. The first 500 epochs skip the audio-to-sign generation flow in the absence of audio. After obtaining a vanilla model, we apply the mapping network $M$ to transform $e_t$ into $e_a$ to continue the training until the model converges. Since PHOENIX-2014T [21] dataset is in *German* sign language, and our pre-trained model is primarily based on *English*, we utilize the penultimate layer features of CLIP along with MLP to align and transform between German and English. As for ECL, we incorporate cycles among the three modalities, namely audio-to-sign, text-to-sign, and audio-to-text, greatly enhancing the efficiency of data utilization. We adopt the commonly used exponential moving average [19] strategy with diffusion parameters [19] to ensure smoother, more robust training. For details, please refer to the supplementary.

**Inference.** The model can perform SLP from audio or text independently. Inference for each modality involves executing the sequence sampling of the diffusion model. Taking text-to-sign as an example, the process starts with CLIP encoding the text into features. These text features are then fed into the sign predictor, which sequentially generates a sequence noise prediction. The completion of this sampling process results in the generation of the desired sign sequence. The process for generating signs from speech is similar. We take the average of twenty generations to mitigate deviation.

**Reproducibility.** Our method is implemented using PyTorch on 2 RTX 4090 GPUs, with a training time of about 12 hours and an average inference time of 0.3 seconds[2]. Following [288], we remove data with word count exceeding 20.

## 5.3 Experiments

We assess the performance of MS2SL under text-to-sign and audio-to-sign configurations. Evaluating these different settings enables us to gauge the model's effectiveness.

### 5.3.1 Experimental Setup

**Datasets.** We conduct experiments on two continuous sign language datasets:

[2] https://hechang25.github.io/MS2SL

| Methods | How2Sign | | | | | PHOENIX-2014T | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BLEU-4 | BLEU-3 | BLEU-2 | BLEU-1 | ROUGE | BLEU-4 | BLEU-3 | BLEU-2 | BLEU-1 | ROUGE |
| Back-translation | $10.89^{\pm.03}$ | $13.32^{\pm.01}$ | $16.71^{\pm.06}$ | $22.38^{\pm.05}$ | $23.23^{\pm.03}$ | $20.53^{\pm.01}$ | $25.13^{\pm.03}$ | $32.81^{\pm.04}$ | $44.01^{\pm.02}$ | $45.61^{\pm.03}$ |
| PT [226] | $2.01^{\pm.02}$ | $3.86^{\pm.04}$ | $7.04^{\pm.00}$ | $13.69^{\pm.04}$ | $13.81^{\pm.03}$ | $11.32^{\pm.02}$ | $12.91^{\pm.01}$ | $19.04^{\pm.05}$ | $31.36^{\pm.01}$ | $32.46^{\pm.01}$ |
| MOMP [228] | $2.34^{\pm.04}$ | $3.92^{\pm.01}$ | $7.63^{\pm.02}$ | $13.68^{\pm.06}$ | $13.83^{\pm.05}$ | $11.19^{\pm.03}$ | $13.14^{\pm.02}$ | $19.64^{\pm.01}$ | $32.22^{\pm.04}$ | $32.96^{\pm.02}$ |
| Ham2Pose [8] | $2.93^{\pm.06}$ | $4.07^{\pm.04}$ | $7.31^{\pm.02}$ | $12.38^{\pm.03}$ | $13.29^{\pm.01}$ | $11.71^{\pm.03}$ | $13.22^{\pm.03}$ | $20.16^{\pm.05}$ | $33.39^{\pm.00}$ | $34.02^{\pm.04}$ |
| T2M-GPT [288] | $3.53^{\pm.03}$ | $5.14^{\pm.01}$ | $7.92^{\pm.05}$ | $12.87^{\pm.05}$ | $13.99^{\pm.03}$ | $11.66^{\pm.02}$ | $13.35^{\pm.07}$ | $21.19^{\pm.00}$ | $35.24^{\pm.02}$ | $35.44^{\pm.03}$ |
| MS2SL *w/o* ECL | $3.76^{\pm.02}$ | $6.03^{\pm.02}$ | $8.05^{\pm.04}$ | $14.51^{\pm.05}$ | $15.10^{\pm.06}$ | $12.03^{\pm.02}$ | $14.32^{\pm.04}$ | $21.72^{\pm.03}$ | $35.36^{\pm.06}$ | $35.68^{\pm.08}$ |
| MS2SL-T2S | $\mathbf{4.26}^{\pm.04}$ | $\mathbf{6.84}^{\pm.02}$ | $\mathbf{9.17}^{\pm.05}$ | $\mathbf{14.67}^{\pm.03}$ | $\mathbf{16.38}^{\pm.03}$ | $\mathbf{12.77}^{\pm.06}$ | $\mathbf{15.81}^{\pm.07}$ | $\mathbf{22.04}^{\pm.03}$ | $\mathbf{36.41}^{\pm.01}$ | $\mathbf{36.63}^{\pm.03}$ |

Table 5.1: **Comparisons of text-to-sign with the state-of-the-art methods (§5.3.2) on How2Sign [56] and PHOENIX-2014T [21].** Each metric undergoes 20 times, with the averages reported. The best and second-best results are marked in Red and Blue, respectively.

- How2Sign [56] is a challenging multimodal American sign language dataset with a 16k-word vocabulary and comprehensive annotations. It includes 1,176 entries with audio and has train/dev/test splits of 31165/1741/2357.

- PHOENIX-2014T [21], a widely applied German weather sign language dataset, contains 2,887 words, 1,066 sign annotations, with train/dev/test splits of 7096/519/642.

**Evaluation Metrics.** Following [226], we adopt back-translation approach for evaluating, *i.e.*, we leverage the cutting-edge SLT model [22] to ingeniously translate back from generated signs to text. Subsequently, we calculate BLEU [193] and ROUGE [137] scores, which are commonly used metrics for SLP and machine translation. We apply ROUGE-L F1-Score and report BLEU-1 to BLEU-4 for translation performance at different phrase lengths.

**Competitors.** For text-to-sign generation stream, we consider four SOTA competitors:

- Ham2Pose [8], which employs transformer and diffusion model, animates HamNoSys (a sign notation) into sign poses.

- T2M-GPT [288] combines VQ-VAE [259] and CLIP [207] for motion generation.

- PT [226] translates discrete spoken sentences into sign sequences.

- MOMP [228] divides sign language production into two sub-tasks: latent sign representation and animation imitation.

| Methods | BLEU-4 | BLEU-3 | BLEU-2 | BLEU-1 | ROUGE |
|---|---|---|---|---|---|
| T2A2S (§5.3.1) | $0.98^{\pm.04}$ | $1.32^{\pm.02}$ | $3.71^{\pm.02}$ | $8.38^{\pm.01}$ | $8.52^{\pm.00}$ |
| A2T2S (§5.3.1) | $1.02^{\pm.02}$ | $1.47^{\pm.01}$ | $4.66^{\pm.05}$ | $9.49^{\pm.08}$ | $9.60^{\pm.04}$ |
| MS2SL *w/o* ECL | $1.24^{\pm.07}$ | $1.63^{\pm.03}$ | $4.71^{\pm.01}$ | $10.59^{\pm.01}$ | $11.04^{\pm.03}$ |
| MS2SL-A2S | **1.67**$^{\pm.01}$ | **1.94**$^{\pm.03}$ | **5.90**$^{\pm.02}$ | **11.77**$^{\pm.05}$ | **12.16**$^{\pm.01}$ |

Table 5.2: **Audio-to-sign results on How2Sign (§5.3.2).**

As for the audio-to-sign stream, since there are not specific methods, we extend MS2SL to multiple implementations for a thorough evaluation, including audio-to-sign, audio-to-text-to-sign, and text-to-audio-to-sign. For audio-to-text-sign, we apply WeNet [280] to translate audio into text, followed by the generation of signs. Conversely, for text-to-audio-to-sign, we employ DeepVoice [72] to convert text into audio for subsequent sign generation.

### 5.3.2 Comparison to State-of-the-art

**Quantitative Results.** We present the comparative analysis results in Table 5.1 on How2Sign and PHOENIX-2014T test set. MS2SL demonstrates impressive gains against the four robust methods, establishing a new benchmark for competitive performance. In the generation of text-to-sign, our approach yields a ROUGE of 14.67, marking a notable increase of 2.39 over its counterpart (T2M-GPT, which has a 13.99 ROUGE). Furthermore, MS2SL combined with ECL surpasses the standalone by 1.28. How2Sign [56] and PHOENIX-2014T [21] are datasets of different scales, demonstrating the robustness of our method and the burgeoning potential of diffusion models in generating long sign sequences.

Table 5.2 reports the audio-to-sign results on How2Sign, noting that PHOENIX-2014T is not included here due to the absence of audio data. Our method significantly enhance performance, achieving notable improvements (*i.e.*, BLEU-1 increase from 9.49 to 11.77, ROUGE from 9.60 to 12.16). The ECL strategy also enhances ROUGE by 1.12. Considering the scarcity of audio modality data, this achievement is particularly noteworthy and shows its real-world applicability. We can also conclude that it is difficult to obtain a well-performing model by training solely with the limited audio data in How2Sign. This also highlights the urgency of utilizing non-co-occurring triplets.

**Qualitative Comparison.** Fig. 5.3 presents visual results on How2Sign [56]. It demonstrates that our method can produce signs that are more closely aligned with their semantic meaning. After meticulous examination, it is evident that MS2SL surpasses other models in generating actions with smoother transitions, heightens expressiveness, greater diversity,


Figure 5.3: **Results examples (§5.3.2): Left column:** text-to-sign generation stream, **right column:** audio-to-sign generation stream. Under given conditions, our MS2SL can generate signs sequences that are more semantically consistent with the spoken description.

and superior adherence to physical constraints. Some noise and jitter are noted in the audio-to-sign generation stream. The main reason is that our method focuses on translating complete spoken content into sign sequences, whereas previous studies [8, 226, 227] target the creation of discrete lexical symbol or phrase. The challenge of training models to convey extended semantic content and long sequences often leads to incoherent movements.

| Methods | How2Sign | PHOENIX-2014T |
|---|---|---|
| PT [226] | 1.29 | 1.54 |
| Ham2Pose [8] | 1.97 | 1.73 |
| A2T2S (§5.3.1) | 1.87 | 2.09 |
| T2M-GPT [288] | 2.19 | 2.20 |
| MS2SL | **2.65** | **3.21** |

Table 5.3: **User study (§5.3.2).**

**Discussion.** To ensure generated hand gestures align with natural biomechanics, two strategies can be used to address common defects, such as unnatural finger bending. First, kinematic constraints can be introduced to limit the range of motion at each joint, ensuring gestures remain within physiologically realistic limits. For example, defining maximum and minimum bending angles for finger joints prevents excessive motion. Additionally, skeletal structure constraints can be applied to maintain natural joint positions and angles, preventing unrealistic poses. To further improve gesture fluidity, applying a sliding average to each joint's position or angle can smooth the motion trajectories, reducing irregularities or noise-induced unnatural movements. This method not only smooths transitions but also enhances the stability and continuity of the gesture, avoiding abrupt changes. Together, these methods ensure the generated gestures are both biomechanically realistic and more natural and fluid overall.

**User Study.** Given the challenge of finding sign language experts, who require extensive training, we conduct a user study with 10 hearing volunteers. We ask the volunteers to compare sign sequences generated by different methods. We slow down sign sequence playback for easier comparison by volunteers. Slowing down the playback allowed volunteers to observe the details of each gesture, such as hand positions and pose transitions, enabling better evaluation of the fluency and accuracy of the sign sequences. However, the study did not specifically test whether volunteers could understand the meaning of the generated sign language simply by watching the sequences. Volunteers select the sequence closer to the ground truth and assign a score. Our scoring range is from 1 to 5, with higher scores indicating closer proximity to the ground truth. Most participants report that the sign sequences generated by MS2SL are smoother and more accurate (Table 5.3). According to user feedback, MS2SL excels in expression clarity and pose accuracy, making it a preferred choice and enhancing the overall usage experience.

**Discussion.** While most volunteers report that the sign language sequences generated by

| Methods | BLEU-4 | BLEU-3 | BLEU-2 | BLEU-1 | ROUGE |
|---|---|---|---|---|---|
| Audio | $0.98^{\pm.04}$ | $1.32^{\pm.02}$ | $3.71^{\pm.02}$ | $8.38^{\pm.01}$ | $8.52^{\pm.00}$ |
| Text | $1.74^{\pm.00}$ | $2.41^{\pm.02}$ | $3.43^{\pm.07}$ | $8.62^{\pm.03}$ | $9.57^{\pm.01}$ |
| T2A2S | $1.85^{\pm.03}$ | $2.35^{\pm.03}$ | $4.26^{\pm.02}$ | $8.52^{\pm.08}$ | $9.28^{\pm.03}$ |
| MS2SL | **4.26**$^{\pm.04}$ | **6.84**$^{\pm.02}$ | **9.17**$^{\pm.05}$ | **14.67**$^{\pm.03}$ | **16.38**$^{\pm.03}$ |

Table 5.4: **Ablation study of different modalities data (§5.3.2).**

| Methods | BLEU-4 | BLEU-3 | BLEU-2 | BLEU-1 | ROUGE |
|---|---|---|---|---|---|
| $0k$ | $3.76^{\pm.06}$ | $6.03^{\pm.02}$ | $8.05^{\pm.05}$ | $14.51^{\pm.04}$ | $15.10^{\pm.02}$ |
| $5k$ | $3.79^{\pm.06}$ | $6.23^{\pm.02}$ | $8.17^{\pm.05}$ | $14.62^{\pm.04}$ | $15.56^{\pm.02}$ |
| $10k$ | $3.82^{\pm.03}$ | $6.37^{\pm.03}$ | $8.31^{\pm.02}$ | $14.57^{\pm.06}$ | $15.87^{\pm.00}$ |
| $15k$ | **4.26**$^{\pm.04}$ | **6.84**$^{\pm.02}$ | **9.17**$^{\pm.05}$ | **14.67**$^{\pm.03}$ | **16.38**$^{\pm.03}$ |

Table 5.5: **Ablation study of embedding consistency learning (§5.3.2).**

| Steps | BLEU-4 | BLEU-3 | BLEU-2 | BLEU-1 | ROUGE |
|---|---|---|---|---|---|
| 0 | $0.62^{\pm.02}$ | $2.08^{\pm.03}$ | $4.16^{\pm.07}$ | $9.57^{\pm.03}$ | $9.72^{\pm.05}$ |
| 5 | $1.09^{\pm.04}$ | $2.42^{\pm.06}$ | $5.24^{\pm.04}$ | $10.44^{\pm.00}$ | $10.90^{\pm.01}$ |
| 10 | $4.26^{\pm.04}$ | $6.84^{\pm.02}$ | $9.17^{\pm.05}$ | $14.67^{\pm.03}$ | $16.38^{\pm.03}$ |
| 15 | $4.04^{\pm.01}$ | $6.23^{\pm.04}$ | $9.58^{\pm.02}$ | $15.26^{\pm.02}$ | $17.33^{\pm.01}$ |
| 20 | $4.87^{\pm.01}$ | $6.66^{\pm.04}$ | $9.67^{\pm.06}$ | $15.45^{\pm.02}$ | $17.24^{\pm.01}$ |

Table 5.6: **Ablation study of diffusion model (§5.3.2).**

MS2SL are clearer and more accurate than those from other methods, the study does not explore whether volunteers could understand the exact meaning of the signs. Slowing down the playback can help volunteers distinguish the sign trajectories, and clearer, more accurate generation might improve understanding. Therefore, due to experimental limitations, the focus is primarily on evaluating the similarity of the generated sequences to the ground truth in terms of motion trajectories, rather than exploring comprehension of the sign meaning.

### 5.3.3 Ablation Study

We conduct careful profiling of the impact of each module within MS2SL on How2Sign.

**Data in Different Modalities.** We primarily conduct four experiments: audio-to-sign, text-to-sign, text-to-audio-to-sign, and MS2SL, to compare and analyze the role of different

| Pre-trained | BLEU-4 | BLEU-3 | BLEU-2 | BLEU-1 | ROUGE |
|---|---|---|---|---|---|
| WavLM [33] | $1.63^{\pm.06}$ | $1.79^{\pm.02}$ | $6.12^{\pm.01}$ | $10.94^{\pm.00}$ | $11.43^{\pm.02}$ |
| HuBERT [91] | $1.67^{\pm.07}$ | $1.94^{\pm.04}$ | $5.90^{\pm.02}$ | $11.77^{\pm.06}$ | $12.16^{\pm.01}$ |
| CLIP [207] | $4.26^{\pm.04}$ | $6.84^{\pm.02}$ | $9.17^{\pm.05}$ | $14.67^{\pm.03}$ | $16.38^{\pm.03}$ |
| BERT [52] | $4.11^{\pm.04}$ | $6.91^{\pm.02}$ | $10.27^{\pm.01}$ | $13.37^{\pm.05}$ | $16.52^{\pm.06}$ |

Table 5.7: **Ablation study of different pretrained models (§5.3.2).**

modalities. As shown in Table 5.4, although direct generation from audio-to-sign and text-to-sign can yield appropriate results, MS2SL significantly outperforms them. Removal of text data leads to a 6.29 decrease in BLEU-1, highlighting its crucial role. The mediating role of text leads to an increase 0.76 in ROUGE. Multimodal data yields superior results compared to its unimodal counterpart, enriching the learning process with more diverse information.

**Embedding Consistency Learning.** We investigate the impact of the cyclical consistency training presented in § 5.2.3, and the results are illustrated in Table 5.5. We note that common training method performs comparably to baseline models, while cyclical consistency boosts model performance akin to adding substantial training data. Compared to the alternative only with single modality, MS2SL approach shows a 1.12 increase in BLEU-2 and a 1.28 increase in ROUGE, demonstrating the synergistic effect of integrating data from multiple modalities. We further pay particular attention to the impact of dataset size. We also observe a direct correlation between dataset size and model accuracy. For smaller datasets (under $10k$ samples), the accuracy plateau around 15.5. Several insights can be drawn: i) Performances improve as more training data is used. ii) Over $10k$ unpaired data entries, the signs might be of good quality, but the model cannot further improve on a large scale, possibly due to the scarcity of audio. This trend shows that more data notably improves sequence generation, even without clear semantic boundaries.

**Diffusion Model.** As shown in Table 5.6, implementing the diffusion model lead to a significant enhancement. The quality metrics, such as BLEU-1 and ROUGE, improved by 5.1 and 6.66, respectively, compared to non-diffusion model approach. Our study explores denoising steps ranging from 5 to 20, revealing a discernible trade-off between generation quality and computational efficiency. Compared to a fixed 10-step denoising process, the 20-step process unsteadily improve 0.78 in BLEU-1 by approximately 5.3% with a disproportionate increase in computational load. Thus, in this chapter, 10 is set as the default number of denoising steps. This choice provides a balance between performance and computational efficiency, ensuring effective denoising without excessive processing time.

**Pre-trained Models.** We select four widely used models, including, CLIP (text), BERT (text) [52], HuBERT (audio) [91] and WavLM (audio) [33], to assess their impact on performance. As shown in Table 5.7, for audio-to-sign generation, the impact of HuBERT and WavLM on performance is minor, with negligible differences observed between the two pre-trained models. GPT outperforms CLIP models in text-related tasks, with a slight improvement of up to 0.14 in ROUGE. This may be because BERT focuses on natural language processing, leading to enhanced text understanding capabilities.

## 5.4 Conclusion

We explore a unified framework that combines diffusion and pretrained models to generate sign language from spoken depictions. We surpass other competitors and solidify this classic framework as a highly competitive method for SLP. MS2SL effectively handles diverse modalities of data for analysis and decoupling. Despite its advancements, our model struggles with maintaining contextual flow in generation, and MS2SL cannot handle lengthy data, which is a future focus. Our research pioneers direct sign language generation from speech, offering some insights to advance the community.

CHAPTER 6

# HYBRID MODEL COLLABORATION FOR SIGN LANGUAGE TRANSLATION

The capability to translate sign language effectively is crucial for bridging communication barriers between the deaf and the hearing. In this chapter, we propose an innovative framework, VRG-SLT, for translating sign language into spoken language, facilitating communication between signing and non-signing communities. VRG-SLT utilizes a hierarchical VQ-VAE to convert continuous sign sequences into discrete representations, referred as sign codes, which are subsequently aligned with text by a fine-tuned pre-trained language model. Additionally, we employ RAG to extend and enhance the language model, producing more semantically coherent and precise text. We demonstrate that the collaborative hybrid model VRG-SLT, equipped with a hierarchical VQ-VAE and the rich prior knowledge embedded in pre-trained large language models, achieves state-of-the-art performance on prominent benchmarks like How2Sign and PHOENIX-2014T. Moreover, the incorporation of additional understanding and knowledge through RAG further improves the accuracy of the generated text.

## 6.1 Introduction

Sign languages play a crucial role in facilitating communication among deaf individuals, characterized by unique linguistic traits [74, 189, 247]. Unlike spoken languages, they rely on visual cues like gestures, body movements, facial expressions, and eye movements to convey

semantic information [16, 103, 136, 222]. Sign language translation (SLT) involves converting sign gestures from video clips into spoken descriptions [21, 22, 36, 50, 133, 177, 298, 300], facilitating communication freedom and accessibility of information for both sign and non-sign language users. In practice, SLT highlights its versatility and significant value across various scenarios [81], such as public service broadcasts, and personal assistants, *etc.*

Building effective and accurate sign language translation systems commonly encounters the following obstacles: 1) **Data scarcity**: The collection of sign language data is particularly challenging, owing to its limited user population and the considerable costs and complexities of data gathering and annotation. For instance, the How2Sign dataset [56] contains only 30,000 pairs, hampering effective model training and knowledge acquisition. 2) **Unique syntax**: Sign language, inherently distinct from spoken language, possesses its own grammar, word formation, and lexicon. These differences, especially in word order, make transcription between the two languages complex. 3) **Multimodal contexts**: Sign language is a multimodal form of communication that combines manual and non-manual actions, such as facial expressions and body postures, to convey detailed and precise information.

Previous research generally divides SLT into two distinct tasks [36]: **Sign2Notation (or sign language recognition, SLR)**, the conversion of sign language videos to lexical representations (*e.g.*, gloss); and **Notation2Text**, the translation of these representations into the target text language, a process akin to conventional machine translation but dealing with complex, multimodal sign inputs. SLR [4, 48, 118, 134, 188, 204, 300] endeavors to decipher successive signs as discrete gloss lexicon, yet disregards the differing grammar and linguistic structure of sign language from spoken language, resulting in a lack of semantic coherence and fluidity in complete sentences. Notation2Text [21, 22, 56, 133, 298, 300] struggles to fully capture non-verbal elements of sign language, such as facial expressions, and relies on extensive bilingual corpora, posing a significant challenge for resource-scarce sign languages. SLT seeks to convert sign language videos into spoken sentences, ensuring accuracy and comprehensibility by accounting for grammatical and word order differences. Recent efforts treat sign language translation as a comprehensive task and introduce external domain knowledge for optimization [36], yet the accuracy and generalization remain inadequate for real-world applications. Large Language Models (LLMs) [17, 45, 52, 54, 125, 131, 145, 197, 250, 279, 291] exhibit strong comprehension and extensive prior knowledge, thereby lessening the dependency on large-scale domain-specific data. However, their integration into sign language translation remains insufficiently explored.

In this chapter, we concentrate on incorporating sign language gestures into large language models to translate sign language sentences into spoken text (Fig. 6.1). Presently,

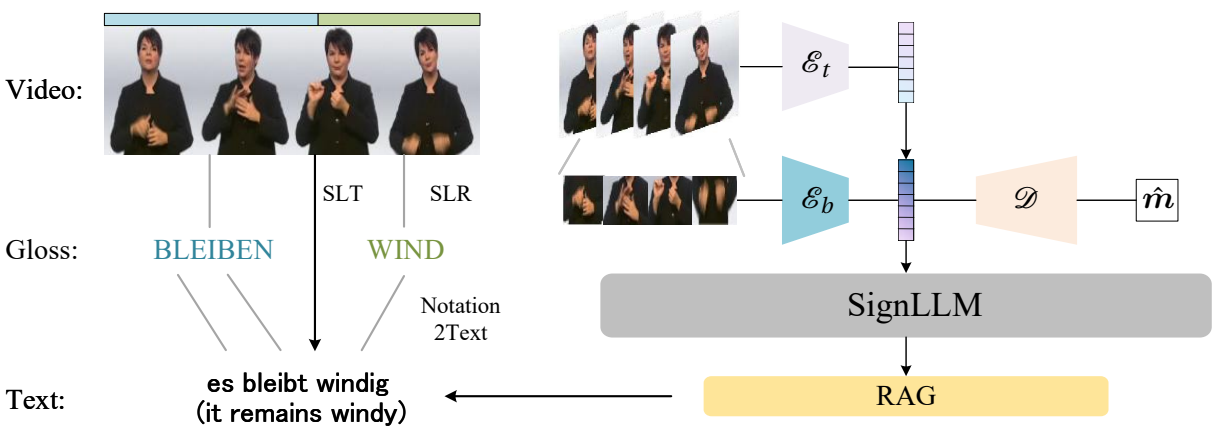


Figure 6.1: We propose VRG-SLT, a hybrid collaborative sign language translation framework that integrates a hierarchical VQ-VAE (Sign-tokenizer) with the pretrained language model FLAN-T5. Initially, Sign-tokenizer vectorizes continuous sign language videos into discrete codes stored in a sign codebook. Subsequently, the sign codebook and a spoken codebook form a unified 'sign-spoken vocabulary' for jointly fine-tuning FLAN-T5, enabling it to comprehend both sign and spoken languages. Finally, we employ a RAG strategy to calibrate and refine the initial outputs (§6.1).

the prevailing LLMs are text-centric and are incapable of directly translating from sign language to text. How can we jointly train sign videos with text? The key is to construct a visual encoder that embeds sign clips into hidden representations and aligns them with text during the finetuning of large models. Our approach integrates insights from VQ-VAE-2 [215] and text-to-motion technologies [100, 288], utilizing a hierarchical VQ-VAE [259] and a pretrained language model (FLAN-T5 [44]) to efficiently convert sign expressions into spoken sequences. Furthermore, we enhance these translations by incorporating a Retrieval-Augmented Generation (RAG) [132] strategy, further improving the performance of VRG-SLT beyond preliminary results. Concretely, we propose a two-stage pipeline VRG-SLT, aimed at facilitating the interpretation of sign language motions. VRG-SLT initially employs a sign-specific VQ-VAE (Sign-tokenizer) to encode raw sign segments into discrete vectors, which are subsequently converted into text by FLAN-T5. Sign-tokenizer generates a discrete and fixed-size 'sign vocabulary' by effectively encoding sign sequences into latent codes. Traditional VQ-VAE prioritize full-body motion, making them less suitable for sign language that emphasizes hand and torso features. Inspired by VQ-VAE-2, we designed Sign-tokenizer as a two-level vectorization network: the top level captures body information, such as the motion trajectories of the shoulders and elbows, while the bottom level focuses on modeling the movements of hands. The hand information is then injected into the top level's global information for sign reconstruction. This integration allows for joint optimization of the

entire Sign-tokenizer. Subsequently, FLAN-T5 processes the sign codes to jointly learn and bridge the syntax and grammar of the 'sign-spoken language' with corresponding textual descriptions. One notable limitation is that LLMs such as T5 may lack sufficient or in-depth domain knowledge, tending to produce inaccurate or unrealistic responses (known as 'hallucination'). Thus, to rectify incorrect answers, we adopt a RAG strategy that retrieves pertinent knowledge and polishes the initial translations for enhanced fluency and accuracy.

We are pioneering the integration of hierarchical VQ-VAE and LLMs into SLT, bolstered by RAG for enhanced generation. VRG-SLT notably surpasses competitors on benchmarks such as How2Sign [56] and PHOENIX-2014T [21]. Our contributions are summarized as follows: (1) We propose VRG-SLT, a collaborative hybrid model for sign language translation that treats sign movements as a 'unique language' and injects them into LLMs for joint training with text. (2) We introduce a Sign-tokenizer that captures both body and hand trajectory characteristics. Its hierarchical structure adeptly handles intricate details and diverse contextual movements. (3) We integrate RAG strategy into VRG-SLT, retrieving and combining relevant knowledge to produce more accurate and content-rich output.

## 6.2 Method

As illustrated in Fig. 6.2, VRG-SLT comprises a Sign-tokenizer, a sign-aware language model, and a RAG module. Sign-tokenizer (§6.2.1) employs a hierarchical VQ-VAE-2 to encode raw sign sequences into discrete tokens for a codebook. These tokens, along with spoken texts, form a new 'motion-language vocabulary', which is then integrated into the fine-tuning of LLMs. The sign-aware language model (§6.2.2) learns to understand motion tokens from corresponding textual descriptions. Concurrently, RAG (§6.2.3) accesses relevant knowledge to refine output text and alleviate hallucinations. In particular, Sign-tokenizer consists of a sign encoder $\mathscr{E}$ and a sign decoder $\mathscr{D}$. Sign-tokenizer first encodes a sign motion sequence $\boldsymbol{m}^{1:M}$ of $M$ frames into $L$ motion codes $\boldsymbol{z}^{1:L}$, and decodes $\boldsymbol{z}^{1:L}$ back into a reconstructed motion sequence $\hat{\boldsymbol{m}}^{1:M} = \mathscr{D}(\boldsymbol{z}^{1:L}) = \mathscr{D}(\mathscr{E}(\boldsymbol{m}^{1:M}))$. Here, $L = M/l$, $l$ denotes the temporal downsampling rate. The goal of VRG-SLT is to generate corresponding verbal text $\hat{\boldsymbol{t}}^{1:N}$ with $N$ words conditioned on the sign sequence $\boldsymbol{m}$, denoted as $\hat{\boldsymbol{t}}^{1:N} = SignLLM(\boldsymbol{z}^{1:L})$.

### 6.2.1 Sign Tokenizer

To represent sign in discrete tokens, we pre-train a sign tokenizer based on the vector quantized variational autoencoders architecture in [77, 243, 259, 288]. The Sign-tokenizer,

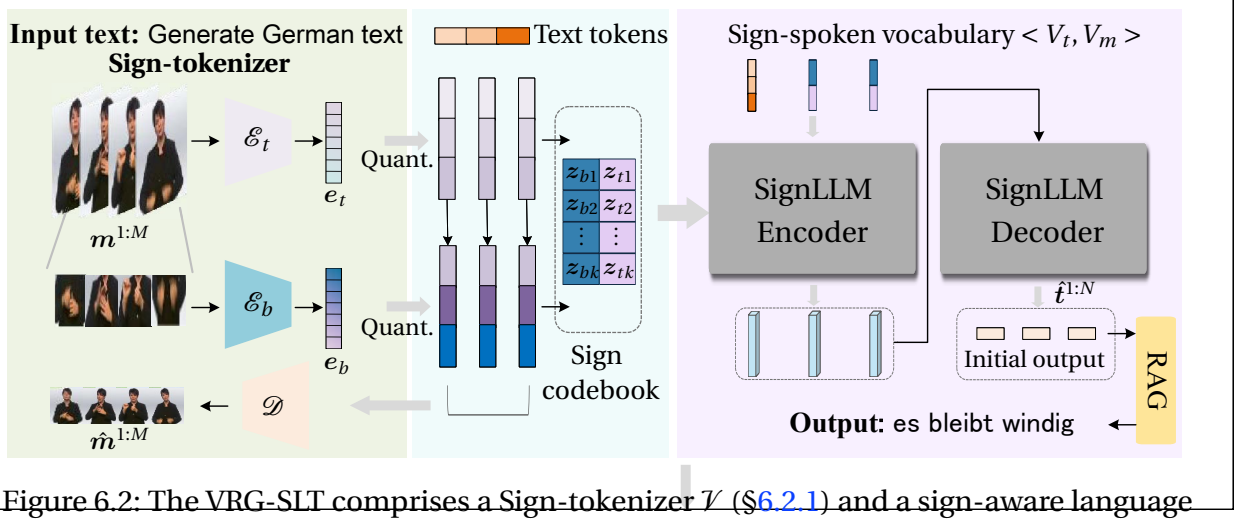


Figure 6.2: The VRG-SLT comprises a Sign-tokenizer $\mathcal{V}$ (§6.2.1) and a sign-aware language model, SignLLM (§6.2.2). The Sign-tokenizer encodes sign actions into a *sign codebook* and, together with the text tokenizer, creates a unified vocabulary $\mathcal{V}$. Using SignLLM, we perform joint learning of sign and spoken languages for sign language translation. The two encoders of the Sign-tokenizer encode global body movements and detailed hand features, respectively, achieving a comprehensive and precise understanding of sign motion. Finally, we refine the initial output using a Retrieval-Augmented Generation (RAG) strategy (§6.2.3).

featuring a hierarchical architecture with two encoders $\mathscr{E}_t$, $\mathscr{E}_b$ and a decoder $\mathscr{D}$, is tailored to capture sign characteristics comprehensively. The encoders and quantizers generate highly informative discrete sign codes, while the decoder reconstructs these codes into sign sequences $\hat{\boldsymbol{m}}^{1:M}$. Sign-tokenizer can effectively represent sign movements as a special language, facilitating the integration of sign and spoken sentences in generative pre-trained models $SignLLM$. The Sign-tokenizer, once trained, applies quantization to the upper ($\boldsymbol{e}_t$) and lower ($\boldsymbol{e}_b$) levels of each input. The quantized representations, $\boldsymbol{e}_t$ and $\boldsymbol{e}_b$, are utilized by the VQVAE to establish a joint probability density for overarching semantic features $\boldsymbol{p}_t$ and the conditional probability density for detailed local mappings $\boldsymbol{p}_b$. The generation process concludes by sampling quantized codebook vectors from $\boldsymbol{p}_t$ for global consistency and $\boldsymbol{p}_b$ for local detail, which are then fed into the decoder $\mathscr{D}$ to generate reconstructed signs.

Specifically, both the sign encoders first applies 1D convolutions to the frame-wise sign motions $m^{1:M}$ along the time dimension, generating latent vectors $\boldsymbol{e}_t = \mathscr{E}_t(\boldsymbol{m}^{1:M})$ and $\boldsymbol{e}_b = \mathscr{E}_b(\boldsymbol{m}^{1:M})$. These latent vectors are then discretized into codebook entries $z$ by quantization. The learnable codebook $Z = \boldsymbol{z}_{i=1}^{K}$ contains $K$ embedding vectors, each of dimension $d$. The quantization function $Q(\cdot)$ replaces each row vector with its nearest codebook entry in $Z$, as indicated:

$$\boldsymbol{z}_i = Q(\hat{\boldsymbol{z}}^i) := \arg\min_{\boldsymbol{z}_k \in Z} \|\hat{\boldsymbol{z}}_i - \boldsymbol{z}_k\|_2 . \tag{6.1}$$

After quantization, the sign decoder $\mathscr{D}$ projects $\boldsymbol{z}^{1:L}$ back to raw motion space as $\hat{\boldsymbol{m}}^{1:M}$ with $M$ frames. We train our motion tokenizer using the method outlined in [77, 288] to synchronize the vector space of the codebook with the encoder output, with three distinct loss functions for optimization. The codebook loss applies only to codebook variables, drawing the selected codebook vector closer to the encoder output $\mathscr{E}(\boldsymbol{m})$. The commitment loss applies solely to the encoder weights, ensuring the encoder output remains close to the chosen codebook vector to minimize frequent shifts between code vectors. The overall objective is described in Eq. 6.2, where $\boldsymbol{z}$ represents the quantized code for training sample $\boldsymbol{m}$, $\mathscr{E}_{\sqcup}$ and $\mathscr{E}_{\lfloor}$ is the encoding function, $\mathscr{D}$ is the decoding function, $sg$ denotes a stop-gradient operation that prevents gradients from flowing into its argument, and $\beta_1$ and $\beta_2$ is a hyperparameter that controls resistance to changes in the encoder's code output.

(6.2)

$$\mathscr{L}_{\mathscr{V}} = \|\boldsymbol{m} - \mathscr{D}(\boldsymbol{z})\|_2 + \|\mathscr{E}_t(\boldsymbol{m}) - \boldsymbol{z}_t\|_2 + \|\mathscr{E}_b(\boldsymbol{m}) - \boldsymbol{z}_b\|_2 + \beta_1\|\boldsymbol{z}_t - \mathscr{E}_t(\boldsymbol{m})\|_2 + \beta_2\|\boldsymbol{z}_b - \mathscr{E}_b(\boldsymbol{m})\|_2.$$

We empirically set $\beta_1$ and $\beta_2$ to 1, respectively. To enhance the quality of the generated motion, we also employ L1 smooth loss, velocity regularization, EMA, and codebook reset [215].

### 6.2.2 SignLLM

With this Sign-tokenizer, a sign motion $m^{1:M}$ can be converted into a sign token sequence $z^{1:L}$, which facilitates joint representation with similar vocabulary embeddings in language models [123, 187, 212]. The unified vocabulary enables simultaneous learning from sign and spoken language, allowing Sign-tokenizer and SignLLM to perform hybrid collaborative operations. Unlike previous text-to-motion approaches [35, 77, 288] that adopt separate modules for text and sign sequence processing, our approach aims to integrate text and sign motion processing in a unified manner. To achieve this, we merge the original text vocabulary $V_t = \{v_t\}$ with the sign vocabulary $V_m = \{v_m\}$, which preserves the order in our sign codebook $Z$. The sign vocabulary $V_m$ contains special tokens like boundary indicators, such as </sos> and </eos> for the start and end of sign, respectively. With the unified text-sign vocabulary $V =< V_t, V_m >$, we can handle sign and text data in a general format, where both input and output "words" are drawn from the same vocabulary. These tokens can represent spoken language, sign motion, or a combination of both. As a result, our method enables flexible representation of diverse sign-related outputs within a single $SignLLM$.

The model takes 'sign codes' as context or conditioning to produce an output spoken text, which represents an item in the unified vocabulary. The input sequence is $\{\boldsymbol{x}_s\}_{i=1}^{N}$, where $\boldsymbol{x}_s \in V$ and $N$ is the input length. The target output is $\boldsymbol{t}^{1:N}$, where $\boldsymbol{t} \in V$ and $N$ is

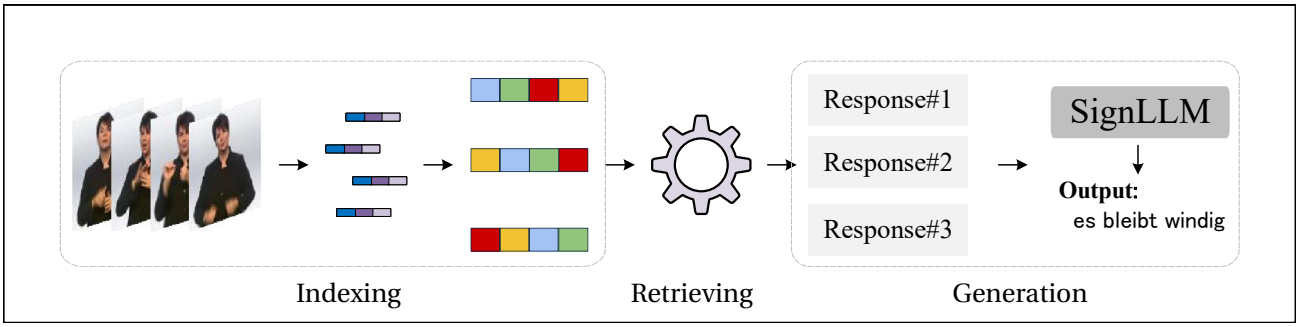


Figure 6.3: Following the classic RAG, we integrate the retrieval step into the generative model, pulling relevant documents from knowledge base to refine the initial output.

the output length. As depicted in Fig. 6.2, the source tokens enter SignLLM encoder, and SignLLM decoder predicts the probability distribution of the next token at each step, $p_\theta(\boldsymbol{t} \mid \boldsymbol{x}_s) = \prod_i p_\theta\left(\boldsymbol{t}^i \mid \boldsymbol{t}^{<i}, \boldsymbol{x}_s\right)$. The objective during training is to maximize the log-likelihood:

$$\mathcal{L}_{LM} = -\sum_{i=0}^{L_t-1} \log p_\theta\left(\boldsymbol{t}^i \mid \boldsymbol{t}^{<i}, \boldsymbol{x}_s\right). \tag{6.3}$$

By optimizing this objective, VRG-SLT can capture the underlying patterns and relationships from data distribution, thereby facilitating the accurate predict of target tokens. During inference, the target tokens are sampled recursively from the predicted distribution $p_\theta\left(\hat{\boldsymbol{t}}^i \mid \hat{\boldsymbol{t}}^{<i}, \boldsymbol{x}_s\right)$ until the end token (i.e., </s>) is reached. This approach sequentially generates the target sequence, where each token is probabilistically derived from both the previously generated tokens and the original input. Consequently, SignLLM maintains semantic consistency while achieving high accuracy.

### 6.2.3 RAG

RAG [132] has become a paradigm in the LLM field for enhancing generative tasks. Specifically, RAG involves an initial retrieval step where the LLM queries external data sources for relevant information before proceeding to answer questions or generate text. This strategy not only guides the generation phase but also uses retrieved evidence to enhance the accuracy and relevance of responses, reducing content errors known as 'hallucinations'.

Following the classic RAG workflow, we also engage in the crucial steps of indexing, retrieving, and generating relevant content or responses.

Indexing: Data is segmented into manageable chunks and transformed into vectors using a well-balanced embedding model.

Retrieving: When a query is received, the system transcodes the query into vectors using the initial encoding model, as shown in Fig. 6.3 It then calculates similarity scores with the indexed vectorized chunks, retrieving the top-K (*i.e.*,3) chunks with the highest similarity.

Generation: The model synthesizes the query and retrieved knowledge into a prompt to generate responses. Responses can vary by different motion codes or prompt text.

### 6.2.4 Training Scheme

FLAN-T5, originally pre-trained with a text-based vocabulary $V_t$, is aligned with sign language through the introduction of a sign-specific vocabulary $V_m$. Our training steps, depicted in Fig. 6.4, consists of three stages: (1) training a sign tokenizer to represent signs with discrete codes; (2) finetuning on sign language to bridge sign motion and language; and (3) tuning the output with RAG[1].

**Training of Sign-tokenizer.** Initially, Sign-tokenizer is trained using the objective defined in Eq. 6.2, enabling any sign sequence $\boldsymbol{m}^{1:M}$ to be represented as a sequence of motion tokens, which integrates seamlessly with textual information. After optimization, the Sign-tokenizer remains unchanged throughout the rest of the pipeline.

**SignLLM Finetuning.** The SignLLM are then trained and fine-tuned on the unified text-sign vocabulary $V =< V_t, V_m >$. We utilize existing sign language datasets (such as How2Sign [56] and PHOENIX-2014T [21]) as a foundation to create a guided sign-action dataset. As explored in prior works [52, 187, 211, 212], we also adopt an objective inspired by [212] where 15% of input tokens $X_s$ are randomly replaced with a sentinel token. The target sequence is then constructed by extracting the dropped-out spans, delimited by the same sentinel tokens, with an additional sentinel token marking the end of the sequence. We establish the relationship between motion and language using paired sign-text datasets [76, 200]. Through training, our model can establish the relationship between text and motion.

**RAG Tuning.** For example, in the case of motion captioning, the prompt might say: "Generate English text: <sign_tokens>" or "Generate German text: <sign_tokens>". Here, <sign_tokens> represents tokens from Sign-tokenizer. We utilize the SQuAD database [52] for general knowledge expansion and ECMWF [86] for weather data. After retrieving relevant information, we combine the initial output with the retrieved knowledge and input them into an large model for refinement. BERT [52] is employed for retrieving relevant information.

## 6.3 Experiments

**Dataset.** We assess performance on prevailing benchmarks **How2Sign** [56] and **PHOENIX-2014T** [21]. How2Sign is a comprehensive American Sign Language dataset, comprising

[1] https://vrg-slt.github.io/VRG-SLT-demos

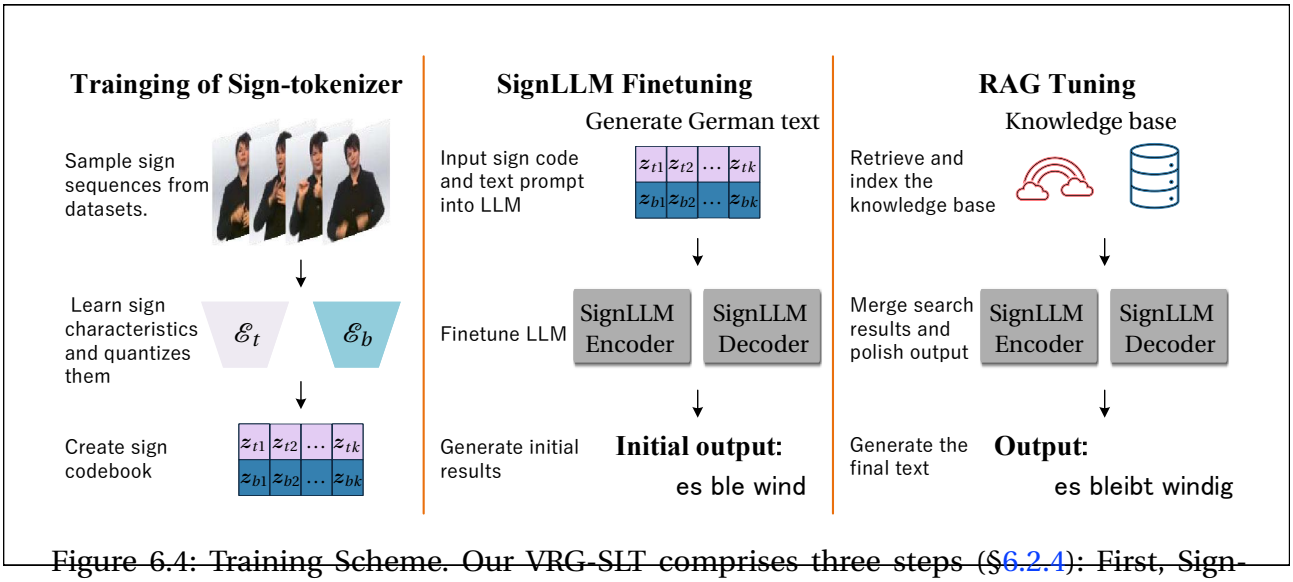


Figure 6.4: Training Scheme. Our VRG-SLT comprises three steps (§6.2.4): First, Sign-tokenizer learns a codebook for discrete sign representations. Next, we train SignLLM using a mix of spoken and sign data. Finally, we polish the initial output using RAG.

approximately 80 hours of sign language videos with corresponding caption annotations. It consists of 31,164, 1,740, and 2,356 sign-video-text triplets for training, validation, and testing, respectively. PHOENIX-2014T, a German Sign Language (DGS) dataset, consists of weather forecast segments extracted from TV broadcasts, with 7,096, 519, and 642 video-text pairs allocated for training, validation, and testing, respectively.

**Evaluation Metric.** Drawing from previous research [21, 22, 298, 300], we utilize ROUGE [137] and BLEU [193] metrics to evaluate precision and fluency of translated content. ROUGE metric primarily measures the overlap between generated text and reference text, focusing on recall, while BLEU emphasizes accuracy and the matching of n-grams.

| Methods | How2Sign | | | | | PHOENIX-2014T | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ROUGE↑ | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ | ROUGE↑ | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ |
| SL-Luong [21] | 18.75 | 19.46 | 9.53 | 4.67 | 3.21 | 31.80 | 32.24 | 19.03 | 12.83 | 9.58 |
| TSPNet-Joint [133] | 16.84 | 17.93 | 11.71 | 6.59 | 4.07 | 34.96 | 36.10 | 23.12 | 16.88 | 13.41 |
| SL-Transf [22] | 21.92 | 24.74 | 13.66 | 8.20 | 5.18 | – | 46.61 | 33.73 | 26.19 | 21.32 |
| STMC-T [300] | 25.40 | 29.38 | 15.27 | 8.68 | 6.05 | 46.65 | 46.98 | 36.09 | 28.70 | 23.65 |
| TIN-Transf+BT* [298] | 26.33 | 28.20 | 15.02 | 9.24 | 6.28 | 49.54 | 50.80 | 37.75 | 29.72 | 24.32 |
| SLRT [36] | 31.27 | 30.10 | 18.13 | 10.43 | 7.98 | 52.65 | 53.97 | 41.75 | 33.84 | 28.39 |
| **VRG-SLT (Ours)** | **33.38** | **35.61** | **20.35** | **13.12** | **8.53** | **53.92** | **55.74** | **43.31** | **36.59** | **30.17** |

Table 6.1: Compared with state-of-the-art methods on How2Sign and PHOENIX-2014T. Methods marked with an asterisk (*) first perform SLR and then Notation2Text. Our method outperforms existing competitors in terms of both accuracy and contextual relevance (§6.3.1).

**Implementation Details.** We set the codebook of Sign-tokenizer as $K \in \mathbb{R}^{512\times1024}$ for most comparisons. The encoders incorporate a temporal downsampling rate $l$ of 4. Diverging from the conventional VQ-VAE-2, the bottom encoder of our Sign-tokenizer utilizes inputs from both raw hand actions and hidden features from the top encoder. We utilize FLAN-T5-base [212] as the underlying architecture for our language model, with a baseline model consisting of 12 layers in both the transformer encoder and decoder. Moreover, all our models employ the AdamW optimizer for training. The Sign-tokenizer are trained utilizing a $10^{-3}$ learning rate and a 512 mini-batch size, while our SignLLM have a $2\times10^{-4}$ learning rate for the finetuning stage and a 32 mini-batch size. Sign-tokenizer undergoes $100k$ iterations of training, while SignLLM undergoes $200k$ iterations.

### 6.3.1 Comparisons with SOTA methods

We develop a comprehensive sign language translation framework, treating sign motion uniquely as a language and incorporating a hierarchical VQ-VAE and SignLLM, further enhanced with RAG for enhanced accuracy. We utilize the $80M$ pre-trained *Flan-T5-Base*[44, 212] model as the backbone, finetuning it through the unified codebook (§6.2.2) for all subsequent comparisons. The results are calculated with a 95% confidence interval from 10 repeated runs. The results in Table 6.1 indicate that VRG-SLT delivers strong performance across all metrics, demonstrating its cross-language learning ability and semantic consistency. Notably, it achieves a BLEU-4 score of 30.17 on PHOENIX-2014T dataset, exceeding the nearest competitor by 1.78 points, and scores 53.92 in ROUGE-L, surpassing others by 1.27 points. These outcomes suggest our model's ability to more effectively decode and render sign language nuances into accurate and fluid translations. Our model prioritizes contextual coherence, leveraging LLM's strong capability for context modeling to produce coherent, semantically consistent sentences. This is reflected in its ROUGE scores, which measure how well the translated text covers the reference text vocabulary. Sign language includes significant non-verbal information like expressions and body language, essential for full meaning. Our model captures these details through the encoding abilities of our hierarchical VQ-VAE, enabling translations that go beyond mere words to include emotional and emphatic cues, significantly benefiting BLEU scores.

### 6.3.2 Ablation Studies

To systematically evaluate the contributions of each component of the VRG-SLT model, we design and conduct ablation studies targeting the parameter count of pretrained LLMs, the

| Methods | How2Sign | | | | | PHOENIX-2014T | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ROUGE↑ | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ | ROUGE↑ | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ |
| FLAN-T5-base | 33.38 | 35.61 | 20.35 | 13.12 | 8.53 | 53.92 | 55.74 | 43.31 | 36.59 | 30.17 |
| FLAN-T5-small | 34.06 | 35.32 | 21.17 | 13.97 | 9.20 | 54.71 | **57.92** | 43.84 | 37.22 | 31.36 |
| FLAN-T5-large | **34.81** | 35.61 | 21.54 | **14.30** | 9.73 | 55.80 | 56.39 | 43.04 | 38.52 | 32.44 |
| FLAN-T5-XL | 34.73 | **36.38** | **21.80** | 13.76 | **9.88** | **56.59** | 55.05 | **44.17** | **38.94** | **32.83** |

Table 6.2: Ablation study for large-scale pre-trained model sizes (§6.3.2).

| Methods | How2Sign | | | | | PHOENIX-2014T | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ROUGE↑ | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ | ROUGE↑ | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ |
| VQ-VAE | 30.80 | 34.08 | 18.16 | 10.01 | 8.04 | 50.10 | 52.93 | 41.96 | 34.52 | 28.61 |
| VQ-VAE-2 | 32.47 | 35.11 | 19.48 | 12.92 | 8.29 | **53.05** | 53.62 | 43.13 | 35.23 | 29.58 |
| Sign-Tokenizer | **33.38** | **35.61** | **20.35** | **13.12** | **8.53** | 53.92 | **55.74** | **43.31** | **36.59** | **30.17** |
| Sign-Tokenizer-128 | 27.94 | 30.66 | 15.39 | 10.68 | 6.13 | 48.27 | 50.64 | 36.87 | 33.39 | 26.81 |
| Sign-Tokenizer-256 | 31.16 | 34.47 | 19.61 | 11.81 | 7.42 | 50.53 | 52.19 | 39.53 | 34.84 | 26.25 |
| Sign-Tokenizer-512 | **34.84** | 34.65 | **20.93** | 12.60 | 8.33 | 52.36 | 53.27 | 41.94 | 35.26 | 29.30 |
| Sign-Tokenizer-1024 | 33.38 | **35.61** | 20.35 | **13.12** | **8.53** | **53.92** | **55.74** | **43.31** | **36.59** | **30.17** |

Table 6.3: Ablation study for different VQ-VAE structures (§6.3.2).

| Methods | How2Sign | | | | | PHOENIX-2014T | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ROUGE↑ | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ | ROUGE↑ | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ |
| *w/o* RAG | 32.16 | 33.41 | 19.39 | 12.83 | 7.63 | 52.25 | 54.90 | 42.70 | 35.11 | 29.49 |
| Before SignLLM | 32.04 | **36.61** | 19.61 | 13.10 | 8.10 | 54.83 | 54.26 | 42.97 | 36.08 | 29.70 |
| After SignLLM | **33.38** | 35.61 | **20.35** | **13.12** | **8.53** | **53.92** | **55.74** | **43.31** | **36.59** | **30.17** |

Table 6.4: Ablation study on RAG strategy (§6.3.2).

structure of the Sign-tokenizer, and RAG strategies. These ablation studies, which involve selectively removing or altering specific model features or structures, aim to elucidate the impact of each component.

**Pre-trained Model Sizes.** We evaluate the performance across the four publicly accessible pre-trained models from FLAN-T5 (Table 6.2). Our experimental results with the FLAN-T5-small show a compelling balance between size and performance. FLAN-T5-small achieve competitive accuracy rates in our tests, showing only a marginal decrease in performance compared to its larger counterparts. The FLAN-T5-small model excels in speed, showing

an approximate 37% boost in inference speed, while the FLAN-T5-XL model surpasses in accuracy with a high of 56.59 in ROUGE. The results make FLAN-T5-small an appealing choice for applications where efficiency is paramount. The slight trade-off in translation accuracy is more than offset by the gains in speed and resource efficiency, indicating that FLAN-T5-small is particularly well-suited for resource-constrained environments.

**Sign-tokenizer.** To evaluate the impact of differentVQ-VAE structures on the accuracy, we experiment with VQ-VAE, VQ-VAE-2, and hierarchical VQ-VAE (Table 6.3). To ensure the reliability of the results, we perform multiple runs for each configuration and took the average. The basic VQ-VAE model, although effective in encoding visual information, fell short in accurately translating complex gestures, achieving only a 34.08 BLEU-1. The improved VQ-VAE-2, with its more detailed encoding layers, raise BLEU-1 to 35.11. Further, our adoption of the hierarchical VQ-VAE, evolved from VQ-VAE-2, significantly enhances the capture of sign language details, boosting translation BLEU-1 to 35.61, thus proving its superiority in handling complex sign language data.

**Codebook Size.** In our pursuit to refine SLT accuracy, we vary the codebook sizes within our VQ-VAE framework and observe significant differences in ROUGE and BLEU scores (Table 6.3). In the VQ-VAE model, the size of the embedding space directly influences the model's ability to quantize the input data. Initially, with a codebook size of 256, the model scored 48.27 in ROUGE and 50.64 in BLEU-1. Doubling the codebook size to 512 improved the ROUGE to 50.53 and BLEU to 52.19. However, when the embedding space reach a certain size (*i.e.*, 1024), performance improvement plateau, where the scores escalated to 53.92 for ROUGE and 55.74 for BLEU-1. These results underscore the importance of a larger codebook in capturing a broader array of features necessary for accurate translation. Larger embedding spaces provide finer quantization but also introduce higher computational complexity and storage requirements. Thus, an code space size around 1024 offers a reasonable trade-off, providing good reconstruction performance while maintaining low computational cost.

**RAG.** In our ablation study on the application of RAG, we evaluate three configurations (Table 6.4): without RAG, RAG applied before the LLM, and RAG applied after the LLM. Our findings indicate significant differences in translation accuracy across these setups. The model without RAG achieved a BLEU-1 score of 54.90 and a ROUGE score of 52.25, serving as our baseline. Integrating RAG before the LLM results in improved performance, with the BLEU score rising to 54.83 and the ROUGE score to 54.26. However, placing RAG after the LLM yielded the best results, with a BLEU score of 53.92 and a ROUGE score of 55.74. This configuration leverage the contextual retrieval capabilities of RAG to enhance the LLM's output. The RAG can enhance the knowledge coverage by retrieving from external

knowledge bases, which is particularly useful for generating knowledge-based answers. Without retrieval enhancement, RAG-SLT relies on pre-trained knowledge, leading to insufficient handling of new information. The RAG also facilitate better context understanding by retrieving relevant documents, thus enhancing response relevance.

## 6.4 Conclusion

Our study represents the first attempt, to the best of our knowledge, to employ large language models for sign language translation tasks. We present VRG-SLT as a unified framework for sign language translation, generating spoken descriptions based on prompt-driven instructions. Extensive experiments on our How2Sign and PHOENIX-2014T datasets demonstrate competitive performance and validate the efficacy of each module. Our findings underscore the substantial benefits that each component contributes to the overall framework. The hierarchical VQ-VAE plays a vital role in effectively encoding visual gestures into a compressed representation, which proves critical, while the LLM facilitates a robust linguistic framework that enriches the translation process with deep syntactic and semantic understanding. Collectively, these components push the boundaries of traditional SLT methods, achieving a BLEU-1 score improvement from 53.97 to 55.74 and a ROUGE score from 52.65 to 53.92. The improvements note in our empirical results show the potential of the integration to produce reliable and contextually translation. The hybrid collaborative training of VQ-VAE and LLMs paves the way for developing more nuanced and accurate communication tools for the deaf community, highlighting the potential of integrating multiple advanced technologies.

CHAPTER 7

# CONCLUSION AND FUTURE WORK

This study involves disciplines such as computer science, linguistics, and audio processing. Sign language, a unique visual language, relies not only on hand movements but also incorporates multimodal information like facial expressions and body postures. Converting spoken language into sign language is a challenging task that requires understanding the verbal content accurately and expressing the original tones naturally. These requirements make the translation process complex, particularly when adapting to various linguistic styles and habits. By exploring four key technologies—voice conversion, reverberation style transfer, sign language production and translation—we aim to facilitate seamless communication between hearing and hearing-impaired individuals. Through extensive analysis and experimental validation, this study is poised to make substantial progress in automatic translation between speech and sign language. It not only to improve communication efficiency and quality of life for the hearing-impaired but also to provide new theoretical and practical foundations for the study of multimodal interaction.

**Social Impact.** In the context of today's diverse communication landscape, providing effective communication methods for specific groups has significant social value. As globalization deepens, diversity and inclusivity have become key drivers of societal progress. For millions of hearing-impaired individuals, effective communication is not just crucial for engaging in society but also for unlocking personal potential. This research significantly enhances the efficiency and accuracy of speech-and-sign language conversion, offering more natural and intuitive communication methods for the hearing-impaired. This improvement not only enhances their social interaction experience but could also have profound impacts on

education and public services. For example, in education, this technology could be used to develop resources specifically for hearing-impaired students, making information and knowledge more accessible. In public services, it can greatly increase the convenience and independence of hearing-impaired individuals in medical, legal, and everyday situations.

**Future Work.** In future endeavors, our research will delve deeper into improving the flexibility and adaptability of the translation system and how to expand it to support more languages and styles of sign language. Additionally, we are committed to applying our technology in practical settings, such as remote education and public services, particularly in scenarios that require cross-lingual and cross-cultural communication, to further the societal inclusion of the hearing-impaired.

# BIBLIOGRAPHY


[1] M. ABE, S. NAKAMURA, K. SHIKANO, AND H. KUWABARA, *Voice conversion through vector quantization*, in International Conference on Acoustics, Speech, Signal Processing, 1988.

[2] N. ADALOGLOU, T. CHATZIS, I. PAPASTRATIS, A. STERGIOULAS, G. T. PAPADOPOULOS, V. ZACHAROPOULOU, G. J. XYDOPOULOS, K. ATZAKAS, D. PAPAZACHARIOU, AND P. DARAS, *A comprehensive study on deep learning-based methods for sign language recognition*, IEEE Transactions on Multimedia, 24 (2022), pp. 1750–1762.

[3] K. AKUZAWA, K. ONISHI, K. TAKIGUCHI, K. MAMETANI, AND K. MORI, *Conditional deep hierarchical variational autoencoder for voice conversion*, in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2021.

[4] S. ALBANIE, G. VAROL, L. MOMENI, T. AFOURAS, J. S. CHUNG, N. FOX, AND A. ZISSERMAN, *BSL-1K: scaling up co-articulated sign language recognition using mouthing cues*, in European Conference on Computer Vision, 2020.

[5] J. B. ALLEN AND D. A. BERKLEY, *Image method for efficiently simulating small-room acoustics*, The Journal of the Acoustical Society of America, 65 (1979), pp. 943–950.

[6] A. ALMAHAIRI, S. RAJESWAR, A. SORDONI, P. BACHMAN, AND A. C. COURVILLE, *Augmented cyclegan: Learning many-to-many mappings from unpaired data*, in International Conference on Machine Learning, 2018.

[7] S. ARIK, J. CHEN, K. PENG, W. PING, AND Y. ZHOU, *Neural voice cloning with a few samples*, in Advances in Neural Information Processing Systems, 2018.

[8] R. S. ARKUSHIN, A. MORYOSSEF, AND O. FRIED, *Ham2pose: Animating sign language notation into pose sequences*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[9] D. F. ARMSTRONG AND S. WILCOX, *Origins of sign languages*, Deaf studies language and education, (2003), pp. 305–318.

[10] A. ARNAB, M. DEHGHANI, G. HEIGOLD, C. SUN, M. LUCIC, AND C. SCHMID, *Vivit: A video vision transformer*, in International Conference on Computer Vision, 2021.

[11] A. ASAI, Z. WU, Y. WANG, A. SIL, AND H. HAJISHIRZI, *Self-rag: Learning to retrieve, generate, and critique through self-reflection*, in International Conference on Learning Representations, 2024.

[12] O. AVRAHAMI, D. LISCHINSKI, AND O. FRIED, *Blended diffusion for text-driven editing of natural images*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[13] P. BACHMAN, O. ALSHARIF, AND D. PRECUP, *Learning with pseudo-ensembles*, in Advances in Neural Information Processing Systems, 2014.

[14] A. BAEVSKI, S. SCHNEIDER, AND M. AULI, *vq-wav2vec: Self-supervised learning of discrete speech representations*, in International Conference on Learning Representations, 2020.

[15] S. BILBAO, *Modeling of complex geometries and boundary conditions in finite difference/finite volume time domain room acoustics simulation*, IEEE Transactions on Audio, Speech, and Language Processing, 21 (2013), pp. 1524–1533.

[16] D. BRENTARI, *Sign language phonology*, The handbook of phonological theory, (2011), pp. 691–721.

[17] T. B. BROWN, B. MANN, N. RYDER, M. SUBBIAH, J. KAPLAN, P. DHARIWAL, A. NEELAKANTAN, P. SHYAM, G. SASTRY, A. ASKELL, S. AGARWAL, A. HERBERT-VOSS, G. KRUEGER, T. HENIGHAN, R. CHILD, A. RAMESH, D. M. ZIEGLER, J. WU, C. WINTER, C. HESSE, M. CHEN, E. SIGLER, M. LITWIN, S. GRAY, B. CHESS, J. CLARK, C. BERNER, S. MCCANDLISH, A. RADFORD, I. SUTSKEVER, AND D. AMODEI, *Language models are few-shot learners*, in Advances in Neural Information Processing Systems, 2020.

[18] H. BU, J. DU, X. NA, B. WU, AND H. ZHENG, *Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline*, in Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, 2017.

[19] Z. CAI, A. RAVICHANDRAN, S. MAJI, C. C. FOWLKES, Z. TU, AND S. SOATTO, *Exponential moving average normalization for self-supervised and semi-supervised learning*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[20] N. C. CAMGÖZ, S. HADFIELD, O. KOLLER, AND R. BOWDEN, *Subunets: End-to-end hand shape and continuous sign language recognition*, in International Conference on Computer Vision, 2017.

[21] N. C. CAMGÖZ, S. HADFIELD, O. KOLLER, H. NEY, AND R. BOWDEN, *Neural sign language translation*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[22] N. C. CAMGÖZ, O. KOLLER, S. HADFIELD, AND R. BOWDEN, *Sign language transformers: Joint end-to-end sign language recognition and translation*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[23] R. CAMPBELL, M. MACSWEENEY, AND D. WATERS, *Sign language and the brain: a review*, Journal of deaf studies and deaf education, 13 (2008), pp. 3–20.

[24] C. CAO, Z. REN, C. SCHISSLER, D. MANOCHA, AND K. ZHOU, *Interactive sound propagation with bidirectional path tracing*, ACM Transactions on Graphics, 35 (2016), pp. 1–11.

[25] W. CHAN AND I. R. LANE, *Deep convolutional neural networks for acoustic modeling in low resource languages*, in International Conference on Acoustics, Speech, Signal Processing, 2015.

[26] H. CHANG, H. ZHANG, L. JIANG, C. LIU, AND W. T. FREEMAN, *Maskgit: Masked generative image transformer*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[27] S. CHANG, H. PARK, J. CHO, H. PARK, S. YUN, AND K. HWANG, *Subspectral normalization for neural audio data processing*, in ICASSP, 2021.

[28] T. CHAVDAROVA AND F. FLEURET, *SGAN: an alternative training of generative adversarial networks*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[29] B. CHEN, C. DU, AND K. YU, *Neural fusion for voice cloning*, IEEE Transactions on Audio, Speech, and Language Processing, 30 (2022), pp. 1993–2001.

[30] C. CHEN, R. GAO, P. CALAMIA, AND K. GRAUMAN, *Visual acoustic matching*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[31] C. CHEN, U. JAIN, C. SCHISSLER, S. V. A. GARI, Z. AL-HALAH, V. K. ITHAPU, P. ROBINSON, AND K. GRAUMAN, *Soundspaces: Audio-visual navigation in 3d environments*, in European Conference on Computer Vision, 2020.

[32] C. CHEN, W. SUN, D. HARWATH, AND K. GRAUMAN, *Learning audio-visual dereverberation*, in International Conference on Acoustics, Speech, Signal Processing, 2023.

[33] S. CHEN, C. WANG, Z. CHEN, Y. WU, S. LIU, Z. CHEN, J. LI, N. KANDA, T. YOSHIOKA, X. XIAO, J. WU, L. ZHOU, S. REN, Y. QIAN, Y. QIAN, J. WU, M. ZENG, X. YU, AND F. WEI, *Wavlm: Large-scale self-supervised pre-training for full stack speech processing*, IEEE Journal of Selected Topics in Signal Processing, 16 (2022), pp. 1505–1518.

[34] X. CHEN, Y. DUAN, R. HOUTHOOFT, J. SCHULMAN, I. SUTSKEVER, AND P. ABBEEL, *Infogan: Interpretable representation learning by information maximizing generative adversarial nets*, in Advances in Neural Information Processing Systems, 2016.

[35] X. CHEN, B. JIANG, W. LIU, Z. HUANG, B. FU, T. CHEN, AND G. YU, *Executing your commands via motion diffusion in latent space*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[36] Y. CHEN, F. WEI, X. SUN, Z. WU, AND S. LIN, *A simple multi-modality transfer learning baseline for sign language translation*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[37] K. L. CHENG, Z. YANG, Q. CHEN, AND Y. TAI, *Fully convolutional networks for continuous sign language recognition*, in European Conference on Computer Vision, 2020.

[38] Y. CHENG, F. WEI, J. BAO, D. CHEN, AND W. ZHANG, *Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[39] D. G. CHILDERS, K. WU, D. HICKS, AND B. YEGNANARAYANA, *Voice conversion*, Speech Communication, 8 (1989), pp. 147–158.

[40] J. CHOI, S. KIM, Y. JEONG, Y. GWON, AND S. YOON, *ILVR: conditioning method for denoising diffusion probabilistic models*, in International Conference on Computer Vision, 2021.

[41] Y. CHOI, Y. UH, J. YOO, AND J.-W. HA, *Stargan v2: Diverse image synthesis for multiple domains*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[42] J. CHOU AND H. LEE, *One-shot voice conversion by separating speaker and content representations with instance normalization*, in Conference of the International Speech Communication Association, 2019.

[43] J. CHOU, C. YEH, H. LEE, AND L. LEE, *Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations*, in Conference of the International Speech Communication Association, 2018.

[44] H. W. CHUNG, L. HOU, S. LONGPRE, B. ZOPH, Y. TAY, W. FEDUS, Y. LI, X. WANG, M. DEHGHANI, S. BRAHMA, ET AL., *Scaling instruction-finetuned language models*, Journal of Machine Learning Research, 25 (2024), pp. 1–53.

[45] K. CLARK, M. LUONG, Q. V. LE, AND C. D. MANNING, *ELECTRA: pre-training text encoders as discriminators rather than generators*, in International Conference on Learning Representations, 2020.

[46] K. CLARK, M. LUONG, C. D. MANNING, AND Q. V. LE, *Semi-supervised sequence modeling with cross-view training*, in Conference on Empirical Methods in Natural Language Processing, 2018.

[47] P. R. COHEN AND S. L. OVIATT, *The role of voice input for human-machine communication.*, proceedings of the National Academy of Sciences, 92 (1995), pp. 9921–9927.

[48] R. CUI, H. LIU, AND C. ZHANG, *Recurrent convolutional neural networks for continuous sign language recognition by staged optimization*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[49] S. DAVIS AND P. MERMELSTEIN, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE transactions on acoustics, speech, and signal processing, 28 (1980), pp. 357–366.

[50] M. DE COSTER, D. SHTERIONOV, M. VAN HERREWEGHE, AND J. DAMBRE, *Machine translation from signed to spoken languages: State of the art and challenges*, Universal Access in the Information Society, (2023), pp. 1–27.

[51] L. DENG, G. HINTON, AND B. KINGSBURY, *New types of deep neural network learning for speech recognition and related applications: An overview*, in ICASSP, 2013.

[52] J. DEVLIN, M. CHANG, K. LEE, AND K. TOUTANOVA, *BERT: pre-training of deep bidirectional transformers for language understanding*, in North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[53] P. DHARIWAL AND A. NICHOL, *Diffusion models beat gans on image synthesis*, in Advances in Neural Information Processing Systems, 2021.

[54] L. DONG, N. YANG, W. WANG, F. WEI, X. LIU, Y. WANG, J. GAO, M. ZHOU, AND H. HON, *Unified language model pre-training for natural language understanding and generation*, in Advances in Neural Information Processing Systems, 2019.

[55] J. DU, X. NA, X. LIU, AND H. BU, *AISHELL-2: transforming mandarin ASR research into industrial scale*, CoRR, abs/1808.10583 (2018).

[56] A. C. DUARTE, S. PALASKAR, L. VENTURA, D. GHADIYARAM, K. DEHAAN, F. METZE, J. TORRES, AND X. GIRÓ-I-NIETO, *How2sign: A large-scale multimodal dataset for continuous american sign language*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[57] A. EPHRAT, I. MOSSERI, O. LANG, T. DEKEL, K. WILSON, A. HASSIDIM, W. T. FREEMAN, AND M. RUBINSTEIN, *Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation*, in Special Interest Group on Graphics and Interactive Techniques, 2018.

[58] O. ERNST, S. E. CHAZAN, S. GANNOT, AND J. GOLDBERGER, *Speech dereverberation using fully convolutional networks*, in European Signal Processing Conference, 2018.

[59] P. ESSER, R. ROMBACH, AND B. OMMER, *Taming transformers for high-resolution image synthesis*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[60] H. FEI, Y. GUO, B. LI, D. JI, AND Y. REN, *Adversarial shared-private model for cross-domain clinical text entailment recognition*, Knowledge-Based Systems, 221 (2021), p. 106962.

[61] H. FEI, Y. REN, AND D. JI, *Retrofitting structure-aware transformer language model for end tasks*, in Conference on Empirical Methods in Natural Language Processing, 2020.

[62] H. FEI, M. ZHANG, B. LI, AND D. JI, *End-to-end semantic role labeling with neural transition-based model*, in Association for the Advancement of Artificial Intelligence, 2021.

[63] J. FENLON AND E. WILKINSON, *Sign languages in the world*, Sociolinguistics and Deaf communities, 1 (2015), p. 5.

[64] J. FORSTER, C. SCHMIDT, O. KOLLER, M. BELLGARDT, AND H. NEY, *Extensions of the sign language recognition and translation corpus rwth-phoenix-weather*, in Proceedings of the Ninth International Conference on Language Resources and Evaluation, 2014.

[65] D. FRIED, R. HU, V. CIRIK, A. ROHRBACH, J. ANDREAS, L. MORENCY, T. BERG-KIRKPATRICK, K. SAENKO, D. KLEIN, AND T. DARRELL, *Speaker-follower models for vision-and-language navigation*, in Advances in Neural Information Processing Systems, 2018.

[66] S. FU, C. YU, T. HSIEH, P. PLANTINGA, M. RAVANELLI, X. LU, AND Y. TSAO, *Metricgan+: An improved version of metricgan for speech enhancement*, in Conference of the International Speech Communication Association, 2021.

[67] S.-W. FU, C.-F. LIAO, Y. TSAO, AND S.-D. LIN, *Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement*, in International Conference on Machine Learning, 2019.

[68] T. FUNKHOUSER, N. TSINGOS, I. CARLBOM, G. ELKO, M. SONDHI, J. E. WEST, G. PINGALI, P. MIN, AND A. NGAN, *A beam tracing method for interactive architectural acoustics*, The Journal of the acoustical society of America, 115 (2004), pp. 739–756.

[69] A. C. GADE, *Acoustics in halls for speech and music*, Springer handbook of acoustics, (2014), pp. 317–366.

[70] H. GAMPER AND I. J. TASHEV, *Blind reverberation time estimation using a convolutional neural network*, in International Workshop on Acoustic Signal Enhancement, 2018.

[71] R. GAO AND K. GRAUMAN, *2.5d visual sound*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[72] A. GIBIANSKY, S. Ö. ARIK, G. F. DIAMOS, J. MILLER, K. PENG, W. PING, J. RAIMAN, AND Y. ZHOU, *Deep voice 2: Multi-speaker neural text-to-speech*, in Advances in Neural Information Processing Systems, 2017.

[73] R. GIRDHAR, A. EL-NOUBY, Z. LIU, M. SINGH, K. V. ALWALA, A. JOULIN, AND I. MISRA, *Imagebind one embedding space to bind them all*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[74] N. S. GLICKMAN AND W. C. HALL, *Language deprivation and deaf mental health*, Routledge, 2018.

[75] I. J. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. C. COURVILLE, AND Y. BENGIO, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, 2014.

[76] C. GUO, S. ZOU, X. ZUO, S. WANG, W. JI, X. LI, AND L. CHENG, *Generating diverse and natural 3d human motions from text*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[77] C. GUO, X. ZUO, S. WANG, AND L. CHENG, *TM2T: stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts*, in European Conference on Computer Vision, 2022.

[78] R. HADSELL, S. CHOPRA, AND Y. LECUN, *Dimensionality reduction by learning an invariant mapping*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2006.

[79] K. HAN, Y. WANG, D. WANG, W. S. WOODS, I. MERKS, AND T. ZHANG, *Learning spectral mapping for speech dereverberation and denoising*, IEEE Transactions on Audio, Speech, and Language Processing, 23 (2015), pp. 982–992.

[80] A. HAO, Y. MIN, AND X. CHEN, *Self-mutual distillation learning for continuous sign language recognition*, in International Conference on Computer Vision, 2021.

[81] R. HARRIS, H. M. HOLMES, AND D. M. MERTENS, *Research ethics in sign language communities*, Sign Language Studies, 9 (2009), pp. 104–131.

[82] T. HASHIMOTO, D. SAITO, AND N. MINEMATSU, *Many-to-many and completely parallel-data-free voice conversion based on eigenspace dnn*, IEEE Transactions on Audio, Speech, and Language Processing, 27 (2019), pp. 332–341.

[83] D. HE, Y. XIA, T. QIN, L. WANG, N. YU, T. LIU, AND W. MA, *Dual learning for machine translation*, in Advances in Neural Information Processing Systems, 2016.

[84] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[85] E. HEIM, *Constrained generative adversarial networks for interactive image generation*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[86] H. HERSBACH, B. BELL, P. BERRISFORD, S. HIRAHARA, A. HORÁNYI, J. MUÑOZ-SABATER, J. NICOLAS, C. PEUBEY, R. RADU, D. SCHEPERS, ET AL., *The era5 global reanalysis*, Quarterly Journal of the Royal Meteorological Society, 146 (2020), pp. 1999–2049.

[87] G. HICKOK, U. BELLUGI, AND E. S. KLIMA, *The neurobiology of sign language and its implications for the neural basis of language*, Nature, 381 (1996), pp. 699–702.

[88] G. HINTON, L. DENG, D. YU, G. E. DAHL, A.-R. MOHAMED, N. JAITLY, A. SENIOR, V. VANHOUCKE, P. NGUYEN, T. N. SAINATH, ET AL., *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, IEEE Signal processing magazine, 29 (2012), pp. 82–97.

[89] J. HO, A. JAIN, AND P. ABBEEL, *Denoising diffusion probabilistic models*, in Advances in Neural Information Processing Systems, 2020.

[90] S. HOCHREITER AND J. SCHMIDHUBER, *Long short-term memory*, Neural Computing, 9 (1997), pp. 1735–1780.

[91] W. HSU, B. BOLTE, Y. H. TSAI, K. LAKHOTIA, R. SALAKHUTDINOV, AND A. MOHAMED, *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*, IEEE Transactions on Audio, Speech, and Language Processing, 29 (2021), pp. 3451–3460.

[92] H. HU, W. ZHAO, W. ZHOU, Y. WANG, AND H. LI, *Signbert: Pre-training of hand-model-aware representation for sign language recognition*, in International Conference on Computer Vision, 2021.

[93] H. HU, W. ZHOU, AND H. LI, *Hand-model-aware sign language recognition*, in Association for the Advancement of Artificial Intelligence, 2021.

[94] R. HU AND A. SINGH, *Unit: Multimodal multitask learning with a unified transformer*, in International Conference on Computer Vision, 2021.

[95] X. HUANG AND S. BELONGIE, *Arbitrary style transfer in real-time with adaptive instance normalization*, in International Conference on Computer Vision, 2017.

[96] E. J. HWANG, J. KIM, AND J. C. PARK, *Non-autoregressive sign language production with gaussian space*, in Proceedings of the British Machine Vision Conference, 2021.

[97] N. B. IBRAHIM, H. H. ZAYED, AND M. M. SELIM, *Advances, challenges and opportunities in continuous sign language recognition*, Journal of Engineering and Applied Sciences, 15 (2020), pp. 1205–1227.

[98] A. IMASHEV, M. MUKUSHEV, V. KIMMELMAN, AND A. SANDYGULOVA, *A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The K-RSL*, 2020.

[99] S. JI, W. XU, M. YANG, AND K. YU, *3d convolutional neural networks for human action recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (2013), pp. 221–231.

[100] B. JIANG, X. CHEN, W. LIU, J. YU, G. YU, AND T. CHEN, *Motiongpt: Human motion as a foreign language*, in Advances in Neural Information Processing Systems, 2023.

[101] S. JIANG, B. SUN, L. WANG, Y. BAI, K. LI, AND Y. FU, *Sign language recognition via skeleton-aware multi-model ensemble*, CoRR, abs/2110.06161 (2021).

[102] Q. JIN, A. R. TOTH, T. SCHULTZ, AND A. W. BLACK, *Voice convergin: Speaker de-identification by voice transformation*, in International Conference on Acoustics, Speech, Signal Processing, 2009.

[103] R. E. JOHNSON AND S. K. LIDDELL, *Toward a phonetic representation of signs: Sequentiality and contrast*, Sign Language Studies, 11 (2011), pp. 241–274.

[104] N. K. KAHLON AND W. SINGH, *Machine translation from text to sign language: a systematic review*, Universal Access in the Information Society, 22 (2023), pp. 1–35.

[105] H. KAMEOKA, T. KANEKO, K. TANAKA, AND N. HOJO, *Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks*, in IEEE Spoken Language Technology Workshop, 2018.

[106] ——, *Acvae-vc: Non-parallel voice conversion with auxiliary classifier variational autoencoder*, IEEE Transactions on Audio, Speech, and Language Processing, 27 (2019), pp. 1432–1443.

[107] T. KANEKO, H. KAMEOKA, K. TANAKA, AND N. HOJO, *Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion*, in Conference of the International Speech Communication Association, 2019.

[108] T. KARRAS, T. AILA, S. LAINE, AND J. LEHTINEN, *Progressive growing of gans for improved quality, stability, and variation*, in International Conference on Learning Representations, 2018.

[109] T. KARRAS, S. LAINE, AND T. AILA, *A style-based generator architecture for generative adversarial networks*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[110] L. KE, X. LI, Y. BISK, A. HOLTZMAN, Z. GAN, J. LIU, J. GAO, Y. CHOI, AND S. S. SRINIVASA, *Tactical rewind: Self-correction via backtracking in vision-and-language navigation*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[111] P. KHOSLA, P. TETERWAK, C. WANG, A. SARNA, Y. TIAN, P. ISOLA, A. MASCHINOT, C. LIU, AND D. KRISHNAN, *Supervised contrastive learning*, in Advances in Neural Information Processing Systems, 2020.

[112] J. KIM, M. GOPAKUMAR, S. CHOI, Y. PENG, W. LOPES, AND G. WETZSTEIN, *Holographic glasses for virtual reality*, in Special Interest Group on Graphics and Interactive Techniques, 2022.

[113] S. KIM, T. HORI, AND S. WATANABE, *Joint ctc-attention based end-to-end speech recognition using multi-task learning*, in International Conference on Acoustics, Speech, Signal Processing, 2017.

[114] T. KIM, M. CHA, H. KIM, J. K. LEE, AND J. KIM, *Learning to discover cross-domain relations with generative adversarial networks*, in International Conference on Machine Learning, 2017.

[115] D. P. KINGMA AND M. WELLING, *Auto-encoding variational bayes*, in International Conference on Learning Representations, 2014.

[116] T. KO, V. PEDDINTI, D. POVEY, M. L. SELTZER, AND S. KHUDANPUR, *A study on data augmentation of reverberant speech for robust speech recognition*, in International Conference on Acoustics, Speech, Signal Processing, 2017.

[117] K. KOBAYASHI, T. TODA, G. NEUBIG, S. SAKTI, AND S. NAKAMURA, *Statistical singing voice conversion based on direct waveform modification with global variance*, in Conference of the International Speech Communication Association, 2015.

[118] O. KOLLER, N. C. CAMGÖZ, H. NEY, AND R. BOWDEN, *Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 42 (2020), pp. 2306–2320.

[119] H. KON AND H. KOIKE, *Estimation of late reverberation characteristics from a single two-dimensional environmental image using convolutional neural networks*, Journal of the Audio Engineering Society, 67 (2019), pp. 540–548.

[120] J. KONG, J. KIM, AND J. BAE, *Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis*, in Advances in Neural Information Processing Systems, 2020.

[121] Z. KONG, W. PING, J. HUANG, K. ZHAO, AND B. CATANZARO, *Diffwave: A versatile diffusion model for audio synthesis*, in International Conference on Learning Representations, 2021.

[122] A. A. KRESSNER, A. WESTERMANN, AND J. M. BUCHHOLZ, *The impact of reverberation on speech intelligibility in cochlear implant recipients*, The Journal of the Acoustical Society of America, 144 (2018), pp. 1113–1122.

[123] T. KUDO AND J. RICHARDSON, *Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*, in Conference on Empirical Methods in Natural Language Processing, 2018.

[124] H. KUTTRUFF AND E. MOMMERTZ, *Room acoustics*, in Handbook of engineering acoustics, London, U.K.: Springer, 2012, pp. 239–267.

[125] Z. LAN, M. CHEN, S. GOODMAN, K. GIMPEL, P. SHARMA, AND R. SORICUT, *ALBERT: A lite BERT for self-supervised learning of language representations*, in International Conference on Learning Representations, 2020.

[126] A. B. L. LARSEN, S. K. SØNDERBY, H. LAROCHELLE, AND O. WINTHER, *Autoencoding beyond pixels using a learned similarity metric*, in International Conference on Machine Learning, 2016.

[127] D. LEE, C. KIM, S. KIM, M. CHO, AND W. HAN, *Autoregressive image generation using residual quantization*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[128] K. LEE, C. PARK, N. KIM, AND J. LEE, *Accelerating recurrent neural network language model based online speech recognition system*, in International Conference on Acoustics, Speech, Signal Processing, 2018.

[129] S. LEE, W. PING, B. GINSBURG, B. CATANZARO, AND S. YOON, *Bigvgan: A universal neural vocoder with large-scale training*, in International Conference on Learning Representations, 2023.

[130] S. E. LEVINSON, L. R. RABINER, AND M. M. SONDHI, *An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition*, Bell Labs Technical Journal, 62 (1983), pp. 1035–1074.

[131] M. LEWIS, Y. LIU, N. GOYAL, M. GHAZVININEJAD, A. MOHAMED, O. LEVY, V. STOYANOV, AND L. ZETTLEMOYER, *BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, in Association for Computational Linguistics, 2020.

[132] P. S. H. LEWIS, E. PEREZ, A. PIKTUS, F. PETRONI, V. KARPUKHIN, N. GOYAL, H. KÜTTLER, M. LEWIS, W. YIH, T. ROCKTÄSCHEL, S. RIEDEL, AND D. KIELA, *Retrieval-augmented generation for knowledge-intensive NLP tasks*, in Advances in Neural Information Processing Systems, 2020.

[133] D. LI, C. XU, X. YU, K. ZHANG, B. SWIFT, H. SUOMINEN, AND H. LI, *Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation*, in Advances in Neural Information Processing Systems, 2020.

[134] D. LI, X. YU, C. XU, L. PETERSSON, AND H. LI, *Transferring cross-domain knowledge for video sign language recognition*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[135] Y. A. LI, A. ZARE, AND N. MESGARANI, *Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion*, in Conference of the International Speech Communication Association, 2021.

[136] S. K. LIDDELL AND R. E. JOHNSON, *American sign language: The phonological base*, Sign language studies, 64 (1989), pp. 195–277.

[137] C.-Y. LIN, *Rouge: A package for automatic evaluation of summaries*, in Text summarization branches out, 2004, pp. 74–81.

[138] K. LIN, X. WANG, L. ZHU, K. SUN, B. ZHANG, AND Y. YANG, *Gloss-free end-to-end sign language translation*, in Association for Computational Linguistics, 2023.

[139] T.-Q. LIN, H.-Y. LEE, AND H. TANG, *Melhubert: A simplified hubert on mel spectrograms*, in ASRU, 2023.

[140] X. V. LIN, X. CHEN, M. CHEN, W. SHI, M. LOMELI, R. JAMES, P. RODRIGUEZ, J. KAHN, G. SZILVASY, M. LEWIS, L. ZETTLEMOYER, AND S. YIH, *RA-DIT: retrieval-augmented dual instruction tuning*, in International Conference on Learning Representations, 2024.

[141] L. LIU, Y. LU, M. YANG, Q. QU, J. ZHU, AND H. LI, *Generative adversarial network for abstractive text summarization*, in Association for the Advancement of Artificial Intelligence, 2018.

[142] R. LIU, Y. GE, C. L. CHOI, X. WANG, AND H. LI, *Divco: Diverse conditional image synthesis via contrastive generative adversarial network*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[143] S. LIU, Y. CAO, S. KANG, N. HU, X. LIU, D. SU, D. YU, AND H. MENG, *Transferring source style in non-parallel voice conversion*, in Conference of the International Speech Communication Association, 2020.

[144] X. LIU, D. H. PARK, S. AZADI, G. ZHANG, A. CHOPIKYAN, Y. HU, H. SHI, A. ROHRBACH, AND T. DARRELL, *More control for free! image synthesis with semantic diffusion guidance*, in IEEE/CVF Winter Conference on Applications of Computer Vision, 2023.

[145] Y. LIU, M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER, AND V. STOYANOV, *Roberta: A robustly optimized BERT pretraining approach*, CoRR, abs/1907.11692 (2019).

[146] C. LO, S. FU, W. HUANG, X. WANG, J. YAMAGISHI, Y. TSAO, AND H. WANG, *Mosnet: Deep learning-based objective assessment for voice conversion*, in Conference of the International Speech Communication Association, 2019.

[147] B. LOGAN ET AL., *Mel frequency cepstral coefficients for music modeling.*, in ISMIR, 2000.

[148] Z. LONG, Y. ZHENG, M. YU, AND J. XIN, *Enhancing zero-shot many to many voice conversion via self-attention VAE with structurally regularized layers*, in International Conference on Artificial Intelligence for Industries, 2022.

[149] I. LOSHCHILOV AND F. HUTTER, *Decoupled weight decay regularization*, in International Conference on Learning Representations, 2019.

[150] J. LU, A. KANNAN, J. YANG, D. PARIKH, AND D. BATRA, *Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model*, in Advances in Neural Information Processing Systems, 2017.

[151] X. LU, W. WANG, J. SHEN, Y.-W. TAI, D. J. CRANDALL, AND S. C. HOI, *Learning video object segmentation from unlabeled videos*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[152] A. LUGMAYR, M. DANELLJAN, A. ROMERO, F. YU, R. TIMOFTE, AND L. V. GOOL, *Repaint: Inpainting using denoising diffusion probabilistic models*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[153] Y. LUO AND Y. YANG, *Large language model and domain-specific model collaboration for smart education*, Frontiers of Information Technology & Electronic Engineering, 25 (2024), pp. 333–341.

[154] H. LUONG AND J. YAMAGISHI, *NAUTILUS: A versatile voice cloning system*, IEEE Transactions on Audio, Speech, and Language Processing, 28 (2020), pp. 2967–2981.

[155] J. MA, W. WANG, Y. YANG, AND F. ZHENG, *Ms2sl: Multimodal spoken data-driven continuous sign language production*, in Association for Computational Linguistics, 2024.

[156] ——, *Mutual learning for acoustic matching and dereverberation via visual scene-driven diffusion*, in European Conference on Computer Vision, 2024.

[157] J. MA, Z. ZHENG, H. FEI, F. ZHENG, T.-S. CHUA, AND Y. YANG, *Subband-based generative adversarial network for non-parallel many-to-many voice conversion*, arXiv preprint arXiv:2207.06057, (2022).

[158] W. MACK, S. DENG, AND E. A. P. HABETS, *Single-channel blind direct-to-reverberation ratio estimation using masking*, in Conference of the International Speech Communication Association, 2020.

[159] A. MALLEN, A. ASAI, V. ZHONG, R. DAS, D. KHASHABI, AND H. HAJISHIRZI, *When not to trust language models: Investigating effectiveness of parametric and non-parametric memories*, in Association for Computational Linguistics, 2023.

[160] S. MALPICA, A. SERRANO, J. GUERRERO-VIU, D. MARTIN, E. BERNAL, D. GUTIERREZ, AND B. MASIÁ, *Auditory stimuli degrade visual performance in virtual reality*, in Special Interest Group on Graphics and Interactive Techniques, 2022.

[161] W. MANN, C. R. MARSHALL, K. MASON, AND G. MORGAN, *The acquisition of sign language: The impact of phonetic complexity on phonology*, Language Learning and Development, 6 (2010), pp. 60–86.

[162] X. MAO, Q. LI, H. XIE, R. Y. LAU, Z. WANG, AND S. PAUL SMOLLEY, *Least squares generative adversarial networks*, in International Conference on Computer Vision, 2017.

[163] S. MATHEW, S. NADEEM, S. KUMARI, AND A. E. KAUFMAN, *Augmenting colonoscopy using extended and directional cyclegan for lossy image translation*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[164] S. A. MEHDI AND Y. N. KHAN, *Sign language recognition using sensor gloves*, 2002.

[165] C. MENG, Y. HE, Y. SONG, J. SONG, J. WU, J. ZHU, AND S. ERMON, *Sdedit: Guided image synthesis and editing with stochastic differential equations*, in International Conference on Learning Representations, 2022.

[166] L. MENG, J. XU, X. TAN, J. WANG, T. QIN, AND B. XU, *Mixspeech: Data augmentation for low-resource automatic speech recognition*, in Conference of the International Speech Communication Association, 2021.

[167] Y. MIN, A. HAO, X. CHAI, AND X. CHEN, *Visual alignment constraint for continuous sign language recognition*, in International Conference on Computer Vision, 2021.

[168] M. MIRZA AND S. OSINDERO, *Conditional generative adversarial nets*, CoRR, abs/1411.1784 (2014).

[169] G. MITTAL, J. H. ENGEL, C. HAWTHORNE, AND I. SIMON, *Symbolic music generation with diffusion models*, in International Society for Music Information Retrieval Conference, 2021.

[170] T. MIYATO, S. MAEDA, M. KOYAMA, AND S. ISHII, *Virtual adversarial training: A regularization method for supervised and semi-supervised learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41 (2019), pp. 1979–1993.

[171] M. MIYOSHI AND Y. KANEDA, *Inverse filtering of room acoustics*, IEEE Transactions on Audio, Speech, and Language Processing, 36 (1988), pp. 145–152.

[172] M. P. MOELLER, *Early intervention and language development in children who are deaf and hard of hearing*, Pediatrics, 106 (2000), pp. e43–e43.

[173] S. M. MOHAMMAD, *NLP scholar: An interactive visual explorer for natural language processing literature*, in Association for Computational Linguistics, 2020.

[174] S. H. MOHAMMADI AND A. KAIN, *An overview of voice conversion systems*, Speech Communication, 88 (2017), pp. 65–82.

[175] J. X. MORRIS, V. KULESHOV, V. SHMATIKOV, AND A. M. RUSH, *Text embeddings reveal (almost) as much as text*, in Conference on Empirical Methods in Natural Language Processing, 2023.

[176] M. MÜLLER, *Fundamentals of music processing: Audio, analysis, algorithms, applications*, vol. 5, Springer, 2015.

[177] M. MÜLLER, Z. JIANG, A. MORYOSSEF, A. RIOS, AND S. EBLING, *Considerations for meaningful sign language machine translation based on glosses*, in Association for Computational Linguistics, 2023.

[178] P. MURGAI, M. RAU, AND J.-M. JOT, *Blind estimation of the reverberation fingerprint of unknown acoustic environments*, in Audio Engineering Society Convention 143, Audio Engineering Society, 2017.

[179] T. NAKATANI, C. BÖDDEKER, K. KINOSHITA, R. IKESHITA, M. DELCROIX, AND R. HAEB-UMBACH, *Jointly optimal denoising, dereverberation, and source separation*, IEEE Transactions on Audio, Speech, and Language Processing, 28 (2020), pp. 2267–2282.

[180] T. NAKATANI, T. YOSHIOKA, K. KINOSHITA, M. MIYOSHI, AND B.-H. JUANG, *Speech dereverberation based on variance-normalized delayed linear prediction*, IEEE Transactions on Audio, Speech, and Language Processing, 18 (2010), pp. 1717–1731.

[181] P. A. NAYLOR, N. D. GAUBITCH, ET AL., *Speech dereverberation*, vol. 2, Springer, 2010.

[182] A. Q. NICHOL AND P. DHARIWAL, *Improved denoising diffusion probabilistic models*, in International Conference on Machine Learning, 2021.

[183] A. Q. NICHOL, P. DHARIWAL, A. RAMESH, P. SHYAM, P. MISHKIN, B. MCGREW, I. SUTSKEVER, AND M. CHEN, *GLIDE: towards photorealistic image generation and editing with text-guided diffusion models*, in International Conference on Machine Learning, 2022.

[184] Z. NIU AND B. MAK, *Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition*, in European Conference on Computer Vision, 2020.

[185] A. NÚÑEZ-MARCOS, O. PEREZ-DE-VIÑASPRE, AND G. LABAKA, *A survey on sign language machine translation*, Expert Systems with Applications, 213 (2023), p. 118993.

[186] A. V. D. OORD, Y. LI, AND O. VINYALS, *Representation learning with contrastive predictive coding*, CoRR, abs/1807.03748 (2018).

[187] L. OUYANG, J. WU, X. JIANG, D. ALMEIDA, C. L. WAINWRIGHT, P. MISHKIN, C. ZHANG, S. AGARWAL, K. SLAMA, A. RAY, J. SCHULMAN, J. HILTON, F. KELTON, L. MILLER, M. SIMENS, A. ASKELL, P. WELINDER, P. F. CHRISTIANO, J. LEIKE, AND R. LOWE, *Training language models to follow instructions with human feedback*, in Advances in Neural Information Processing Systems, 2022.

[188] C. A. PADDEN, *Interaction of morphology and syntax in American Sign Language*, Routledge, 2016.

[189] C. A. PADDEN AND T. L. HUMPHRIES, *Deaf in America*, Harvard University Press, 1988.

[190] D. PALAZ, M. MAGIMAI-DOSS, AND R. COLLOBERT, *Analysis of cnn-based speech recognition system using raw speech as input*, in Conference of the International Speech Communication Association, 2015.

[191] V. PANAYOTOV, G. CHEN, D. POVEY, AND S. KHUDANPUR, *Librispeech: an asr corpus based on public domain audio books*, in International Conference on Acoustics, Speech, Signal Processing, 2015.

[192] V. PANAYOTOV, G. CHEN, D. POVEY, AND S. KHUDANPUR, *Librispeech: An ASR corpus based on public domain audio books*, in International Conference on Acoustics, Speech, Signal Processing, 2015.

[193] K. PAPINENI, S. ROUKOS, T. WARD, AND W. ZHU, *Bleu: a method for automatic evaluation of machine translation*, in Association for Computational Linguistics, 2002.

[194] D. S. PARK, W. CHAN, Y. ZHANG, C. CHIU, B. ZOPH, E. D. CUBUK, AND Q. V. LE, *Specaugment: A simple data augmentation method for automatic speech recognition*, in Conference of the International Speech Communication Association, 2019.

[195] D. S. PARK, Y. ZHANG, Y. JIA, W. HAN, C. CHIU, B. LI, Y. WU, AND Q. V. LE, *Improved noisy student training for automatic speech recognition*, in Conference of the International Speech Communication Association, 2020.

[196] D. PAUL, M. P. SHIFAS, Y. PANTAZIS, AND Y. STYLIANOU, *Enhancing Speech Intelligibility in Text-To-Speech Synthesis Using Speaking Style Conversion*, in Conference of the International Speech Communication Association, 2020.

[197] M. E. PETERS, M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE, AND L. ZETTLEMOYER, *Deep contextualized word representations*, in North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.

[198] L.-A. PETITTO, C. LANGDON, A. STONE, D. ANDRIOLA, G. KARTHEISER, AND C. COCHRAN, *Visual sign phonology: Insights into human reading and language from a natural soundless phonology*, Wiley Interdisciplinary Reviews: Cognitive Science, 7 (2016), pp. 366–381.

[199] S. S. R. PHAYE, E. BENETOS, AND Y. WANG, *Subspectralnet–using sub-spectrogram based convolutional neural networks for acoustic scene classification*, in ICASSP, 2019.

[200] M. PLAPPERT, C. MANDERY, AND T. ASFOUR, *The KIT motion-language dataset*, Big Data, 4 (2016), pp. 236–252.

[201] M. POBAR AND I. IPŠIĆ, *Online speaker de-identification using voice transformation*, in International Convention on Information and Communication Technology, Electronics and Microelectronics, 2014.

[202] A. POLYAK, L. WOLF, Y. ADI, AND Y. TAIGMAN, *Unsupervised cross-domain singing voice conversion*, in Conference of the International Speech Communication Association, 2020.

[203] L. PRIMEWORDS INFORMATION TECHNOLOGY CO., *Primewords chinese corpus set 1*, 2018.
https://www.primewords.cn.

[204] J. PU, W. ZHOU, AND H. LI, *Iterative alignment network for continuous sign language recognition*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[205] K. QIAN, Y. ZHANG, S. CHANG, J. XIONG, C. GAN, D. D. COX, AND M. HASEGAWA-JOHNSON, *Global rhythm style transfer without text transcriptions*, CoRR, abs/2106.08519 (2021).

[206] K. QIAN, Y. ZHANG, S. CHANG, X. YANG, AND M. HASEGAWA-JOHNSON, *Autovc: Zero-shot voice style transfer with only autoencoder loss*, in International Conference on Machine Learning, 2019.

[207] A. RADFORD, J. W. KIM, C. HALLACY, A. RAMESH, G. GOH, S. AGARWAL, G. SASTRY, A. ASKELL, P. MISHKIN, J. CLARK, G. KRUEGER, AND I. SUTSKEVER, *Learning transferable visual models from natural language supervision*, in International Conference on Machine Learning, 2021.

[208] A. RADFORD, J. W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, AND I. SUTSKEVER, *Robust speech recognition via large-scale weak supervision*, in International Conference on Machine Learning, 2023.

[209] A. RADFORD, L. METZ, AND S. CHINTALA, *Unsupervised representation learning with deep convolutional generative adversarial networks*, in International Conference on Learning Representations, 2016.

[210] A. RADFORD, K. NARASIMHAN, T. SALIMANS, I. SUTSKEVER, ET AL., *Improving language understanding by generative pre-training*, (2018).

[211] A. RADFORD, J. WU, R. CHILD, D. LUAN, D. AMODEI, I. SUTSKEVER, ET AL., *Language models are unsupervised multitask learners*, OpenAI blog, 1 (2019), p. 9.

[212] C. RAFFEL, N. SHAZEER, A. ROBERTS, K. LEE, S. NARANG, M. MATENA, Y. ZHOU, W. LI, AND P. J. LIU, *Exploring the limits of transfer learning with a unified text-to-text transformer*, Journal of Machine Learning Research, 21 (2020), pp. 140:1–140:67.

[213] K. RAO, C. HARRIS, A. IRPAN, S. LEVINE, J. IBARZ, AND M. KHANSARI, *Rl-cyclegan: Reinforcement learning aware simulation-to-real*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[214] R. RASTGOO, K. KIANI, AND S. ESCALERA, *Sign language recognition: A deep survey*, Expert Systems with Applications, 164 (2021), p. 113794.

[215] A. RAZAVI, A. VAN DEN OORD, AND O. VINYALS, *Generating diverse high-fidelity images with VQ-VAE-2*, in Advances in Neural Information Processing Systems, 2019.

[216] J. RENNIES, T. BRAND, AND B. KOLLMEIER, *Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet*, The Journal of the Acoustical Society of America, 130 (2011), pp. 2999–3012.

[217] A. W. RIX, J. G. BEERENDS, M. P. HOLLIER, AND A. P. HEKSTRA, *Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs*, in International Conference on Acoustics, Speech, Signal Processing, 2001.

[218] R. ROMBACH, A. BLATTMANN, D. LORENZ, P. ESSER, AND B. OMMER, *High-resolution image synthesis with latent diffusion models*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[219] C. SAHARIA, W. CHAN, S. SAXENA, L. LI, J. WHANG, E. L. DENTON, S. K. S. GHASEMIPOUR, R. G. LOPES, B. K. AYAN, T. SALIMANS, J. HO, D. J. FLEET, AND M. NOROUZI, *Photorealistic text-to-image diffusion models with deep language understanding*, in Advances in Neural Information Processing Systems, 2022.

[220] M. SAJJADI, M. JAVANMARDI, AND T. TASDIZEN, *Regularization with stochastic transformations and perturbations for deep semi-supervised learning*, in Advances in Neural Information Processing Systems, 2016.

[221] S. SAKAMOTO, A. TANIGUCHI, T. TANIGUCHI, AND H. KAMEOKA, *Stargan-vc+asr: Stargan-based non-parallel voice conversion regularized by automatic speech recognition*, in Conference of the International Speech Communication Association, 2021.

[222] W. SANDLER, *The phonological organization of sign languages*, Lang. Linguistics Compass, 6 (2012), pp. 162–182.

[223] W. SANDLER, *The challenge of sign language phonology*, Annual review of Linguistics, 3 (2017), pp. 43–63.

[224] W. SANDLER AND D. C. LILLO-MARTIN, *Sign language and linguistic universals*, Cambridge University Press, 2006.

[225] A. SARROFF AND R. MICHAELS, *Blind arbitrary reverb matching*, 2020.

[226] B. SAUNDERS, N. C. CAMGÖZ, AND R. BOWDEN, *Progressive transformers for end-to-end sign language production*, in European Conference on Computer Vision, 2020.

[227] ——, *Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks*, International Journal of Computer Vision, 129 (2021), pp. 2113–2135.

[228] ——, *Mixed signals: Sign language production via a mixture of motion primitives*, in International Conference on Computer Vision, 2021.

[229] ——, *Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[230] G.-L. SAVINO, J. MORAES BRAGA, AND J. SCHÖNING, *Velocity: Using voice assistants for cyclists to provide traffic reports*, in ACM International Conference on Multimedia, 2021.

[231] L. SAVIOJA AND N. XIANG, *Simulation-based auralization of room acoustics*, Acoust. Today, 16 (2020), pp. 48–55.

[232] M. R. SCHROEDER, *Natural sounding artificial reverberation*, in Audio Engineering Society Convention, 1961.

[233] M. R. SCHROEDER AND B. F. LOGAN, *"colorless" artificial reverberation*, IRE Transactions on Audio, (1961), pp. 209–214.

[234] P. SELVARAJ, G. N. C., P. KUMAR, AND M. M. KHAPRA, *Openhands: Making sign language recognition accessible with pose-based pretrained models across languages*, in Association for Computational Linguistics, 2022.

[235] J. SERRÀ, S. PASCUAL, AND C. SEGURA PERALES, *Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion*, in Advances in Neural Information Processing Systems, 2019.

[236] M. SHAH, X. CHEN, M. ROHRBACH, AND D. PARIKH, *Cycle-consistency for robust visual question answering*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[237] F. SHI, X. CHEN, K. MISRA, N. SCALES, D. DOHAN, E. H. CHI, N. SCHÄRLI, AND D. ZHOU, *Large language models can be easily distracted by irrelevant context*, in International Conference on Machine Learning, 2023.

[238] Y. SHI, H. BU, X. XU, S. ZHANG, AND M. LI, *Aishell-3: A multi-speaker mandarin tts corpus and the baselines*, CoRR, abs/2010.11567 (2020).

[239] A. SINGH, R. HU, V. GOSWAMI, G. COUAIRON, W. GALUBA, M. ROHRBACH, AND D. KIELA, *FLAVA: A foundational language and vision alignment model*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[240] N. SINGH, J. MENTCH, J. NG, M. BEVERIDGE, AND I. DRORI, *Image2reverb: Cross-modal reverb impulse response synthesis*, in International Conference on Computer Vision, 2021.

[241] B. SISMAN, K. VIJAYAN, M. DONG, AND H. LI, *Singan: Singing voice conversion with generative adversarial networks*, in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2019.

[242] B. SISMAN, J. YAMAGISHI, S. KING, AND H. LI, *An overview of voice conversion and its challenges: From statistical modeling to deep learning*, IEEE Transactions on Audio, Speech, and Language Processing, 29 (2021), pp. 132–157.

[243] L. SIYAO, W. YU, T. GU, C. LIN, Q. WANG, C. QIAN, C. C. LOY, AND Z. LIU, *Bailando: 3d dance generation by actor-critic GPT with choreographic memory*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[244] J. SOHL-DICKSTEIN, E. A. WEISS, N. MAHESWARANATHAN, AND S. GANGULI, *Deep unsupervised learning using nonequilibrium thermodynamics*, in International Conference on Machine Learning, 2015.

[245] A. SOMAYAZULU, C. CHEN, AND K. GRAUMAN, *Self-supervised visual acoustic matching*, in Advances in Neural Information Processing Systems, 2023.

[246] S. S. STEVENS, J. VOLKMANN, AND E. B. NEWMAN, *A scale for the measurement of the psychological magnitude pitch*, The journal of the acoustical society of america, 8 (1937), pp. 185–190.

[247] W. C. STOKOE JR, *Sign language structure: An outline of the visual communication systems of the american deaf*, Journal of deaf studies and deaf education, 10 (2005), pp. 3–37.

[248] J. SU, Z. JIN, AND A. FINKELSTEIN, *Acoustic matching by embedding impulse responses*, in International Conference on Acoustics, Speech, Signal Processing, 2020.

[249] L. SUN, K. LI, H. WANG, S. KANG, AND H. MENG, *Phonetic posteriorgrams for many-to-one voice conversion without parallel data training*, 2016.

[250] Y. SUN, S. WANG, Y. LI, S. FENG, H. TIAN, H. WU, AND H. WANG, *ERNIE 2.0: A continual pre-training framework for language understanding*, in Association for the Advancement of Artificial Intelligence, 2020.

[251] R. SUNKARA AND T. LUO, *No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects*, 2022.

[252] H. TAN, L. YU, AND M. BANSAL, *Learning to navigate unseen environments: Back translation with environmental dropout*, 2019.

[253] K. TAN, Y. XU, S. ZHANG, M. YU, AND D. YU, *Audio-visual speech separation and dereverberation with a two-stage multimodal network*, IEEE Journal of Selected Topics in Signal Processing, 14 (2020), pp. 542–553.

[254] K. TANAKA, H. KAMEOKA, T. KANEKO, AND N. HOJO, *Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms*, in International Conference on Acoustics, Speech, Signal Processing, 2019.

[255] M. TASKIRAN, M. KILLIOGLU, AND N. KAHRAMAN, *A real-time system for recognition of american sign language by using deep learning*, in Telecommunications and Signal Processing, 2018.

[256] T. TODA, A. W. BLACK, AND K. TOKUDA, *Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory*, IEEE Transactions on Audio, Speech, and Language Processing, 15 (2007), pp. 2222–2235.

[257] D. TRAN, L. D. BOURDEV, R. FERGUS, L. TORRESANI, AND M. PALURI, *Learning spatiotemporal features with 3d convolutional networks*, in International Conference on Computer Vision, 2015.

[258] V. VÄLIMÄKI, J. D. PARKER, L. SAVIOJA, J. O. S. III, AND J. S. ABEL, *Fifty years of artificial reverberation*, IEEE Transactions on Audio, Speech, and Language Processing, 20 (2012), pp. 1421–1448.

[259] A. VAN DEN OORD, O. VINYALS, AND K. KAVUKCUOGLU, *Neural discrete representation learning*, in Advances in Neural Information Processing Systems, 2017.

[260] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER, AND I. POLOSUKHIN, *Attention is all you need*, in Advances in Neural Information Processing Systems, 2017.

[261] R. VERGIN, D. O'SHAUGHNESSY, AND A. FARHAT, *Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition*, IEEE Transactions on speech and audio processing, 7 (1999), pp. 525–532.

[262] A. WAIBEL, T. HANAZAWA, G. E. HINTON, K. SHIKANO, AND K. J. LANG, *Phoneme recognition using time-delay neural networks*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 37 (1989), pp. 328–339.

[263] H. WALSH, B. SAUNDERS, AND R. BOWDEN, *Changing the representation: Examining language representation for neural sign language production*, CoRR, abs/2210.06312 (2022).

[264] C. WANG AND Y.-B. YU, *Cyclegan-vc-gp: Improved cyclegan-based non-parallel voice conversion*, in International Conference on Communication Technology, 2020.

[265] D. WANG AND X. ZHANG, *THCHS-30 : A free chinese speech corpus*, CoRR, abs/1512.01882 (2015).

[266] H. WANG, W. LIANG, J. SHEN, L. V. GOOL, AND W. WANG, *Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[267] N. WANG, Y. SONG, C. MA, W. ZHOU, W. LIU, AND H. LI, *Unsupervised deep tracking*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[268] P. WANG, A. YANG, R. MEN, J. LIN, S. BAI, Z. LI, J. MA, C. ZHOU, J. ZHOU, AND H. YANG, *OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework*, in International Conference on Machine Learning, 2022.

[269] Y. WANG, Y. XIA, T. HE, F. TIAN, T. QIN, C. ZHAI, AND T. LIU, *Multi-agent dual learning*, in International Conference on Learning Representations, 2019.

[270] B. WU, K. LI, F. GE, Z. HUANG, M. YANG, S. M. SINISCALCHI, AND C.-H. LEE, *An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition*, IEEE Journal of Selected Topics in Signal Processing, 11 (2017), pp. 1289–1300.

[271] B. WU, K. LI, M. YANG, AND C. LEE, *A reverberation-time-aware approach to speech dereverberation based on deep neural networks*, IEEE Transactions on Audio, Speech, and Language Processing, 25 (2017), pp. 98–107.

[272] H. WU, J. JIA, H. WANG, Y. DOU, C. DUAN, AND Q. DENG, *Imitating arbitrary talking style for realistic audio-driven talking face synthesis*, in ACM International Conference on Multimedia, 2021.

[273] S. WU, H. FEI, Y. REN, D. JI, AND J. LI, *Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge*, in International Joint Conferences on Artificial Intelligence, 2021.

[274] T. WU, Z. FAN, X. LIU, Y. GONG, Y. SHEN, J. JIAO, H. ZHENG, J. LI, Z. WEI, J. GUO, N. DUAN, AND W. CHEN, *Ar-diffusion: Auto-regressive diffusion model for text generation*, in Advances in Neural Information Processing Systems, 2023.

[275] F.-L. XIE, F. K. SOONG, AND H. LI, *A kl divergence and dnn-based approach to voice conversion without parallel training sentences*, in Conference of the International Speech Communication Association, 2016.

[276] Q. XIE, Z. DAI, E. H. HOVY, T. LUONG, AND Q. LE, *Unsupervised data augmentation for consistency training*, in Advances in Neural Information Processing Systems, 2020.

[277] Q. XIE, X. TIAN, G. LIU, K. SONG, L. XIE, Z. WU, H. LI, S. SHI, H. LI, F. HONG, H. BU, AND X. XU, *The multi-speaker multi-style voice cloning challenge 2021*, in International Conference on Acoustics, Speech, Signal Processing, 2021.

[278] J. YAMAGISHI, C. VEAUX, K. MACDONALD, ET AL., *Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)*, (2019).

[279] Z. YANG, Z. DAI, Y. YANG, J. G. CARBONELL, R. SALAKHUTDINOV, AND Q. V. LE, *Xlnet: Generalized autoregressive pretraining for language understanding*, in Advances in Neural Information Processing Systems, 2019.

[280] Z. YAO, D. WU, X. WANG, B. ZHANG, F. YU, C. YANG, Z. PENG, X. CHEN, L. XIE, AND X. LEI, *Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit*, in Conference of the International Speech Communication Association, 2021.

[281] Z. YI, H. R. ZHANG, P. TAN, AND M. GONG, *Dualgan: Unsupervised dual learning for image-to-image translation*, in International Conference on Computer Vision, 2017.

[282] Q. YIN, W. TAO, X. LIU, AND Y. HONG, *Neural sign language translation with sf-transformer*, in International Conference on Innovation in Artificial Intelligence, 2022.

[283] C. YOSHINAGA-ITANO, *From screening to early identification and intervention: Discovering predictors to successful outcomes for children with significant hearing loss*, Journal of deaf studies and deaf education, 8 (2003), pp. 11–30.

[284] L. YU, W. ZHANG, J. WANG, AND Y. YU, *Seqgan: Sequence generative adversarial nets with policy gradient*, in Association for the Advancement of Artificial Intelligence, 2017.

[285] H. YUAN, Z. YUAN, C. TAN, F. HUANG, AND S. HUANG, *Seqdiffuseq: Text diffusion with encoder-decoder transformers*, CoRR, abs/2212.10325 (2022).

[286] D. ZENG, H. LIU, H. LIN, AND S. GE, *Talking face generation with expression-tailored generative adversarial network*, in ACM International Conference on Multimedia, 2020.

[287] H. ZHANG AND Y. LIN, *Improve few-shot voice cloning using multi-modal learning*, in International Conference on Acoustics, Speech, Signal Processing, 2022.

[288] J. ZHANG, Y. ZHANG, X. CUN, Y. ZHANG, H. ZHAO, H. LU, X. SHEN, AND S. YING, *Generating human motion from textual descriptions with discrete representations*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[289] J. ZHANG, W. ZHOU, C. XIE, J. PU, AND H. LI, *Chinese sign language recognition with adaptive HMM*, in IEEE International Conference on Multimedia and Expo, 2016.

[290] Y. ZHANG, T. XIANG, T. M. HOSPEDALES, AND H. LU, *Deep mutual learning*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[291] Z. ZHANG, X. HAN, Z. LIU, X. JIANG, M. SUN, AND Q. LIU, *ERNIE: enhanced language representation with informative entities*, in Association for Computational Linguistics, 2019.

[292] Y. ZHAO, W. HUANG, X. TIAN, J. YAMAGISHI, R. K. DAS, T. KINNUNEN, Z. LING, AND T. TODA, *Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion*, CoRR, abs/2008.12527 (2020).

[293] Y. ZHAO, M. KURUVILLA-DUGDALE, AND M. SONG, *Voice conversion for persons with amyotrophic lateral sclerosis*, IEEE journal of biomedical and health informatics, 24 (2019), pp. 2942–2949.

[294] Y. ZHAO, D. WANG, B. XU, AND T. ZHANG, *Monaural speech dereverberation using temporal convolutional networks with self attention*, IEEE Transactions on Audio, Speech, and Language Processing, 28 (2020), pp. 1598–1607.

[295] Y. ZHAO, Z. WANG, AND D. WANG, *Two-stage deep learning for noisy-reverberant speech enhancement*, IEEE Transactions on Audio, Speech, and Language Processing, 27 (2019), pp. 53–62.

[296] Z. ZHAO, Y. XIA, T. QIN, L. XIA, AND T. LIU, *Dual learning: Theoretical study and an algorithmic extension*, in Asian Conference on Machine Learning, 2020.

[297] C. ZHENG, T. VUONG, J. CAI, AND D. PHUNG, *Movq: Modulating quantized vectors for high-fidelity image generation*, in Advances in Neural Information Processing Systems, 2022.

[298] H. ZHOU, W. ZHOU, W. QI, J. PU, AND H. LI, *Improving sign language translation with monolingual data by sign back-translation*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[299] H. ZHOU, W. ZHOU, Y. ZHOU, AND H. LI, *Spatial-temporal multi-cue network for continuous sign language recognition*, in Association for the Advancement of Artificial Intelligence, 2020.

[300] ——, *Spatial-temporal multi-cue network for sign language recognition and translation*, IEEE Transactions on Multimedia, 24 (2022), pp. 768–779.

[301] J.-Y. ZHU, T. PARK, P. ISOLA, AND A. A. EFROS, *Unpaired image-to-image translation using cycle-consistent adversarial networks*, in International Conference on Computer Vision, 2017.