Bridging Facial Imagery and Vocal Reality: Stable Diffusion-Enhanced Voice Generation

Yueqian Lin[1], *Dong Liu*[1], *Yunfei Xu*[2], *Hongbin Suo*[2], *Ming Li*[1]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China
[2]Guangdong OPPO Mobile Telecommunications Corp., Ltd.

ming.li369@dukekunshan.edu.cn

Abstract

Generating novel voices in speech synthesis is a challenging task with potential for creating versatile voices that are needed in entertainment and research. One of the primary obstacles in this area is the lack of well-annotated voice descriptions for expressive speech corpora. Our research aims to address this issue by representing speaker styles from vision. We introduce Stable Diffusion-Enhanced Voice Generation (SD-EVG), which leverages Stable Diffusion to generate *imaginary* facial images for new voice generation. To create a reference set of facial images based on *realistic* voices, SD-EVG employs a transformer encoder and a Stable Diffusion decoder to visualize the speaker's face. Subsequently, SD-EVG uses a KNN-based approach to map facial features to speech style for voice generation. The experiments demonstrate that the voices generated from the imagined facial data have better potential at capturing speech style than text-based methods for the same descriptions.[1]

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Text-to-speech (TTS) synthesis has witnessed significant advancements in recent years, fueled by the development of emerging applications such as audiobooks and virtual assistants [1]. This evolution has spanned from modeling the speech of individual speakers [2] to accommodating multi-speaker scenarios [3], with the naturalness and robustness of the generated speech utterances being markedly enhanced by the advent of larger model and dataset size [4, 5]. A notable milestone in this progression is the modeling of speaker timbre, which has evolved from utilizing singular speaker embeddings [6] to the capability of customizing timbre with just a reference speech utterance with transcription [7].

Despite advancements, natural language descriptions' application in specifying voice characteristics (e.g., pitch, gender, race, emotion) faces a critical limitation due to the scarcity of extensively annotated speech data [8]. This limitation hinders the potential for generating novel, high-quality voice timbres from text prompts. Given the rich and well-collected image-text and audio-visual datasets [9, 10, 11, 12], exploring alternative modalities such as vision for voice generation becomes compelling. Thus, we propose an image-based approach to voice generation that capitalizes on the naturally occurring synergy between facial images and voicesa rich yet underutilized source for enhancing speaker style representation. Unlike the text prompt method, which struggles to encapsulate the full spectrum of desired voice characteristics, leveraging high-quality images promises a more detailed representation of speaker styles. Our approach advocates the utilization of 512×512 images, generated via image generative models such as Stable Diffusion [13], to accurately capture and represent the speech style, thus facilitating a more effective bridge among textual description, visual representation, and vocal style in TTS systems.

The overview of our proposed voice generation pipeline enhanced by Stable Diffusion (SD-EVG) is shown in Figure 1. It uses Stable Diffusion as a bridge for three pipelines: voice-to-face, prompt-to-face, and face-to-voice. The introduction of an image generation model is twofold, as 1) some expressive speech datasets do not include facial images for the speakers in the recordings, so we need to use a voice-to-face model for training data preparation, and 2) to realize the new voice generation, a prompt-to-face model is needed. And findings show that Stable Diffusion, though initially designed for text-to-image, is an ideal base model for realizing face generation from different modalities [14]. Furthermore, SD-EVG can receive facial input from sources beyond Stable Diffusion. This allows for potential extensions to be implemented, including faces in the wild. The main contributions of this paper are as follows:

- We propose a voice-to-face pipeline to augment existing *realistic* speech corpora without the speaker's facial information with an *imaginary* face in the training stage.
- We develop a face-to-speech pipeline that transforms facial features into speech style embedding using k-nearest neighbors to bridge multiple modalities' feature latent spaces.
- We evaluate the performance of SD-EVG within both the voice-to-face and face-to-voice pipelines. We also conduct a comparative analysis between our proposed methodology, which involves a prompt-to-face approach followed by face-based voice generation, and the direct prompt-to-voice generation method through a case study.

2. Related Work

2.1. Stable Diffusion

Stable Diffusion is a set of latent diffusion models [13] that can generate images given an image or text input due to its ability to model conditional distributions in the form $p(\boldsymbol{z}|\boldsymbol{e})$, representing a significant advancement in generative models. We use Equation (1) to abstract the model input and output:

$$\boldsymbol{x} = G(\boldsymbol{z}; \boldsymbol{e}), \tag{1}$$

where $\boldsymbol{x}$ is the synthesized image, $\boldsymbol{z}$ is a latent representation vector, $\boldsymbol{e}$ is a guided embedding vector representing the conditional information for image generation, and G is the Stable Diffusion model. Specifically, in the text-to-image pipeline

[1] A demo website featuring the generated face and speech utterances is available at https://sd-evg.github.io.

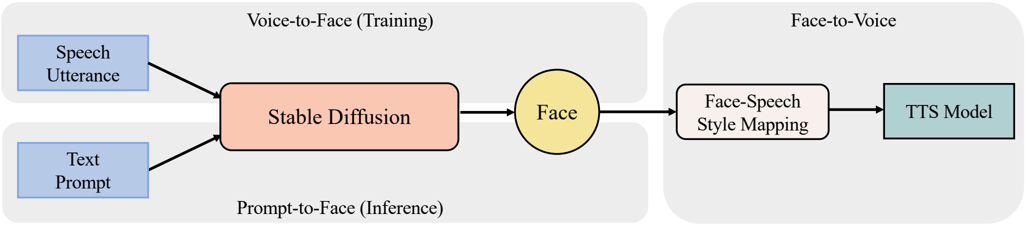

Figure 1: *Overview of Stable Diffusion-Enhanced Voice Generation.*

of Stable Diffusion, $\boldsymbol{z}$ is a random latent variable and $\boldsymbol{e}$ is the output of a transformer-based text encoder taking an input text prompt. In terms of the text-guided image-to-image translation pipeline [15], the initial latent $\boldsymbol{z}$ is the reference image added with noise, while $\boldsymbol{e}$ is the same as in text-to-image pipeline. Additionally, methodologies similar to those developed by OpenAI [16] for transforming image embeddings back into images allow Stable Diffusion to extend its capabilities to include processing of image embeddings, extracted by foundation models, through further fine-tuning.

2.2. Speech and Vision Alignment

In the field of multimodal learning, the integration of visual and linguistic data through transformers, as exemplified by the Contrastive Language-Image Pre-training (CLIP) model [17], represents a significant advance in improving model performance across a wide range of tasks. This methodological innovation trains visual models for transfer learning using natural language supervision derived from a large corpus of image-text pairs. CLIP distinguishes itself by embedding images and text in a shared semantic space and using contrastive learning to align semantically similar pairs while discriminating dissimilar ones. This strategy facilitates CLIP's ability to apply its pre-trained knowledge to various computer vision tasks with minimal additional supervision.

Although contrastive learning has been applied to align speech and vision, this field is still emerging. SpeechCLIP, as introduced by Shih et al. [18], pioneers the integration of speech models with vision through the use of paired image-speech data, enhancing speech models with visual context. Building upon SpeechCLIP's foundation, subsequent research has broadened the scope to include multilingual speech-to-image retrieval [19], image-to-speech captioning [20], and the enhancement of speech systems with visual data [21]. Despite these advances, there appears to be a gap in the literature regarding applying contrastive learning to exploit facial identity and styling features within CLIP's image encoder for speech style extraction. This gap underscores a compelling research opportunity further to enhance the integration of speech and vision modalities.

3. Methodology

SD-EVG comprises three primary pipelines: voice-to-face, face-to-voice, and prompt-to-voice. The voice-to-face pipeline is used during the training phase to compensate for the absence of facial images in speech datasets. The face-to-voice pipeline is used both in the training and the inference stage, facilitating the conversion from facial images to corresponding speech styles. The prompt-to-voice pipeline is used naturally in the inference stage, where it leverages Stable Diffusion's text-to-image capability to generate a face image from a text prompt, which is then used to guide the synthesis of speech. The voice-to-face and face-to-voice pipelines are discussed in detail below.

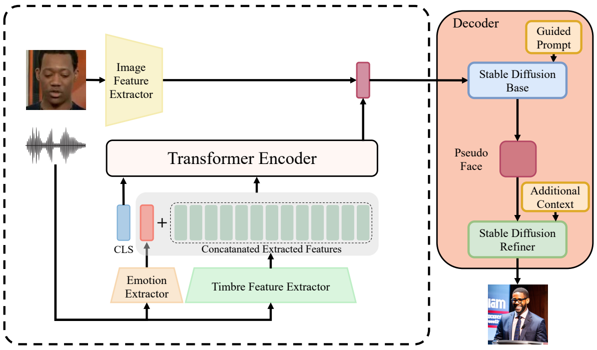

Figure 2: *Architecture of the Voice-to-Face Pipeline.*

3.1. Voice-to-Face Pipeline

In our approach to leveraging the capabilities of Stable Diffusion model (G) for the generation of imaginary faces (F) from a realistic speech corpus (C), we focus on conditioning the generative process on the timbre features of specific utterances (u in C). To achieve this, we utilize WavLM [22] and emotion2vec [23] for extracting the emotional and timbre features from the utterances. These extracted features are then concatenated and input into a transformer encoder layer, which outputs a vector representation V_{style} of the utterance. This speech-to-vector transformation is represented by the equation:

$$V_{\text{style}} = \text{Transformer}(E_{\text{emotion}}(u) + E_{\text{timbre}}(u)). \tag{2}$$

During training, as illustrated within the dashed lines of Figure 2, both visual and auditory features are extracted using their respective extractors, with the CLIP image encoder being used for image feature extraction. The goal is to compute V_{style} from both the image feature extractor and the speech-to-vector process, which then serves as a crucial component for calculating the contrastive loss during training. This process aligns multimodal data by contrasting pairs of related and unrelated data. During inference, V_{style} directly guides the Stable Diffusion in generating a facial image. This image, imagined from the voice input, is regarded as a direct embodiment of the speaker's speech style.

As described in Section 2.1, Stable Diffusion can simultaneously take in both image and text guidance input. To enhance its capability to capture the speaker's style with guided feature descriptions and higher resolution, the pipeline's decoder module is designed to initially accept the inferred V_{style} from the

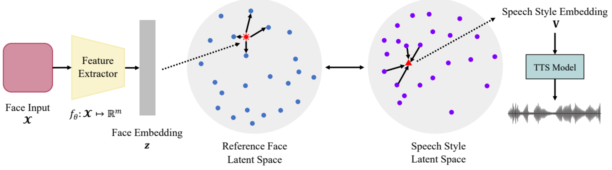

Figure 3: *Face-Speech Mapping Based on k-nearest neighbors.*

transformer encoder, generating a pseudo facial image with a guided prompt ("face" in our case) using a base-level Stable Diffusion model. This pseudo facial image can then undergo further refinement through another iteration of Stable Diffusion, guided by additional contextual inputs for enhanced resolution and detailed feature expression.

3.2. Face-to-Voice Pipeline

Once a facial image representing the speaker's speech style has been generated (or is already available), a trained zero-shot Text-to-Speech (TTS) system can be utilized. This system is capable of generating speech utterances from a reference speech style embedding V_{speech}. Therefore, the primary objective of the face-to-voice pipeline is to translate the abundant style information captured in the facial image into a corresponding speech style embedding V_{speech}.

The key idea of our proposed face-to-voice pipeline (Figure 3) begins with extracting face embeddings using a facial feature extractor, thereby establishing a face latent space for reference. Simultaneously, we employ a speech-style embedding extractor to infer paired speech styles from the face, constructing a speech-style latent space. We hypothesize that faces positioned closely within the face latent space are more likely to exhibit similar vocal characteristics. For inference, k-nearest neighbors (KNN) regression is used to achieve this mapping, employing a strategy similar to that described in [24]. It is noteworthy that this process is non-parametric, necessitating no further training.

Specifically, a facial feature extractor $f_\theta : \mathcal{X} \to \mathbb{R}^m$ maps an input facial image $\boldsymbol{x} \in \mathcal{X}$ to a face embedding $\boldsymbol{z} = f_\theta(\boldsymbol{x})$. Given a set of face embeddings $\mathcal{Z}$ and corresponding speech style embeddings $\boldsymbol{v}$, KNN regression is employed to generate the speech style embedding $\boldsymbol{v}$ corresponding to $\boldsymbol{z}$. The KNN method entails identifying the k nearest neighbors $\mathcal{N}_k(\boldsymbol{z}) \subseteq \mathcal{Z}$ using a distance metric d, acquiring their speech style embeddings $\mathcal{V}_k(\boldsymbol{z})$, and calculating a weighted sum:

$$\boldsymbol{V}(\boldsymbol{z}) = \sum_{i=1}^{k} w_i \boldsymbol{v}_i \tag{3}$$

where $w_i = \frac{1/d(\boldsymbol{z},\boldsymbol{z}_i)}{\sum_{j=1}^{k} 1/d(\boldsymbol{z},\boldsymbol{z}_j)}$ to ensure contributions are proportional to the proximity of each neighbor.

4. Experiments

4.1. Experimental Setup

For the voice-to-face pipeline, the base version of emotion2vec is utilized [23] and a speaker-verification-fine-tuned version of WavLM is used [22], acting as the emotion and timbre feature extractors for speech utterances. During training, the image feature extractor employed uses CLIP's ViT-L/14 Transformer architecture[2]. For face generation, the unclip version of Stable Diffusion 2.1[3] serves as the base model, and the SDXL-turbo[4] functions as the refiner model for enhanced detail. In the face-to-voice pipeline, a pretrained Inception ResNet [25] architecture-based face recognition system[5] is adopted for facial feature extraction, while the style encoder module from StyleTTS [26] is used to extract speech styles from utterances, with its zero-shot TTS module facilitating voice synthesis.

We mainly use two datasets in the experiments, for voice-to-face learning we use an audio-visual dataset VoxCeleb1 [11], which contains 153516 utterances from 1,251 celebrities, and for face-to-voice pipeline we use the VoxCeleb1, and the ESD dataset [27], which contains 5 different emotion categories, each category containing 350 utterances, for a total of 17,500 utterances evenly distributed across 10 speakers, to test the ability of SD-EVG in emotional speech generation. For all experiments, the number of k-nearest neighbors is fixed at 4.

To assess our system's efficacy, separate evaluations of the voice-to-face and face-to-voice pipelines are conducted, followed by a case study on emotional speech generation using only prompt information, compared against the direct prompt-to-voice method. Specifically, the evaluation of SD-EVG covers three distinct aspects, involving 30 participants recruited for subjective assessments.

1. **Voice-to-face fitness assessment.** Evaluation involves providing 20 diverse speech utterances from speakers in VoxBlink [12], a recently introduced large-scale audio-visual dataset from individuals on YouTube, and asking participants to rank according to faces most likely to be the speaker, from four other randomly generated faces. This aims to test SD-EVG's voice-to-face effectiveness for public data.

2. **Face-to-voice quality assessment.** For the face-to-voice pipeline comparison, the naturalness is evaluated against other models that generate speech from face images. Twelve unseen speakers' faces are prepared, mirroring the setting in [28]. Each speaker's face generates a speech utterance, and evaluators conduct a Mean Opinion Score (MOS) test on a 5-point scale (ranging from 1 (unnatural, unable to hear clearly) to 5 (natural, human-like speech)), and we also calculate the Word Error Rate (WER) using Whisper [29]. This aims to test SD-EVG's face-to-voice quality in zero-shot scenarios.

3. **Case study in emotional speech generation.** For a comprehensive evaluation of SD-EVG, the ESD dataset serves as the sole reference dataset for the face-to-voice pipeline in this test. Note that the ESD dataset only contains speech data so we use the voice-to-face pipeline first to generate the *imaginary* faces while the additional context in Figure 2 is the emotion label. For comparison with the prompt-to-voice method, the original face feature extractor is replaced by the base BERT model [30] which, having 768 feature dimensions, exceeds the 512 dimensions of the face feature extractor, allowing for prompt-based speech style generation. We use a preference test for the comparison of the two systems, i.e., we generate prompts describing the speaker's emotion in the ESD dataset and 20 test prompts by GPT 4 [31] and evaluators are asked to select the speech utterances generated by the test prompts that best match the given emotion.

[2] https://huggingface.co/openai/clip-vit-large-patch14

[3] https://huggingface.co/stabilityai/stable-diffusion-2-1-unclip

[4] https://huggingface.co/stabilityai/sdxl-turbo

[5] https://github.com/timesler/facenet-pytorch

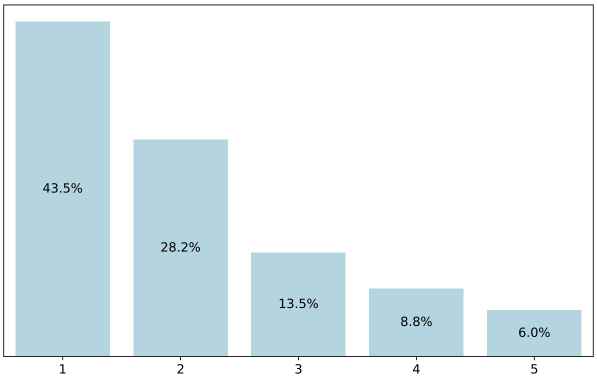

Figure 4: *Frequency of preference rankings for SD-EVG. The x-axis represents SD-EVG's rank in the preference test.*

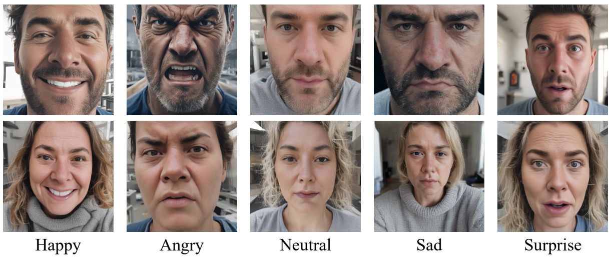

Figure 5: *Generated faces from ESD dataset using the proposed SD-EVG. Each row represents a different speaker.*

4.2. Results and Discussion

Figure 4 illustrates SD-EVG's ranking in the subjective evaluation of voice-to-face fitness. In testing, more than 70% of the generated faces were ranked first or second out of five, demonstrating the alignment of the faces with the corresponding voices. Table 1 presents the MOS scores and WER for the proposed SD-EVG system alongside three recent face-to-voice models, illustrating comparative face-to-voice quality. SD-EVG exhibits a slightly lower MOS score than the highest baseline but surpasses others, and it demonstrates a lower WER compared to all baselines, indicating its good performance in speech naturalness and intelligibility. Also, it should be noted that due to the exploratory nature of this work, a KNN regressor is employed for face-to-voice style mapping without further training, in contrast to the baselines, which utilize well-trained face encoders.

Model	MOS (↑)	WER (↓)
Face-TTS [32]	2.62 (± 0.07)	0.125
Grad-StyleSpeech [33]	3.36 (± 0.06)	0.086
Face-StyleSpeech [28]	**3.57 (± 0.06)**	0.092
SD-EVG (proposed)	3.45 (± 0.06)	**0.046**

Table 1: *MOS and WER of different face-to-voice models. 95% confidence intervals of MOS are presented in parentheses.*

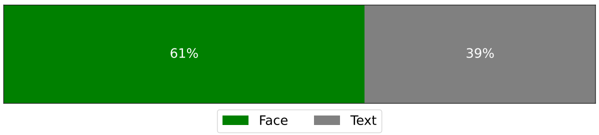

Figure 6: *Face-text preference test.*

As for the emotional speech generation, Figure 5 displays the generated faces of two speakers by SD-EVG from the ESD dataset, showcasing the same speaker under different speaking emotions. It can be observed that while not explicitly indicating the speaker's gender, SD-EVG is able to generate facial images that correspond to the speaker's gender in most cases. The average cosine similarity of the extracted face embeddings within each speaker across the ESD dataset is calculated to be 43%. This suggests that SD-EVG can capture and preserve the speaker's identity to a certain extent (given the imaginative nature of Stable Diffusion, an accurate and consistent voice-to-face mapping cannot be guaranteed).

Figure 6 shows the preference rate of SD-EVG compared to the text-based voice generation method in terms of emotional voice generation. Although the text-based feature extractor has more dimensions than the face-based one, the face-based method outperforms the text-based method in voice generation. We also note that stronger emotions such as sadness or surprise perform much better than the text-based method. This suggests that using facial cues in new voice generation enhances expressiveness by tapping into the unique style and emotional information conveyed by faces, as well as leveraging the detailed and rich facial data within Stable Diffusion.

4.3. Further Discussion

The results in Section 4.2 suggest SD-EVG's competitive performance in voice-to-face alignment and face-to-voice naturalness. Moreover, SD-EVG's ability to capture and convey emotional features through facial cues highlights the potential of image-based voice generation methodologies in enhancing expressiveness within TTS systems. The substantial model size of Stable Diffusion and the utilization of the KNN approach for face-to-voice style mapping enable the generation of a diverse array of voices, tailored to specific facial features. Additionally, further research on Stable Diffusion's adaptability reveals opportunities for other applications of SD-EVG, such as achieving identity-preserving speech style customization with Stable Diffusion's image-to-image translation capability.

5. Conclusion

Our proposed SD-EVG framework leverages the power of Stable Diffusion for the generation of new voices through voice-to-face, face-to-voice, and prompt-to-voice pipelines. Our approach uniquely integrates visual cues in the form of generated facial images to enhance the stylistic fidelity of voice generation, addressing the challenge of capturing nuanced timbre descriptions textually. Experimental evaluations demonstrate the effectiveness of SD-EVG in producing voices that accurately reflect desired emotional states and speech styles, and outperform our baselines in terms of naturalness and intelligibility. By pioneering the use of visual information from vision foundation models in voice synthesis, SD-EVG marks an exploration toward more authentic and versatile voice generation.

6. Acknowledgement

This research is funded in part by the National Natural Science Foundation of China (62171207), Guangdong Science and Technology Plan (2023A1111120012) and OPPO.

7. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, T. Qin, and T.-Y. Liu, "Multispeech: Multi-speaker text to speech with transformer," *arXiv preprint arXiv:2006.04664*, 2020.

[4] T. A. Nguyen, W.-N. Hsu, A. d'Avirro, B. Shi, I. Gat, M. Fazel-Zarani, T. Remez, J. Copet, G. Synnaeve, M. Hassid *et al.*, "Expresso: A benchmark and analysis of discrete expressive speech resynthesis," *arXiv preprint arXiv:2308.05725*, 2023.

[5] J. Shi, Y. Lin, X. Bai, K. Zhang, Y. Wu, Y. Tang, Y. Yu, Q. Jin, and S. Watanabe, "Singing voice data scaling-up: An introduction to ace-opencpop and kising-v2," *arXiv preprint arXiv:2401.17619*, 2024.

[6] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Proc. NeurIPS*, vol. 32, 2019.

[7] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[8] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "Prompttts: Controllable text-to-speech with text descriptions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. 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.* Springer, 2014, pp. 740–755.

[10] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[11] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.

[12] Y. Lin, X. Qin, M. Cheng, N. Jiang, G. Zhao, and M. Li, "Voxblink: X-large speaker verification dataset on camera," in *Proc. ICASSP.* IEEE, 2024.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR*, 2022, pp. 10 684–10 695.

[14] X. Xu, Z. Wang, G. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 7754–7765.

[15] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=aBsCjcPu_tE

[16] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML.* PMLR, 2021, pp. 8748–8763.

[18] Y.-J. Shih, H.-F. Wang, H.-J. Chang, L. Berry, H.-y. Lee, and D. Harwath, "Speechclip: Integrating speech with pre-trained vision and language model," in *2022 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2023, pp. 715–722.

[19] L. Berry, Y.-J. Shih, H.-F. Wang, H.-J. Chang, H.-y. Lee, and D. Harwath, "M-speechclip: Leveraging large-scale, pre-trained models for multilingual speech to image retrieval," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[20] M. Kim, J. Choi, S. Maiti, J. H. Yeo, S. Watanabe, and Y. M. Ro, "Towards practical and efficient image-to-speech captioning with vision-language pre-training and multi-modal tokens," *arXiv preprint arXiv:2309.08531*, 2023.

[21] S. Bhati, J. Villalba, L. Moro-Velazquez, T. Thebaud, and N. Dehak, "Leveraging pretrained image-text models for improving audio-visual learning," *arXiv preprint arXiv:2309.04628*, 2023.

[22] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[23] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[24] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Interspeech*, 2023.

[25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, vol. 31, no. 1, 2017.

[26] Y. A. Li, C. Han, and N. Mesgarani, "Styletts: A style-based generative model for natural and diverse text-to-speech synthesis," *arXiv preprint arXiv:2205.15439*, 2022.

[27] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 920–924.

[28] M. Kang, W. Han, and E. Yang, "Face-stylespeech: Improved face-to-voice latent mapping for natural zero-shot speech synthesis from a face image," *arXiv preprint arXiv:2311.05844*, 2023.

[29] 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.* PMLR, 2023, pp. 28 492–28 518.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019.

[31] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[32] J. Lee, J. S. Chung, and S.-W. Chung, "Imaginary voice: Face-styled diffusion model for text-to-speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[33] M. Kang, D. Min, and S. J. Hwang, "Grad-stylespeech: Any-speaker adaptive text-to-speech synthesis with diffusion models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.