Efficient Generative Multimodal Integration (EGMI): Enabling Audio Generation from Text-Image Pairs through Alignment with Large Language Models

Taemin Kim[*]
Hansung University
taemin6697@gmail.com

Wooyeol Baek[*]
Soongsil University
100wooyeol@gmail.com

Heeseok Oh[†]
Hansung University
ohhs@hansung.ac.kr

Abstract

Multimodal large language models (MLLM) face challenges in leveraging their rich knowledge since spanning different modalities is nontrivial and their contextual ambiguity arises from lack of paired data. In the context of audio generation based on MLLM, the annotation of audio-text paired datasets demands significant human resources due to the complexity of audio data, making such datasets much scarcer and harder to access compared to image-text paired datasets. To address these issues, we propose a novel technique called *efficient generative multimodal integration (EGMI)*, which enables audio generation tasks leveraging only image-text data. Based on pretrained LLM's powerful knowledge on text comprehension, EGMI successfully leverages image-text paired datasets for cross-modal alignment, enabling interactions between audio and image information. We also introduce an efficient mapping network, called the EGMI mapper, and utilize it to attend to image information when generating audio data. Therefore, we have extended the limits of existing methods in terms of scalability and flexibility. Also, we have demonstrated that EGMI maximizes the interaction between cross-modal knowledge, improving alignment, and sample quality.

1 Introduction

Recently, large language models (LLMs) have demonstrated significant ability in natural language comprehension and generation. Despite recent advances, fully understanding multimodal data—including audio, visual, and language—remains challenging. This is largely due to difficulties in cross-modal alignment and integration, compounded by the scarcity of high-quality datasets. Therefore, the newly emerging multimodal large language models (MLLMs) have primarily focused on text-image interactions, while giving relatively little attention to audio data.

While recent studies have explored generating audio with various generative models, they often face difficulties in effectively capturing and integrating relationships between different modalities. They often depend on extra encoders, adapters, or complex objectives, making the fine-tuning process cumbersome. Furthermore, one of the major challenges for MLLMs is the scarcity of paired datasets, especially in the audio domain, where audio-text pairs are much rarer than image-text pairs due to the significant effort required for annotation.

To address these challenges, we propose a novel MLLM-based audio generative model, dubbed *efficient generative multimodal integration* (EGMI). It leverages image-text paired datasets and pretrained audio models while minimizing the use of learnable layers. The newly designed EGMI mapper enhances cross-modal understanding between images and audio by fusing and aligning

[*]Equal Contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Workshop on AI-Driven Speech, Music, and Sound Generation.

features. This helps address the challenge of limited paired audio-text data. Experimental results show that the proposed method significantly improves both the degree of multimodal alignment and the quality of generated samples.

2 Related Work

2.1 End-to-End MLLMs for Both Comprehension and Generation

Research on MLLMs has made significant progress in the text and image domains, with comparatively less emphasis on audio. Early works such as BLIP-2 [Li et al., 2023], Flamingo [Alayrac et al., 2022], and LLaVA [Liu et al., 2024] advanced MLLM's understanding of image data by leveraging high-quality image-text paired datasets, which are easily accessible. SALMONN [Tang et al., 2023] enhances multimodal comprehension by incorporating Whisper [Radford et al., 2023] and BEATs [Chen et al., 2022] encoders into LLM training. While this design improves audio processing, the LLM training procedure requires significant computational resources and large amounts of data. Similarly, Macaw-LLM [Lyu et al., 2023] utilizes separate encoders for images, audio, and video, which also demands extensive training time and large datasets to construct a joint embedding space for all modalities. In contrast, PandaGPT [Su et al., 2023] and ImageBind-LLM [Han et al., 2023] can process various modalities in an unified manner by utilizing image-text paired data, however, but they have not yet extended their capabilities to high-quality generation. To address the issue, there are many recent studies focusing on incorporating MLLM with audio generative models, such as, TANGO [Majumder et al., 2024] and AudioPaLM [Rubenstein et al., 2023]. While NExT-GPT [Wu et al., 2023] can generate across various modalities, but still relies heavily on modality-specific encoders and paired datasets. Additionally, their interaction between modalities during the generation process is insufficient, which can lead to consistency issues and possibly limit the scalability of MLLM.

2.2 Models Using Multimodal Paired Datasets

In multimodal tasks, well-constructed paired datasets are crucial for learning interactions between different modalities. However, high-quality audio-text paired data are limited, creating significant challenges for audio generation tasks. Models like LLaVA [Liu et al., 2024] and GILL [Koh et al., 2024] utilize text-image paired data, while SALMONN [Tang et al., 2023], Macaw-LLM [Lyu et al., 2023], and AudioPaLM [Rubenstein et al., 2023] rely on text-audio paired datasets. Similarly, NExT-GPT [Wu et al., 2023] requires paired data for each modality it generates, which demands substantial resources. But in general, high-quality audio-text paired data is insufficient due to the complexity of describing time-based audio, and the labor-intensive annotation process [Agostinelli et al., 2023]. To address these challenges, EGMI introduces a novel technique that extends the use of image-text paired data for audio generation. Experimental results demonstrate that this approach effectively mitigates the issue of dataset scarcity while improving the model's efficiency and scalability.

3 EGMI

3.1 Model Architecture

The proposed EGMI efficiently bridges the gap between disparate modalities by leveraging a pre-trained MLLM. Here, we utilized PandaGPT [Su et al., 2023] as our baseline model. To map cross-modal embeddings into a unified representation space, we introduce a learnable projection layer that facilitates seamless interactions between the two models. For image embedding and generation, we employ the visual encoder from ImageBind [Girdhar et al., 2023] and Stable Diffusion [Rombach et al., 2022], respectively, while AudioLDM-L [Liu et al., 2023] is used for audio generation. The core module, the EGMI mapper, adopts a simplified transformer encoder-decoder structure to capture cross-modal interactions between visual and audio token embeddings (i.e., `[IMG1-8]` and `[AUD1-8]`). EGMI mapper grounds these interactions within the LLM and aligns them with the text/image generation space. An overview of the entire framework is shown in 1, with a detailed description of the EGMI mapper provided in Section 3.3

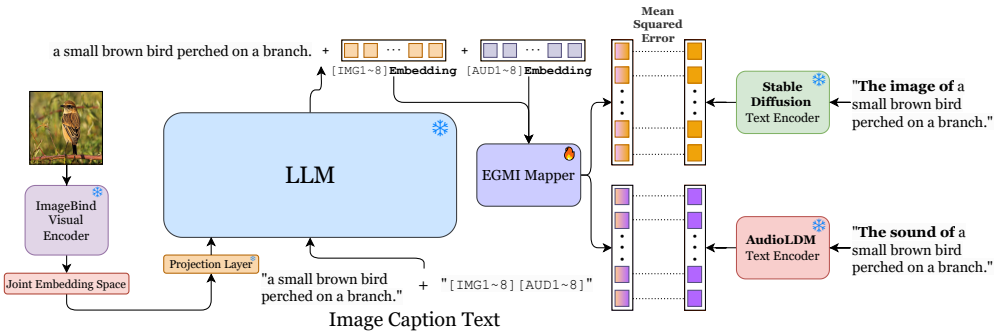

Figure 1: The overall framework of EGMI involves feeding encoded images and text into the LLM, which generates [IMG1-8] and [AUD1-8] tokens. The EGMI mapper grounds these embeddings onto the representations of text encoders within the image and audio generators. EGMI integrates cross-modal data and optimizes alignment for enhanced multimodal performance.

3.2 Training Pipeline

The objective of training EGMI is to generate multimodal outputs using only text-image paired data by effectively learning cross-modal alignment between images and audio, without relying on audio data. The input image is encoded by the ImageBind visual encoder [Girdhar et al., 2023] and subsequently projected for input into the LLM. The special tokens [IMG1-8] and [AUD1-8], which encapsulate visual and potential audio information, are then fused with the text input. The LLM integrates all tokens through the attention mechanism and generates output embeddings. In the generated sequence, the [IMG1-8] and [AUD1-8] tokens appear at the end, following the LLM's autoregressive training mechanism, and serve as embeddings containing both text and image information. Only the embeddings corresponding to these tokens are passed to the EGMI mapper for further processing.The EGMI mapper, built on a transformer encoder-decoder structure, aligns and integrates information by modeling the interaction between audio and image embeddings. This process is primarily used to learn the conditions necessary for audio generation based on image data. To align the EGMI mapper's output with the text encoders of Stable Diffusion [Rombach et al., 2022] and AudioLDM [Liu et al., 2023], we prepend the prefix 'The image of' to the image captions in the training data, allowing pre-extraction of embeddings from the CLIP encoder [Radford et al., 2021]. Similarly, for audio generation, we use the prefix 'The sound of' to extract embeddings from the CLAP encoder [Elizalde et al., 2023]. The entire output embeddings of the EGMI mapper are then aligned to minimize the MSE loss. During the training process, only the special tokens, [IMG1-8] and [AUD1-8], and the EGMI mapper are trained, allowing the model to align with audio generation models.

3.3 EGMI Mapper

As shown in Figure 2, the EGMI mapper is a transformer-based structure designed to map relevant multimodal information by aligning the representations embedded in images and audio, functioning as both an image and audio mapper. We employ learnable queries to capture representative features and generate a fixed-size output from [IMG1-8] and [AUD1-8], similar to the approaches used in GILL [Koh et al., 2024], BLIP-2 [Li et al., 2023], and DETR [Carion et al., 2020]. In this setup, except for the image and audio mappers, the transformer encoder/decoder and learnable query parameters are shared. This configuration effectively facilitates information exchange and alignment between modalities. Here, it is noteworthy that the EGMI mapper learns by processing [IMG] and [AUD] representations in distinct ways. For operations involving CLIP [Radford et al., 2021], the text encoder of Stable Diffusion [Rombach et al., 2022] is utilized. In this case, [IMG] serves as the query in the transformer encoder, while [AUD] functions as the key and value. Conversely, for operations involving CLAP [Elizalde et al., 2023], the AudioLDM [Liu et al., 2023] text encoder is used. Here, [AUD] acts as the query, while [IMG] is treated as the key and value for attention. The EGMI mapper is designed to align multimodal information with the CLIP and CLAP text encoders, aiming to capture interactions across differently configured latent spaces. The loss signals from these encoders

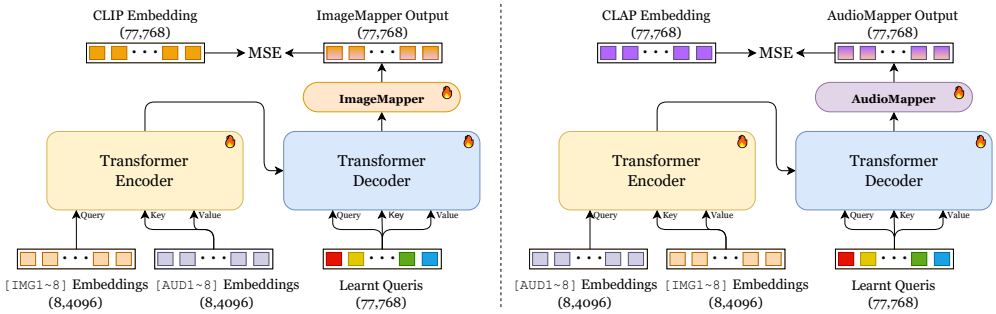

Figure 2: The architecture of the EGMI mapper: the left side aligns with CLIP, while the right side aligns with CLAP. The transformer encoder attends to `[IMG]` and `[AUD]` embeddings distinctly, using them as query, key, and value depending on the target space. Learnable queries are used to extract meaningful information for alignment with the transformer decoder's output in a fixed-size. Image and audio mappers retrieve the integrated visual and audio representations embedded in the transformer decoder's outputs.

are backpropagated to train the transformer and learnable queries, facilitating the effective integration of disparate modalities.

3.4 Generation Process

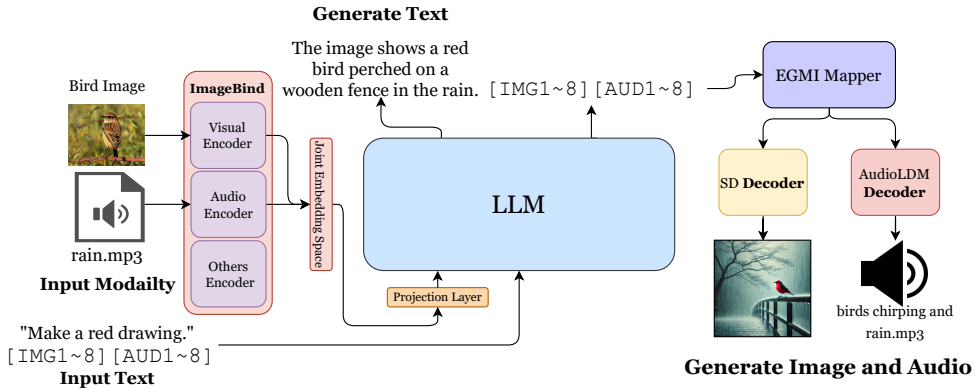

Figure 3: Generation Process. Input images, audio, and text are converted into joint embeddings via ImageBind and then fed into the LLM. The LLM generates `[IMG1-8]` and `[AUD1-8]` tokens containing visual and audio information, along with text outputs. Based on these tokens, the EGMI mapper extracts information as an alternative to the original text encoders of the generative models, passing the conditions to the Stable Diffusion Decoder and AudioLDM Decoder, ultimately generating images and audio.

In Figure 3, the generative process of EGMI, input image, audio, and other modality data are encoded by ImageBind [Girdhar et al., 2023] Encoder to generate joint embeddings. These embeddings are then encoded into modality tokens through a Projection Layer and fed into the LLM. The LLM processes these tokens, which are postfixed with `[IMG]` and `[AUD]`, using its attention mechanism to handle interactions. Based on the input embeddings, the LLM comprehensively understands the multimodal data and generates representations for text, image, and audio outputs. The generated `[IMG]` and `[AUD]` embeddings are then fed into the Transformer Encoder of the EGMI mapper, which extracts integrated information. The extracted embeddings are utilized differently depending on the task: for image generation, `[IMG]` serves as the query, and `[AUD]` as the key and value, while the roles are reversed for audio generation. Each output embedding is processed through the image and audio components of the EGMI mapper, respectively, and then conditioned into Stable Diffusion [Rombach et al., 2022] and AudioLDM [Liu et al., 2023].

4 Experiments

4.1 Text-to-Audio Generation

We compared various models on Text-to-Audio tasks using AudioCaps dataset [Kim et al., 2019], evaluating their performance with FAD (Fréchet Audio Distance) and IS (Inception Score). In Table 1, EGMI achieved FAD of 26.51 and IS of 7.23. Despite not being trained on an audio-paired dataset, EGMI delivered scores comparable to models like AudioLDM-L [Liu et al., 2023]. Although EGMI and NEXT-GPT share the same embedding space, only EGMI is able to process tasks without audio-text paired data, highlighting its significant advantage in data-limited environments.

4.2 Audio-to-Audio Generation

We compared various models on text-based audio editing tasks using VCTK dataset [Yamagishi et al., 2019], evaluating their performance with MCD (Mel Cepstral Distortion) metric. In Table 2, EGMI achieved MCD of 0.361, demonstrating performance comparable to AudioLDM-L [Liu et al., 2023]. Remarkably, this result was achieved by utilizing only the CoCo [Lin et al., 2014] image captioning dataset, without any English speech data. This highlights the robust text editing capabilities of the well-trained MLLM, which can handle more complex alignment without additional LLM training. Unlike other models relying on large paired datasets, EGMI proves that effective cross-modal alignment can be achieved with limited data.

Table 1: Text-to-Audio Performances

Models	**FAD(↓)**	**IS(↑)**
DiffSound [Yang et al.]	47.68	4.01
AudioLDM-S [Liu et al., 2023]	49.48	6.90
AudioLDM-L [Liu et al., 2023]	23.31	8.13
CoDI [Tang et al., 2024]	**22.90**	**8.77**
NExT-GPT [Wu et al., 2023]	23.58	8.35
EGMI(ours)	26.51	7.23

Table 2: Text+Audio-to-Audio Performances

Models	**MCD(↓)**
CampNet [Wang et al., 2022]	0.380
MakeAudio [Huang et al., 2023]	0.375
AudioLDM-L [Liu et al., 2023]	0.349
NExT-GPT [Wu et al., 2023]	**0.300**
EGMI(ours)	0.361

4.3 Ablations on EGMI mapper

In Table 3, by evaluating various mapper designs that incorporate both image and audio features, we identified the most efficient design for multimodal alignment. The Separated mapper uses distinct Transformers for image and audio, while the Cross mapper adds a cross-attention module to the last layer of the separate Transformers to facilitate information sharing. Experimental results revealed that the EGMI mapper outperforms the others in terms of cross-modal alignment.

Table 3: Comparison of mapper designs: Similarity was measured using CLIP for text-image and CLAP for text-audio evaluations. Text-image similarity was assessed on the CoCo dataset [Lin et al., 2014], while text-audio similarity was evaluated on the AudioCaps dataset [Kim et al., 2019].

Models	**CLIP (text-image) similarity**	**CLAP (text-audio) similarity**
Separated mapper	0.611	0.385
Cross mapper	0.582	0.312
EGMI mapper	**0.688**	**0.401**

5 Conclusion

EGMI presents a novel perspective on MLLMs through its innovative approach to audio generation. By aligning cross-modal data using only a text-image paired dataset, EGMI has extended multimodal understanding and generation to the audio domain, pushing the limits of efficiency and flexibility. The EGMI mapper module further demonstrates the model's capability to produce consistent and high-quality multimodal outputs through effective feature fusion. Looking ahead, EGMI shows promise for further advancements, particularly in enhancing scalability for large datasets and improving metrics for evaluating the perceptual quality and diversity of generated outputs.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yanli2023blipa Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *arXiv preprint arXiv:2404.09956*, 2024.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

Tao Wang, Jiangyan Yi, Ruibo Fu, Jianhua Tao, and Zhengqi Wen. Campnet: Context-aware mask prediction for end-to-end text-based speech editing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2241–2254, 2022.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pages 271–350, 2019.

D Yang, J Yu, H Wang, W Wang, C Weng, Y Zou, and D Diffsound Yu. Discrete diffusion model for text-to-sound generation. arxiv 2022. *arXiv preprint arXiv:2207.09983*, 1.