



# VS-Singer: Vision-Guided Stereo Singing Voice Synthesis with Consistency Schrödinger Bridge

Zijing Zhao<sup>1</sup>, Kai Wang<sup>1,2</sup>, Hao Huang<sup>1,2,3,\*</sup>, Ying Hu<sup>1,2</sup>, Liang He<sup>1,4</sup>, Jichen Yang<sup>5</sup>

<sup>1</sup>School of Computer Science and Technology, Xinjiang University, China

<sup>2</sup>Xinjiang Key Laboratory of Multi-lingual Information Technology, China

<sup>3</sup>Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, China

<sup>4</sup>Department of Electronic Engineering, Tsinghua University, China

<sup>5</sup>School of Cyber Security, Guangdong Polytechnic Normal University, China

mirror-zhao@stu.xju.edu.cn, huanghao@xju.edu.cn

## Abstract

To explore the potential advantages of utilizing spatial cues from images for generating stereo singing voices with room reverberation, we introduce VS-Singer, a vision-guided model designed to produce stereo singing voices with room reverberation from scene images. VS-Singer comprises three modules: firstly, a modal interaction network integrates spatial features into text encoding to create a linguistic representation enriched with spatial information. Secondly, the decoder employs a consistency Schrödinger bridge to facilitate one-step sample generation. Moreover, we utilize the SFE module to improve the consistency of audio-visual matching. To our knowledge, this study is the first to combine stereo singing voice synthesis with visual acoustic matching within a unified framework. Experimental results demonstrate that VS-Singer can effectively generate stereo singing voices that align with the scene perspective in a single step.

**Index Terms:** stereo singing voice synthesis, multimodal, consistency schrödinger bridge

## 1. Introduction

Singing voice synthesis (SVS) involves generating singing voice from musical scores and lyrics, which include features such as notes and pitch [1–5]. As the continuous development of diffusion models [6–12], the naturalness and fluency of synthesized singing speech have now approached those of human performances. Grad-TTS [13] introduces stochastic differential equation (SDE) and uses a numerical ordinary differential equation (ODE) solver [14] to solve the reverse SDE, but this reverse process requires thousands of iterations to produce high-quality speech. DiffSinger [1] employs a shallow diffusion mechanism, using an auxiliary acoustic model to generate mel-spectrograms, which reducing the sampling process to just one hundred steps. While these models can generate high-quality audio, the excessive number of sampling iterations results in slow inference speed. To alleviate this issue, CoMo-Speech [15] first pretrains an acoustic model as teacher model, then performs consistency distillation [16] using the pretrained model [17], allowing speech samples to be generated in just one sampling step. However, this two-stage training process significantly increases training costs, and if the consistency model [16] is trained independently without the pre-trained teacher model, its performance declines drastically.

The primary focus of current research in SVS is to improve the naturalness and fluency of synthesized singing voice. However, with the advancement of AR/VR, users have raised their

\*Corresponding author: Hao Huang

This work was supported by National Natural Science Foundation of China (62466055).

expectations, seeking a more immersive auditory experience. To achieve this, mainstream research typically converts mono audio into stereo audio through spatial processing. Some researchers [18–20] proposed matching visual and audio features using attention mechanisms to convert audio into sounds that appear to be recorded in the target environment. Building on this, NVAS [21] designed a network to analyze audio-visual features and convert audio observed from a given perspective into audio for the target perspective. Liu et al. [22] introduced the use of masking layers to create left and right eye views that match the observer’s perspective, enhancing the audio in the corresponding channels. M2SE-VTTS [23] proposes to simultaneously use the RGB and Depth spaces of spatial images to model local and global spatial knowledge. However, there has not yet been an effort to research stereo singing voice synthesis.

To bridge this gap and investigate the potential benefits of leveraging spatial cues from images to generate stereo singing voices with room reverberation, this work proposes a novel method named VS-Singer, which is a unified framework designed to have the capability in both visual acoustic matching and stereo singing voice synthesizing. Specifically, VS-Singer comprises a modal interaction network (MIN), a decoder based on the consistency Schrödinger bridge (CSB) and a spatially-aware feature enhancement module (SFE). The role of the modal interaction network is to analyze audio-visual consistency through the interaction of different modal information and extract spatial features to guide stereo synthesis. Moreover, the CSB is also creatively proposed to generate high-quality stereo singing in one step without the need for a teacher model, which increases the synthesis speed to 2 times that of the baseline model. Additionally, the SFE is employed to enhance the consistency of audio-visual matching. Our approach is validated through extensive experiments on the open-source Openpop [24] and NVAS-SoundSpace [21] corpora. Results demonstrate VS-Singer’s superior capability in inference speed and immersive stereo singing voice synthesis. Audio samples are available at: <https://usinger1.github.io/VS-Singer/>.

The key contributions of our work are summarized as follows:

- To the best of our knowledge, VS-Singer is the first to integrate visual-acoustic matching and singing voice synthesis into a unified framework for synthesizing stereo singing voice with room reverberation.
- We propose a method that fuses Schrödinger bridge with consistency training [16] that reduces the cost of training while boosting performance of the model and the speed of inference.
- Extensive experiments conducted on open-source corpora demonstrate that VS-Singer can efficiently generate stereo singing voices that accurately matches the scene

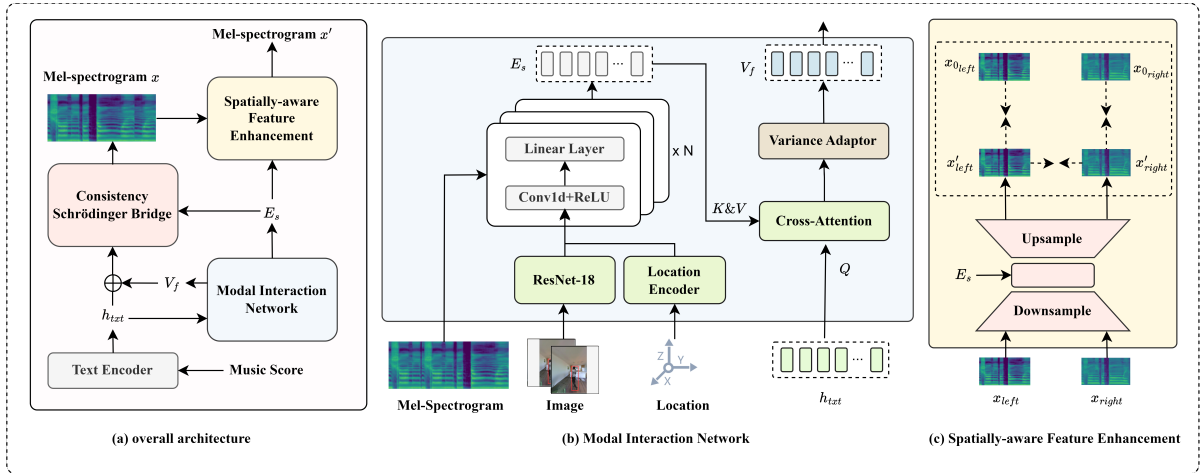


Figure 1: Overview of our proposed model structure.

perspective in one step, particularly achieving a 2x inference speed-up over the comparative cascaded systems.

## 2. Methodology

The overall architecture of the proposed VS-Singer model is illustrated in Fig. 1. Our model consists of three parts: 1) The MIN is used to enable the interactive aggregation of multimodal information sources, models environmental acoustics, generates spatial cues, interacts with text encoding, and generates text hidden sequences with spatial features to guide stereo singing generation. 2) The decoder based on CSB is used to establish optimal transmission and generates stereo samples in a single step. 3) The SFE module plays the role of enhancing the audio difference between left and right channels and accelerates model convergence.

### 2.1. Modal Interaction Network

For visual acoustic matching, it is important to consider the different contributions of different regions of the space to the acoustics. Therefore this paper proposes mimicking the observer’s natural focus on the scene to infer the effect on acoustics. Firstly, mask off the 1/4 region of the left and right sides of the scene image to simulate the left and right eye viewpoints, respectively, and extract the corresponding viewpoint spatial feature  $V_{env} = (V_{left}, V_{right})$  using the pre-trained ResNet-18 [25], where  $(\cdot, \cdot)$  denotes the concatenate operation. Then add the absolute position coding to the visual coding. Because a single 2D image is insufficient for the model to understand the propagation paths of sound waves in space. Therefore, 3D spatial positioning information is introduced to capture the relative position between the target speaker and source view. The target speaker’s position is recorded as translations along the  $x$ ,  $y$ , and  $z$  axes. The format is represented as :

$$V_{loc} = (d, \sin(\alpha), \cos(\alpha)), \quad (1)$$

, where  $d$  represents the Euclidean distance between the target speaker and the source viewpoint, and  $\alpha$  is the rotation angle between the target speaker and the source view on the  $XY$  plane.

The energy of each short-time Fourier transform frame is calculated by calculating the L2-norm of the amplitude of each frame of the binaural channel and quantizing it on a logarithmic

scale, which is encoded to generate an energy vector  $V_e$ . The  $V_{env}$ ,  $V_{loc}$  and  $V_e$  are fed into the modal interaction network to obtain spatial information embedding  $E_s$  aligned with the text. The network includes a convolution stack and a cross-modal attention function. The convolution stack includes convolution layers and pooling layers, which can fuse energy embeddings with spatial information at the frame level and learn the energy changes of stereo at different viewpoints. The cross-modal attention function enables the model to pay attention to different image region features and spatial information, and infer how they affect reverberation and stereo.

In order to learn the interaction between text encoding and spatial features, an attention mechanism is introduced to compute the dot product of text-hidden sequences and spatial information embedding  $E_s$ :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2)$$

where  $Q$  is the text-hidden sequences  $h_{txt}$ , and  $K$  and  $V$  is the spatial information embedding. Finally the attention score obtained by the variance adapter [17] is projected to the text-hidden sequences to obtain spatially relevant text-hidden sequences  $V_f$ .

### 2.2. Consistency Schrödinger Bridge

To ensure the speed and quality of singing voice, we introduce the CSB in this subsection. The consistency model supports one-step sampling and can trade computation time for improved quality. We adopt consistency training as in [16], which supports fast sampling while reducing the cost of training a teacher model. To compensate for the performance loss caused by independent training, we integrate the Schrödinger Bridge (SB) [26–28] into score-based generative models (SGM) framework, establishing a tractable process between between the linguistic representation  $x_1$  (e.g., the hidden sequence) and the clean audio of the left and right channels  $x_0 = (x_{0left}, x_{0right})$ , enhancing the accuracy of marginal density distribution calculations in consistency training, and thereby improving the quality of the samples. According to [27], Schrödinger bridge can be established compatible with the SGM framework. By using the diffusion model setup, this tractable diffusion bridge can be established between a single data point  $x_0$  and the linguistic rep-

resentation  $p_B(x_1|x_0)$ :

$$q(x_t|x_0, x_1) = \mathcal{N}(x_t; \mu_t(x_0, x_1), \Sigma(t)^2), \quad (3)$$

$$\mu_t = \frac{\bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} x_0 + \frac{\sigma_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} x_1, \quad \Sigma(t)^2 = \frac{\bar{\sigma}_t^2 \sigma_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} \cdot I,$$

where  $\sigma_t^2 := \int_0^t \beta_\tau d\tau$  and  $\bar{\sigma}_t^2 := \int_t^1 \beta_\tau d\tau$  are variances accumulated from either sides and  $I$  represents the identity matrix.

When  $x_0$  derived by SGM from the sampled  $x_t$  of  $p_B$  approaches the clean audio  $x_0$  asymptotically, the marginal density of the Schrödinger bridge path will coincide with the marginal density induced by diffusion model [29]. [14] proposed that any SDE corresponds to a probabilistic flow ordinary differential equation (PF-ODE), whose sampling trajectory at  $t$  has the same marginal probability distribution  $p_t(x_t)$ . This means that, given score function  $\nabla \log p_t(x_t)$ , the state at any time on the ODE trajectory can be obtained through an ODE solver. The unbiased estimator of  $\nabla \log p_t(x_t)$  is designed as:

$$\nabla \log p_t(x_t) = -\mathbb{E} \left[ \frac{x_1 - x_t}{\Sigma(t)^2} |x_t \right], \quad (4)$$

where  $x_t$  is obtained from (3). Correspondingly, the PF-ODE of CSB can be designed as:

$$dx_t = \left[ f(x_t, t) + \frac{1}{2} \beta(t)^2 \frac{x_1 - x_t}{\Sigma(t)^2} \right] dt. \quad (5)$$

To estimate the score, this paper trained a model  $f_\theta(x_t, t, V_f)$ , which can map any state on the same PF-ODE trajectory to the same initial state  $x_0$ .

Specifically, given a data point  $x$ , a pair of adjacent data points  $(x_{t_n}, x_{t_{n+1}})$  can be generated on the same PF-ODE trajectory. Subsequently, the parameters  $\theta$  will be updated through the difference between the Exponential Moving Average (EMA) model [16]. The loss of CSB can be defined as:

$$\mathcal{L}_{CSB}^N(\theta, \theta^-) := \lambda(t_n) d(f_\theta(x_{t_{n+1}}, t_{n+1}, V_f, E_s), f_{\theta^-}(x_{t_n}, t_n, V_f, E_s)), \quad (6)$$

where the distance  $d(\cdot, \cdot)$  is calculated using the LPIPS loss [30],  $\lambda(\cdot) \in \mathbb{R}^+$  is a positive weighting function,  $f_{\theta^-}$  is the target network, and  $f_\theta$  is the online network.

In the generation process, the Gaussian noise  $x_n$  and the  $V_f, E_s$  obtained by the spatial interaction fusion module are directly input into the decoder, and the initial data point  $x_0$  is directly calculated from  $x_t$  at any time on the ODE trajectory.

### 2.3. Spatially-aware Feature Enhancement

In order to further enhance the consistency of audio-visual matching, the U-Net network is used to combine the dual-view spatial cues extracted from the modal interaction network with the left and right channel information generated by the CSB. This paper proposes to use enhanced loss  $\mathcal{L}_{\text{enh}}$ , that is, to calculate the L2 loss between the the ground truth  $x_0$  and the generated sample audio  $x'$ , reduce the distance between the distance between enhanced mel and the ground truth that belong to the same channel and expand the distance between different channels.

$$\mathcal{L}_{\text{enh}} = \|x'_{\text{left}} - x_{0\text{left}}\|^2 + \|x'_{\text{right}} - x_{0\text{right}}\|^2 - \|x'_{\text{left}} - x'_{\text{right}}\|^2, \quad (7)$$

## 3. Experiments

### 3.1. Data Preparation

This paper evaluated the proposed method using the public datasets Opencpop [24] and NVAS-SoundSpace [21]. The Opencpop dataset contains 100 Chinese songs performed by a female singer. The NVAS-SoundSpace dataset contains 13,000 hours of binaural audiovisual data. The average difference was calculated between the left and right channel audio and selected videos where this difference exceeded 0.01 to identify qualifying binaural audio. Subsequently, this paper applied convolutional reverberation by transferring the impulse responses extracted from NVAS-SoundSpace audio to the Opencpop audio, generating binaural audio. Following the data split method proposed by DiffSinger [1], 95 songs were selected as the training set and 5 songs as the validation set. According to the experimental scheme of AViTAR [31], the test set is divided into “test-seen” and “test-unseen”.

### 3.2. Experimental Setup

In our experiments, all audio data were resampled to 22,050 Hz, with the hop size set to 128 and the frame size set to 512. The number of mel bins  $H_m$  was 80. The mel-spectrograms were linearly scaled to the range  $[-1, 1]$ , and the  $F_0$  was normalized to have zero mean and unit variance. The deep neural network was trained on a single NVIDIA 4090 GPU using the Adam optimizer, with an initial learning rate of  $10^{-4}$  and a batch size of 16, over a total of 800,000 steps. The maximum number of discrete steps  $N$  was set to 120. We set the minimum processing time  $\epsilon = 0.001$  and the maximum processing time  $T = 0.999$  to ensure numerical stability. ResNet-18 [25], pretrained on ImageNet [32], was used as the visual encoder to extract visual features.

Since there is currently no model that directly synthesizes stereo audio with room reverberation through visual guidance, this paper chose to evaluate the effectiveness of our proposed method by comparing it with a cascade of models: a singing voice synthesis model, a vision-guided environmental acoustic matching model, and a vision-guided mono-to-stereo conversion model. Specifically, the singing voice generated by the synthesis model is processed through the environmental acoustics matching model to add room reverberation, and finally converted into binaural audio. In the experimental results, we will only use the singing synthesis model to represent the cascade model. The cascaded baseline is constructed based on the following models: 1) GT, the ground truth singing audio. 2) GT(voc.), convert GT’s mel-spectrograms back to audio using vocoder. 3) DiffSinger [1], a singing voice synthesis model based on diffusion model. 4) VISinger2 [2], an end-to-end singing voice synthesis model based on VITS [33]. 5) CoMoSpeech [15], a speech synthesis model using consistency distillation [16] with DiffSinger [1] as the teacher model. 6) LeMARA [34], a visual acoustic matching model that adds room reverberation to a given audio based on images. 7) SepStereo [35], a model that converts mono audio to binaural audio using scene images.

### 3.3. Experimental Results and Analysis

In the objective evaluation, this paper measures performance from four aspects: 1) Real-Time Factor (RTF), the time required for the system to synthesize one second waveform. 2) Mel Cepstral Distortion (MCD), this metric measures the degree of Mel-cepstral distortion between the synthesized speech and the refer-

Table 1: Results of comparison with the cascaded systems.

System	NFE	Test-Unseen					Test-Seen				
		MOS $\uparrow$	RTF $\downarrow$	MCD $\downarrow$	LRE $\downarrow$	RTE $\downarrow$	MOS $\uparrow$	RTF $\downarrow$	MCD $\downarrow$	LRE $\downarrow$	RTE $\downarrow$
GT	-	4.35 $\pm$ 0.12	-	-	-	-	4.32 $\pm$ 0.18	-	-	-	-
GT(voc.)	-	4.23 $\pm$ 0.16	-	1.85	0.154	0.005	4.19 $\pm$ 0.15	-	1.82	0.153	0.005
DiffSinger	50	3.62 $\pm$ 0.11	0.203	7.89	0.945	0.073	3.68 $\pm$ 0.09	0.208	7.88	0.936	0.068
VISinger2	-	3.72 $\pm$ 0.06	0.069	7.81	0.939	0.069	3.79 $\pm$ 0.10	0.069	7.74	0.932	0.064
CoMoSpeech	1	3.46 $\pm$ 0.08	0.052	8.61	0.951	0.070	3.51 $\pm$ 0.07	0.055	8.53	0.942	0.063
CoMospeech	4	3.70 $\pm$ 0.09	0.063	7.74	0.948	0.068	3.76 $\pm$ 0.12	0.064	7.72	0.933	0.062
Ours	1	3.48 $\pm$ 0.12	<b>0.024</b>	8.55	0.982	0.077	3.56 $\pm$ 0.10	<b>0.021</b>	8.49	0.968	0.069
Ours	4	<b>3.74 <math>\pm</math> 0.13</b>	0.033	<b>7.71</b>	<b>0.907</b>	<b>0.065</b>	<b>3.81 <math>\pm</math> 0.08</b>	0.030	<b>7.65</b>	<b>0.889</b>	<b>0.058</b>

Table 2: Ablation Study on the Opencpop and NVAS-SoundSpace datasets.

Model	MOS $\uparrow$	MCD $\downarrow$	LRE $\downarrow$	RTE $\downarrow$
VS-Singer	<b>3.81 <math>\pm</math> 0.08</b>	<b>7.65</b>	<b>0.889</b>	<b>0.058</b>
w/o MIN	3.65 $\pm$ 0.11	7.82	1.257	0.084
w/o SB	3.50 $\pm$ 0.06	8.63	0.899	0.062
w/o SFE	3.72 $\pm$ 0.10	7.72	0.921	0.069

ence speech. 3) Left-Right Energy Ratio Error (LRE), evaluates the correctness of spatial sound by measuring the difference between the energy ratio of the left and right channels. 4) RT60 Error (RTE) [31], assesses the correctness of acoustic properties by measuring the reverberation time error for a 60dB decay (RT60). Number of function evaluation (NFE) is the total number of times the denoiser function is evaluated during the generation process. For the subjective evaluation, to evaluate the naturalness of the synthesized speech, 30 native Chinese speakers were asked to rate the samples using MOS on a scale of 1 to 5.

From Table 1, it can be seen that although the values of each metric in the unseen scene are slightly lower than those in the seen scene, our model achieves the best results in each metric. Compared with the baseline model, our model has significantly lower RTF. This mainly attributed to the fact that the consistency model can generate in just one step. Moreover, our model achieved higher MOS and lower MCD than CoMoSpeech, which uses consistency distillation as singing synthesis part, demonstrating that the introduction of the Schrödinger bridge effectively addresses the performance degradation associated with independently training a consistency model. Additionally, it also can be observed that our model possesses excellent spatial awareness, as evidenced by the lowest LER and RTE values. All results indicate that our model can effectively synthesize speech with corresponding room reverberation based on the spatial information provided by the input images, without compromising speech quality.

### 3.4. Ablation Study

In this subsection, an ablation analysis is conducted to evaluate the impact of each module on the model’s performance. From Table 2, it can be observed that removing MIN leads to a sharp increase in RTE and LRE, indicating that MIN effectively helps the model understand the positioning of characters in the scene. The removal of the SFE module shows a similar

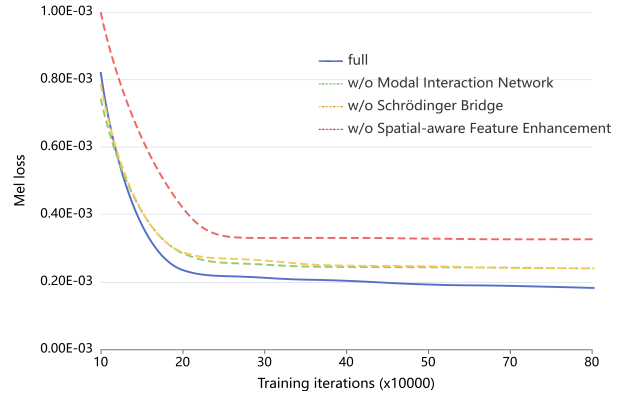


Figure 2: Convergence speed of the models at 800k training steps in the ablation experiment.

trend. Fig. 2 demonstrates that this module accelerates model convergence and aids the model in further perceiving the spatial structure of the scene. As expected, the removal of the Schrödinger bridge while retaining the consistency model results in the lowest MOS and the highest MCD, which suggests that the Schrödinger bridge compensates for the performance degradation caused by training the consistency model independently. All these results indicate the effectiveness of each component in the model.

## 4. Conclusion

This paper proposes VS-Singer, a novel framework that unifies spatial acoustic matching and stereo singing voice generation. VS-Singer consists of a modal interaction network, a decoder based on consistency Schrödinger bridge and a spatially-aware feature enhancement module. The modal interaction network is introduced to add the spatial information into hidden text sequences. Then, a tractable consistency Schrödinger bridge between the linguistic representation and the mel-spectrogram of the binaural channels is established, which boosts performance and speed while reducing training costs. Furthermore, the spatially-aware feature enhancement module can improve its spatial perception ability. Extensive experiments on open-source corpora demonstrate that the proposed method outperforms cascaded systems in both stereo singing voice quality and synthesis speed.

## 5. References

- [1] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, “Diffsinger: Singing voice synthesis via shallow diffusion mechanism,” in *AAAI*, 2022, pp. 11 020–11 028.
- [2] Y. Zhang, H. Xue, H. Li, L. Xie, T. Guo, R. Zhang, and C. Gong, “Visinger2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer,” in *Interspeech*, 2023, pp. 4444–4448.
- [3] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu *et al.*, “Audioopt: Understanding and generating speech, music, sound, and talking head,” in *AAAI*, 2024, pp. 23 802–23 804.
- [4] S. Gao, S. Lei, F. Zhuo, H. Liu, F. Liu, B. Tang, Q. Huang, S. Kang, and Z. Wu, “An end-to-end approach for chord-conditioned song generation,” *arXiv preprint arXiv:2409.06307*, 2024.
- [5] Y. Lei, S. Yang, X. Wang, Q. Xie, J. Yao, L. Xie, and D. Su, “Unisyn: an end-to-end unified model for text-to-speech and singing voice synthesis,” in *AAAI*, 2023, pp. 13 025–13 033.
- [6] T. Karras, M. Aittala, T. Aila, and S. Laine, “Elucidating the design space of diffusion-based generative models,” *Advances in neural information processing systems*, vol. 35, pp. 26 565–26 577, 2022.
- [7] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, “Video diffusion models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.
- [8] J. Liu, C. Li, Y. Ren, Z. Zhu, and Z. Zhao, “Learning the beauty in songs: Neural singing voice beautifier,” in *ACL*, 2022, pp. 7970–7983.
- [9] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, “Align your latents: High-resolution video synthesis with latent diffusion models,” in *CVPR*, 2023, pp. 22 563–22 575.
- [10] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, “An image is worth one word: Personalizing text-to-image generation using textual inversion,” in *ICLR*, 2023.
- [11] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation,” in *CVPR*, 2023, pp. 22 500–22 510.
- [12] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *ICCV*, 2023, pp. 4195–4205.
- [13] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, “Grad-its: A diffusion probabilistic model for text-to-speech,” in *ICML*, 2021, pp. 8599–8608.
- [14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *ICLR*, 2021.
- [15] Z. Ye, W. Xue, X. Tan, J. Chen, Q. Liu, and Y. Guo, “Comospeech: One-step speech and singing voice synthesis via consistency model,” in *MM*, 2023, pp. 1831–1839.
- [16] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, “Consistency models,” in *ICML*, 2023, pp. 32 211–32 252.
- [17] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *arXiv preprint arXiv:2006.04558*, 2020.
- [18] R. Gao and K. Grauman, “2.5 d visual sound,” in *CVPR*, 2019, pp. 324–333.
- [19] W. Lim and J. Nam, “Enhancing spatial audio generation with source separation and channel panning loss,” in *ICASSP*, 2024, pp. 8321–8325.
- [20] Y. H. H. L. Rui Liu, Shuwei He, “Multi-source spatial knowledge understanding for immersive visual text-to-speech,” in *ICASSP*, 2025.
- [21] C. Chen, A. Richard, R. Shapovalov, V. K. Ithapu, N. Neverova, K. Grauman, and A. Vedaldi, “Novel-view acoustic synthesis,” in *CVPR*, 2023, pp. 6409–6419.
- [22] M. Liu, J. Wang, X. Qian, and X. Xie, “Visually guided binaural audio generation with cross-modal consistency,” in *ICASSP*, 2024, pp. 7980–7984.
- [23] Y. H. H. L. Rui Liu, Shuwei He, “Multi-modal and multi-scale spatial environment understanding for immersive visual text-to-speech,” in *AAAI*, 2025.
- [24] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi, “Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis,” in *Interspeech*, 2022, pp. 4242–4246.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [26] Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet, “Diffusion schrödinger bridge matching,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [27] T. Chen, G. Liu, and E. A. Theodorou, “Likelihood training of schrödinger bridge using forward-backward sdes theory,” in *ICLR*, 2022.
- [28] Z. Tang, T. Hang, S. Gu, D. Chen, and B. Guo, “Simplified diffusion schrödinger bridge,” *arXiv preprint arXiv:2403.14623*, 2024.
- [29] G. Liu, A. Vahdat, D. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar, “I<sup>2</sup>sb: Image-to-image schrödinger bridge,” in *ICML*, 2023, pp. 22 042–22 062.
- [30] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *ICLR*, 2018, pp. 586–595.
- [31] C. Chen, R. Gao, P. Calamia, and K. Grauman, “Visual acoustic matching,” in *CVPR*, 2022, pp. 18 858–18 868.
- [32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, 2009, pp. 248–255.
- [33] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *ICML*, 2021, pp. 5530–5540.
- [34] A. Somayazulu, C. Chen, and K. Grauman, “Self-supervised visual acoustic matching,” in *NeurIPS*, 2023.
- [35] H. Zhou, X. Xu, D. Lin, X. Wang, and Z. Liu, “Sep-stereo: Visually guided stereophonic audio generation by associating source separation,” in *ECCV*, 2020, pp. 52–69.