Self-Supervised Audio-Visual Soundscape Stylization

Tingle Li[1], Renhao Wang[1], Po-Yao Huang[2],
Andrew Owens[3], and Gopala Anumanchipalli[1]

[1]University of California, Berkeley [2]FAIR, Meta [3]University of Michigan
https://tinglok.netlify.app/files/avsoundscape

Abstract. Speech sounds convey a great deal of information about the scenes, resulting in a variety of effects ranging from reverberation to additional ambient sounds. In this paper, we manipulate input speech to sound as though it was recorded within a different scene, given an audio-visual conditional example recorded from that scene. Our model learns through self-supervision, taking advantage of the fact that natural video contains recurring sound events and textures. We extract an audio clip from a video and apply speech enhancement. We then train a latent diffusion model to recover the original speech, using another audio-visual clip taken from elsewhere in the video as a conditional hint. Through this process, the model learns to transfer the conditional example's sound properties to the input speech. We show that our model can be successfully trained using unlabeled, in-the-wild videos, and that an additional visual signal can improve its sound prediction abilities.

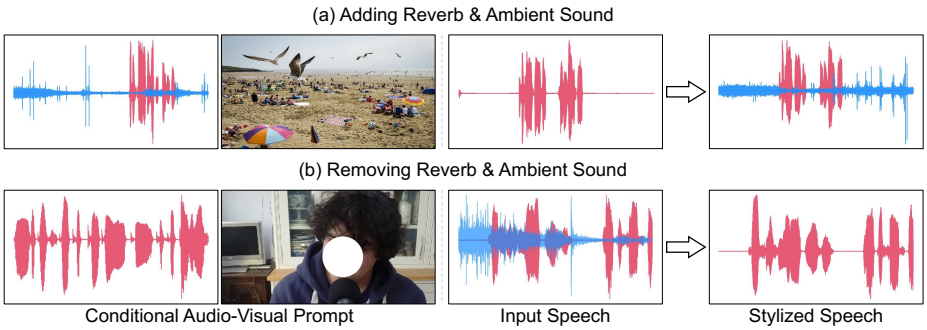


Fig. 1: Audio-visual soundscape stylization. We learn through self-supervision to manipulate input speech (middle) such that it sounds as though it were recorded within a given scene (left). Our approach captures both acoustic properties, such as reverberation, as well as the ambient sounds, such as crashing waves (top). To help convey the results of the stylization, we have used source separation to visualize the speech waveform (shown in red) separately from background sound (shown in blue).

1 Introduction

Speech conveys a tremendous amount about the scene that it was produced in, from the material properties of its surfaces to its ambient sounds [28, 72]. A

major goal of the audio and audio-visual generation communities has been to accurately resynthesize speech to sound as though it were recorded in a different scene [5, 80, 88], thereby capturing these subtle details — a task that has a number of applications, from movie dubbing to virtual reality.

Existing formulations of this problem have largely focused on reproducing room acoustic properties, such as adding reverb by manipulating the room impulse response [5,81]. However, these approaches do not model many of the other ways that a scene can affect a recorded sound. For example, when we walk on the beach (Figure 1), we may experience the whispers of the wind, the cries of seagulls, and the crash of waves — a distinctive ambient sound texture [63] that would be conveyed in any sound recording taken within the scene. Since these aspects of a soundscape are not modeled by existing resynthesis techniques, making a sound fully reflect a scene requires additional postprocessing, such as "mixing in" background noises that fit the scene. This process, often reliant on descriptive language, can be time-consuming and constrained in its ability to convey subtle auditory properties. Moreover, these methods have largely required simulated (or labeled) training data, and are not designed to learn from abundantly available "in the wild" audio-visual data, limiting their scalability.

We propose the *audio-visual soundscape stylization* problem. Given a visual or audio-visual example from a scene and a clean input speech, our goal is to manipulate the input speech such that it could have occurred within the scene, reproducing both the acoustic properties of a scene and the ambient sounds within it. We propose a simple, self-supervised model that can be trained entirely using "in the wild" internet video. Our method is based on *conditional speech de-enhancement*: we randomly sample two nearby audio-visual clips from a video, and remove scene-specific attributes from one of them performing speech enhancement. We then train a model, based on latent diffusion [77], to reverse this speech enhancement process, using the other audio-visual clip as a conditional "hint." In order to perform this task, the model needs to infer the acoustic and ambient properties of the scene, and to successfully transfer them to an input speech. At test time, we give the model a conditional example from the scene whose properties we would like to transfer.

Our model is simple and can be trained entirely using in-the-wild egocentric videos. We show through quantitative evaluations and perceptual studies that our method learns to successfully stylize sounds in a number of challenging in-the-wild scenarios, transferring both the acoustic properties and ambient sounds to input speech. As part of these experiments, we find that our model can successfully transfer sound from visual conditioning, and that visual signals improve our model's ability to stylize audio. We also find that we can outperform existing work on the previously proposed problem of styling sound using room acoustic properties from images [5] while going beyond this work in also transferring ambient sound. Finally, we find that "prompting" our model with specific conditional examples can achieve a desired style, such as approximately converting between near- and far-field speech, and that our model can successfully restyle a variety of non-speech input sounds.

2 Related Work

Stylization in image and audio. The concept of image stylization was pioneered by Hertzmann et al. [32], which restyled input images based on a single user-provided example. Various types have been explored for image stylization, including image [39, 44, 100], text [3, 4, 14], sound [54, 58], and touch [94, 95]. Recent work has addressed a variety of audio stylization tasks, such as voice conversion (via feature disentanglement [87, 90] or adversarial learning [45, 57]), music timbre transfer [38], text-driven audio editing [92], visual acoustic matching [5, 81], and audio effects stylization [84]. In particular, Chen et al. [5] proposed to manipulate the room impulse response based on the surrounding images using a generative adversarial network (GAN). Recently, Somayazulu et al. [81] used an additional GAN to further denoise the target audio to get the paired data for training. In contrast to these works, our method differs by: (i) Generating ambient sounds beyond mere room impulse response; (ii) Using a more expressive diffusion model rather than GAN; (iii) Learning from "in-the-wild" internet videos instead of curated indoor videos.

Sound generation from visual and textual inputs. Generating sound from visual and textual inputs has recently attracted much research attention. For visual-based methods, researchers have explored the generation of sound effects, music, speech, and ambient sound from visual cues such as impacts [68], musical instrument playing [49], dancing [23, 85], lip movements [19, 35, 73], and open-domain images [42, 62, 79, 99]. In contrast to these methods, which focus on generating a specific type of sound from visual input, our method is centered on stylizing soundscapes to fit their environmental context, guided by conditional audio-visual examples. For text-based methods, Yang et al. [93] introduced a discrete diffusion model for generating ambient sound from text descriptions. Kreuk et al. [53] used VQGAN [20] for sound generation. Recently, latent diffusion models [37, 59] have significantly improved generation quality. All these methods necessitate text annotations to establish text-audio pairs, whether at the training [53] or representation [37, 59] level, which can be labor-intensive and limited in expressiveness with plain text descriptions. In contrast, we make an input speech match a given soundscape that is specified by a given audio-visual example, without the need for annotations or text. This allows users to select the (often difficult-to-articulate) auditory properties "by example" rather than through language. Our approach thus provides a complementary learning signal to text-based methods.

Audio-visual learning. The natural correlation between audio and visual frames in videos has facilitated extensive audio-visual research, including representation learning [2, 36, 52, 66, 67, 69, 70], source separation [18, 24, 56, 97, 98], audio source localization [7, 9, 29], audio spatialization [25, 65, 96], visual speech recognition [1], deepfake detection [21], and scene classification [8, 15, 26]. Inspired by this line of work, we aim to stylize input audio to match the soundscape of a single audio-visual conditional example.

3 Audio-Visual Soundscape Stylization

Our goal is to manipulate an input sound such that it could plausibly have been recorded within another scene, given a conditional audio-visual example from that scene. We learn a function $\mathcal{F}_\theta(\boldsymbol{a}_e, \boldsymbol{a}_c, \boldsymbol{i}_c)$ parameterized by θ that stylizes an input sound $\boldsymbol{a}_e$ given the reference audio $\boldsymbol{a}_c$ and its corresponding image $\boldsymbol{i}_c$. We show that $\mathcal{F}_\theta$ can be learned solely from unlabeled audio-visual data.

3.1 Self-Supervised Soundscape Stylization

We propose a self-supervised task that trains a model to stylize input sounds, using audio-only, visual-only, or audio-visual conditioning.

Learning by audio-visual speech de-enhancement. As a self-supervised pretext task [12], we randomly select an audio clip from a training video, apply source separation and enhancement to it, and train a model to undo this operation (*i.e.*, to recover the original sound) after conditioning on another audio-visual example taken from the same training video (Fig. 2). We observe that the background noises and acoustic properties within a video tend to exhibit temporal coherence [16], especially when sound events occur repeatedly. Moreover, similar sound events often share semantically similar visual appearances [40, 69]. By providing the model with a conditional example from another time step in the video, the model is implicitly able to estimate the scene properties and transfer these to the input audio (*e.g.*, the reverb and ambient background sounds). At test time, we will provide a clip taken from a *different* scene as conditioning, forcing the model to match the style of a desired scene.

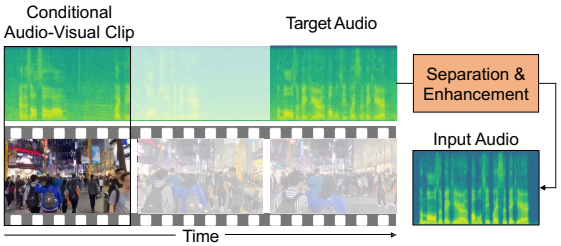


Fig. 2: Soundscape stylization by conditional speech de-enhancement. We randomly select two disjoint clips from a video, designating one as a *conditional* example and the other as the *target*. We then separate and enhance the target audio. Our model's self-supervised pretext task is to remove this enhancement using the other conditional (audio, visual, or audio-visual) signal as a hint. At test time, we stylize an audio clip using a conditional example from the desired scene.

Specifically, we sample non-overlapping clips from a long video, centered at times τ and τ'. One of these clips serves as the conditional audio-visual example, denoted as $\boldsymbol{a}_c$ and $\boldsymbol{i}_c$, while the soundtrack of the other clip is designated as the target audio $\boldsymbol{a}_q$. We then apply both a source separation model [71] to isolate the foreground speech, and a speech enhancement and dereverberation model [43] to further refine the separated speech quality. This process results in the generation of high-fidelity speech $\boldsymbol{a}_e = \mathcal{H}(\boldsymbol{a}_q)$, which sounds as if it were recorded in a soundproofed studio. Please refer to the supplement for the analysis of different

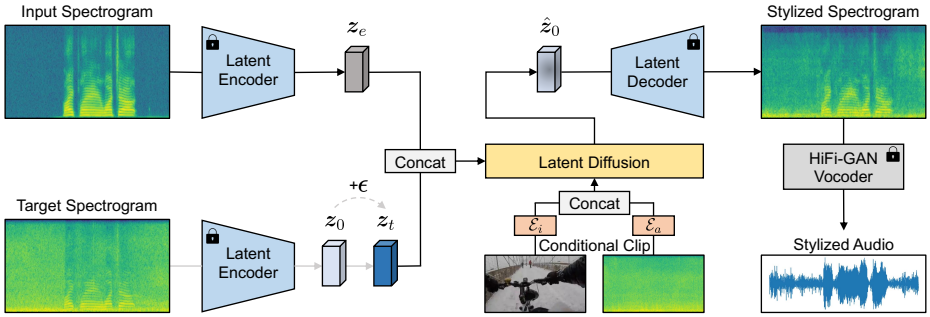


Fig. 3: Model architecture. Given input audio derived from an enhancement model, and the conditional audio-visual clip sampled from the same video, we aim to stylize the input to closely resemble the original signal. We encode both the input and target spectrograms to the latent space using a pre-trained latent encoder, and feed them into a latent diffusion model together with the conditional audio-visual embedding. The goal is to harmonize the encoded latent of the input spectrogram with the target one. Finally, we employ a pre-trained latent decoder followed by a pre-trained HiFi-GAN vocoder to reconstruct the waveform from the latent space. Note that the latent encoder for the target spectrogram is *not* used at test time.

enhancement strategies. For preprocessing, we use an off-the-shelf voice activity detector [86] to ensure that each selected audio clip is likely to contain speech.

Sound stylization model. After training on the pretext task of audio-visual speech de-enhancement, the resulting model is able to tailor its stylization according to the conditional examples, which aligns with the assumption that the conditional example is instructive for the input audio. At test time, we retain the flexibility to substitute the conditional example with a completely different audio-visual clip, enabling the potential for one-to-many stylization.

3.2 Conditional Stylization Model

We describe our conditional soundscape stylization model $\mathcal{F}_\theta$ (Figure 3), which is designed to stylize input audio based on a conditional audio-visual pair and consists of three main components: i) compressing the input audio into a latent space; ii) applying soundscape stylization using the conditional latent diffusion model; iii) reconstructing the waveform from the latent space.

Conditional latent diffusion model. We train a conditional latent diffusion model to stylize soundscapes based on conditional audio-visual examples. Building upon the denoising diffusion probabilistic model [33] and the latent diffusion model [77], our model breaks the generation process into N conditional denoising steps, and improves the efficiency of diffusion models by operating in the latent space. Therefore, our soundscape stylization model $\mathcal{F}_\theta$ can be interpreted as an equally weighted sequence of denoising auto-encoders $\boldsymbol{\epsilon}_\theta$.

Specifically, it takes the encoded latent of the input audio and the conditional audio-visual example as a conditioning signal. To elaborate further, given the

encoded latent of the original audio $\boldsymbol{z}_0 = \text{Enc}(\boldsymbol{a}_q)$, the conditional audio-visual example $(\boldsymbol{a}_c, \boldsymbol{i}_c)$, a random denoising step t, and random noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, our model first generates a noisy version $\boldsymbol{z}_t$ via a noise schedule [83]. We then define the training loss $\mathcal{L}_\theta$ by predicting the noise $\boldsymbol{\epsilon}$ added to the noisy latent, guided by the input audio $\boldsymbol{a}_e$ and the conditional audio-visual pair $(\boldsymbol{a}_c, \boldsymbol{i}_c)$. This can be achieved by minimizing the loss function as follows:

$$\mathcal{L}_\theta = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{a}_c, \boldsymbol{i}_c, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_e, \boldsymbol{a}_c, \boldsymbol{i}_c)\|_2^2 \ , \tag{1}$$

where $\boldsymbol{z}_e = \text{Enc}(\boldsymbol{a}_e)$ is the encoded latent of the input audio $\boldsymbol{a}_e$.

Adding noise to the input. We propose to add Gaussian noise $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ to the enhanced audio at both training and test time. This mixed audio is then employed as the input audio. The primary purpose of this is to mitigate the effect of "audio nostalgia" [13]. This conceals subtle remnants of the original sound that may persist in the enhanced audio, which could potentially help avoid leaking information in the pretext task. Specifically, when the model is trained on the enhanced audio, the generated soundscapes might exhibit a suspicious resemblance to the originals. Conversely, when clean speech is used as input, the output may largely replicate the input. We consider this addition of noise as a type of data augmentation.

Compressing mel-spectrograms. We employ a ResNet-based variational auto-encoder (VAE) [30, 48] to compress the mel-spectrogram $\boldsymbol{a} \in \mathbb{R}^{T \times F}$ into a latent space $\boldsymbol{z} \in \mathbb{R}^{T/r \times F/r \times d}$, where r denotes the compression level, T/r and F/r is a lower-resolution time-frequency bin, and d represents the embedding size at each bin. The VAE is tasked with reconstructing sounds from a dataset, where the bottleneck can then serve as the encoded latent. For our experiments, we adopt a pre-trained VAE model from Liu et al. [59].

Conditional audio-visual representations. The conditional audio-visual example is represented using its latent vector. We employ separate audio and image encoders, denoted as $\mathcal{E}_a(\cdot)$ and $\mathcal{E}_i(\cdot)$, to extract audio embeddings $\mathcal{E}_a(\boldsymbol{a}_c) \in \mathbb{R}^L$ and image embeddings $\mathcal{E}_i(\boldsymbol{i}_c) \in \mathbb{R}^L$, where L represents the embedding size. We use pre-trained encoders like CLIP [74] and CLAP [17] for image and audio representation respectively. Prior to fusion, we apply linear projections to the image and audio embeddings, followed by concatenating and feeding them into the diffusion model through cross-attention mechanism [89].

Classifier-free guidance. We use classifier-free guidance [34] to balance the trade-off between the quality and diversity of generated samples. It involves jointly training the model for both conditional and unconditional denoising. During training, we randomly nullify the conditional model with a fixed probability of 10%. At test time, a guidance scale ($\lambda \geq 1$) is utilized to adjust the score estimates, skewing them towards the conditional $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_e, \boldsymbol{a}_c, \boldsymbol{i}_c)$ and away from the unconditional $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_e, \varnothing, \varnothing)$.

$$\begin{aligned} \tilde{\boldsymbol{\epsilon}}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_e, \boldsymbol{a}_c, \boldsymbol{i}_c) = \lambda \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_e, \boldsymbol{a}_c, \boldsymbol{i}_c) \\ + (1 - \lambda) \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_e, \varnothing, \varnothing) \end{aligned} \tag{2}$$

We find this guidance enhances the output quality and relevance of stylized samples.

Recovering the waveform. Following the estimation of noise $\tilde{\epsilon}_\theta$ from the diffusion model, we retrieve the encoded latent of the stylized mel-spectrogram. This latent is then fed into the VAE decoder to reconstruct the stylized mel-spectrogram. Finally, a pre-trained HiFi-GAN vocoder [50] is employed to reconstruct the waveform.

4 Experiments

4.1 Experimental Setup

We evaluate our model's ability to restyle sounds to match new environments.

Dataset. Our goal is to train and evaluate the model on minimally curated "in-the-wild" videos. To do this, we train and evaluate on two different datasets: *CityWalk* and *Acoustic-AVSpeech*.

- ***CityWalk* dataset**: We collect a *CityWalk* dataset which includes egocentric videos with diverse ambient sounds and acoustic properties, recorded in places like trains, buses, streets, beaches, shopping malls, *etc*. Using the search terms "City Walk + POV" on YouTube, we gather 3,447 videos derived from indoor (28%) and outdoor (72%) scenes. From this collection, we choose a subset of 235 videos for training and testing, with lengths varying between 5 to 225 minutes, totaling 158 hours. We ensure that these videos only contain naturally occurring sounds in the scenes, without any post-edited voice-overs or music. We also guarantee that the sources of training and testing videos do not overlap. Please see the supplement for more dataset details.
- ***Acoustic AVSpeech* dataset** [5]: The *Acoustic-AVSpeech* dataset is a subset of the AVSpeech dataset [18] that contains 3-10 seconds indoor clips of single speakers without interfering ambient sound. Since the visual content of these clips offers useful insights into the geometry and materials of the scenes, it can be utilized to estimate the acoustic properties (but not ambient sound). We use this dataset for a fair comparison with the existing baseline [5] (we are unable to compare with [81] as it is not open source).

Model configurations. We use the VAE and HiFi-GAN vocoder from [59], which are trained on the combination of AudioSet [26], AudioCaps [47], BBC Sound Effect [10] and Freesound [22] datasets. The VAE is configured with a compression level r of 4 and latent channels d of 8. For extracting audio and image embedding, we have two options: i) use a from-scratch ResNet-18 encoder; ii) utilize a fine-tuned CLAP audio encoder [17] derived from our audio-only model, alongside a fixed CLIP image encoder [74]. These encoders are integrated into the diffusion model through late fusion [91] and cross-attention [89]. The diffusion model is based on a U-Net backbone, consisting of four encoder and decoder blocks with downsampling and upsampling and a bottleneck in between. Multi-head attention with 64 head features and 8 heads per layer is applied in the

last three encoder and first three decoder blocks. During the forward process, we employ $N = 1000$ steps and a linear noise schedule, ranging from $\beta_1 = 0.0015$ to $\beta_N = 0.0195$, to generate noise. Additionally, we leverage the DDIM sampling method [83] with 200 sampling steps. For classifier-free guidance, we set the guidance scale λ to 4.5, as described in Equation (2).

Training procedures. To enhance training efficiency, we divide all videos into 10-second video and audio clips. We apply a frame-level voice activity detector [86] to the resulting audio clips to detect speech onset. Subsequently, we randomly select two 2.56-second audio clips from the same source – one for the target audio and the other for the conditional audio. The conditional image is chosen by randomly sampling one video frame within the scope of the selected conditional audio. Our model is trained using the AdamW optimizer [61] with a learning rate of 10^{-4}, $\beta_1 = 0.95$, $\beta_2 = 0.999$, $\epsilon = 10^{-6}$, and a weight decay of 10^{-3} over 200 epochs.

Evaluation metrics. To assess the performance of our models, we use both objective and subjective metrics. Our objective metrics include the *Mean Square Error* (MSE), *RT60 Error* (RTE), *Mean Opinion Score Error* (MOSE), *Perceptual Evaluation of Speech Quality* (PESQ) [76], *Fréchet Audio Distance* (FAD) [46], *Fréchet Distance* (FD) [59], *Kullback-Leibler divergence* (KL), *Word Error Rate* (WER), *Inception Score* (IS) [78], and *Audio-Visual Correspondence* (AVC) [2]. MSE evaluates how closely the stylized audio matches the ground truth in terms of magnitude spectrograms (if the ground truth is available), while RTE measures the MSE between RT60 estimates of generated and target speech, *i.e.*, the differences in the reverb level. MOSE assesses the difference in speech quality between the generated audio and ground truth using MOSNet [60]. FD and FAD measure the similarity between real and generated audio using different classifiers (FAD employs VGGish [31] and FD uses PANNs [51]). KL quantifies the distributional similarity between real and generated audio, while PESQ and IS evaluate the quality and diversity of generated audio. WER measures the intelligibility of the generated audio using a pre-trained speech recognizer [75]. AVC assesses the correlation between audio and image, utilizing features extracted by either OpenL3 [11] or ImageBind (IB) [27].

In addition, we conduct a subjective evaluation through Amazon Mechanical Turk. Human participants are asked to rate audio generated by various methods based on its similarity to the soundscapes in the given audio-visual example. This rate considers four criteria: *overall quality* (OVL), *relation to ambient sounds* (RAM), *relation to acoustic properties* (RAC), and *relation to visuals* (RVI), with scores ranging from 1 (low correlation) to 5 (high correlation). Please see the supplement for more human evaluation details.

Baselines. We consider a variety of baselines for comparison:

– **AViTAR** [5]: AViTAR is a GAN-based method that is initially proposed to generate indoor room impulse responses conditioned on images, which is not directly compatible with our setting. To address this issue, we integrate an additional audio conditioning branch into this model, and retrain it on our dataset for fair comparison.

Method	MSE* (↓)	RTE* (↓)	PESQ* (↑)	FD (↓)	FAD (↓)	KL (↓)	WER (↓)	IS (↑)	AVC (↑)	
									IB	L3
Ground Truth	/	/	/	/	/	/	/	0.22	0.98	
Cap. (aud) [64]	2.34	0.91	1.85	14.30	9.73	1.09	0.29	1.49	0.08	0.80
Cap. (sfx) [64]	2.09	0.86	2.05	12.53	9.12	1.14	0.16	1.51	0.09	0.81
Cap. (img) [55]	2.23	0.89	2.02	17.27	9.24	1.30	0.17	1.46	0.08	0.79
Aud Anlg. [59]	1.37	0.77	2.34	9.33	3.97	0.91	0.12	1.53	0.11	0.82
S & R [71]	1.02	0.71	2.54	8.65	3.34	0.71	**0.11**	1.51	0.12	0.83
AViTAR [5]	0.76	0.32	2.56	7.44	3.02	0.68	0.17	1.47	0.14	0.87
Ours	**0.54**	**0.20**	**2.83**	**5.13**	**1.64**	**0.59**	**0.11**	**2.03**	**0.17**	**0.92**

Table 1: Quantitative objective results on the *CityWalk* dataset. Captioning (Cap.) can be driven by original conditional audio (aud), separated sound effects (sfx), and conditional images (img). Aud Anlg. and S & R refer to Audio Analogy and Separate & Remix respectively. * indicates metrics are evaluated using test set with ground truth.

- **Captioning**: This cascaded approach employs pre-trained image [55] or audio [64] captioning models to generate captions from conditional examples, which are then used to generate sound effects using a text-to-audio model [59].
- **Audio Analogy** [59]: AudioLDM is originally used for text-to-audio synthesis. Here we switch the text input with the audio one, allowing us to perform audio-to-audio analogies. We train an audio-visual conditioning model, where we condition on the isolated sound effects instead of the original audio to enforce that the resulting audio does not include any additional speech.
- **Separate & Remix** [71]: This method adopts a simple "copy and paste" strategy. It uses a pre-trained source separation model to isolate sound effects from the conditional audio, and overlays them onto the input audio at a constant signal-to-noise ratio (SNR) of 8 (we empirically find this can boost both the metrics and auditory perception).

4.2 Comparison to Baselines

Quantitative results. We start by presenting the quantitative results on the *CityWalk* dataset in Table 1. Our model, whether operating in an uni-modal (Table 5) or audio-visual (Table 1) conditioning, consistently outperforms all the baselines across multiple objective metrics. These results suggest that our model excels in generating more realistic soundscapes compared to the baselines. In particular, Separate & Remix is worse than our method, despite the fact that it receives nearly the same ambient sounds from the conditional audio. This is probably because our method can not only manipulate ambient sounds but also acoustic properties. We also find that all three Captioning methods perform worse than Separate & Remix, perhaps due to errors introduced by automatic captioning. Among the Caption-based methods, we observe that using separated sound effects produces more precise captions than the others, leading to the best performance. Although Audio Analogy is trained to resemble the similar ambient sounds of the conditional audio, it still cannot achieve comparable performance to Separate & Remix, whereas our method can. This highlights the importance of acoustic properties when it comes to soundscape stylization. Moreover, our

Method	OVL (↑)	RAM (↑)	RAC (↑)	RVI (↑)
Ground Truth	4.03 ± 0.09	/	/	4.15 ± 0.11
Cap. (aud) [64]	2.58 ± 0.14	2.53 ± 0.12	3.08 ± 0.10	3.13 ± 0.07
Cap. (sfx) [64]	2.77 ± 0.08	3.01 ± 0.07	3.18 ± 0.13	3.22 ± 0.12
Cap. (img) [55]	2.14 ± 0.10	2.22 ± 0.14	3.07 ± 0.15	3.09 ± 0.10
Aud Analg. [59]	3.08 ± 0.13	3.03 ± 0.10	3.15 ± 0.11	3.12 ± 0.13
S & R [71]	3.16 ± 0.09	3.34 ± 0.11	3.22 ± 0.07	3.34 ± 0.08
AViTAR [5]	3.32 ± 0.11	3.48 ± 0.07	3.39 ± 0.12	3.40 ± 0.06
Ours	**3.68 ± 0.14**	**3.72 ± 0.08**	**3.55 ± 0.09**	**3.59 ± 0.06**

Table 2: Quantitative subjective results on the *CityWalk* dataset, where OVL, RAM, RAC, and RVI are presented with 95% confidence intervals.

method surpasses AViTAR, manifesting that the diffusion model excels in producing audio of higher quality compared to the GAN-based counterpart.

To further validate our model's performance, we conduct a human evaluation. We randomly select 100 generated audio samples from the test set, with each sample scored by 40 participants. To prevent random submissions, we include one control set consisting entirely of noise. The participants consistently favor our model's stylized audio, as indicated in Table 2, which aligns with the objective evaluation results. Interestingly, we observe that the RAC metrics of the first three methods (Captioning, Audio Analogy, and Separate & Remix) are on par with each other. This consistency could be attributed to the fact that their output is mixed with the same speech as the input, without considering the difference in acoustic properties. This finding also emphasizes the importance of considering acoustic nuances in soundscape stylization, which our model effectively addresses. Furthermore, while AViTAR notably excels beyond other baselines, the inherent challenges in training GAN lead to its audio quality and similarity being consistently worse than those restyled by our method.

Additionally, to ensure a fair comparison with AViTAR, we retrain our model using the *Acoustic AVSpeech* dataset specifically for the task of visual acoustic matching [5]. As illustrated in Table 3, the quantitative comparison reveals that our method surpasses AViTAR in metrics under seen and unseen settings, thereby demonstrating our method's superior capability in capturing inherent acoustic properties in conditional images.

Method	Seen		Unseen	
	RTE (↓)	MOSE (↓)	RTE (↓)	MOSE (↓)
AViTAR [5]	0.144	0.481	0.183	0.453
Ours	**0.098**	**0.412**	**0.124**	**0.399**

Table 3: Quantitative comparisons of our method and AViTAR on the *Acoustic AVSpeech* dataset.

Qualitative results. We visualize how our results vary under different conditional examples and compare our model with baselines in Figure 4. We also provide additional qualitative results in the supplement. For the caption-based methods, we only present the best model, which relies on isolated sound effects. Notably, we observe that while these methods occasionally align with the provided conditional examples, they often falter in most instances (like the artifacts introduced in the second example). Audio Analogy appears promising for generating ambient sounds that align with the conditional examples but falls

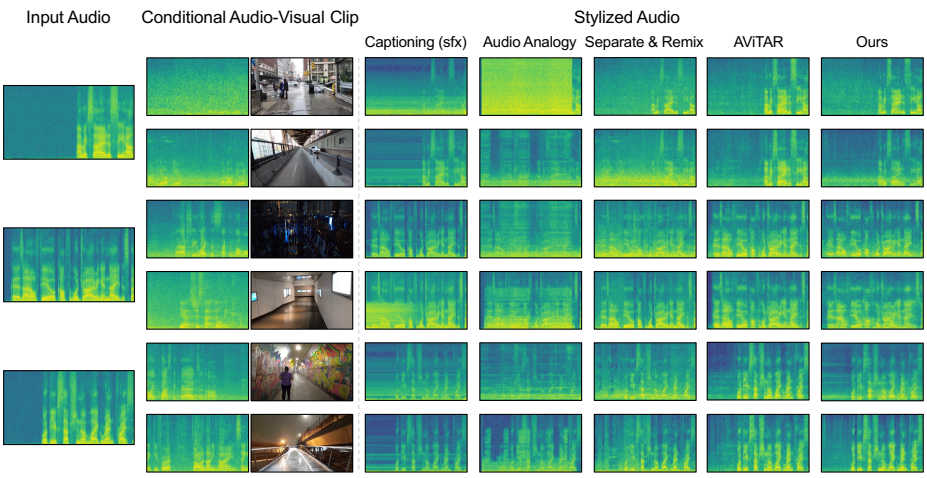


Fig. 4: Model comparison. We show soundscape stylization results for several models, where each input audio is conditioned on two different audio-visual clips.

short in complex scenarios, such as the rainy street in the first example. Separate & Remix tends to directly replicate ambient sounds without considering the specific acoustic environment, leading to less precise acoustic reproduction (like the reverb in the last two examples). AViTAR ranks as the closest approach to our method, but it struggles to capture high frequencies accurately, negatively affecting the overall quality and intelligibility of the output. Our method stands out for its ability to resemble the soundscapes of the conditional example with higher fidelity. For a more direct experience of our model's capabilities, we strongly encourage readers to check out the results video available on the project webpage.

4.3 Ablation Study and Analysis

Table 4 presents the ablation studies on the *CityWalk* dataset. We analyze the following model variants: (i) Using the clean audio from the enhancement model as inputs instead of adding Gaussian noise; (ii) Employing only the source separation model to isolate the target speech that preserves the original acoustic properties (whereas our default method involves employing a speech enhancement model afterward to modify speech properties); (iii) Using only the separated sound effects (no speech) and their corresponding images as conditions; Randomly swapping either (iv) the conditional audio, or (v) the conditional image with another at test time to create misaligned audio-visual conditioning; (vi) Using only the conditional model (no CFG) to stylize input; (vii) Using only the unconditional model to stylize input; (viii) Training a ResNet-18-based audio-visual encoder [30] from scratch rather than using pre-trained CLIP and CLAP. Judging from the results, we draw the following observations:

Adding noise mitigates audio nostalgia. We ask whether adding noise will help mitigate the effect of "audio nostalgia" described in Section 3.2. Table 4

Method	RTE* (↓)	PESQ* (↑)	FD (↓)	FAD (↓)	KL (↓)	IS (↑)
(i) Clean Input	0.41	2.49	6.12	2.66	0.75	1.49
(ii) Separation-only Cond.	0.60	2.48	6.11	2.44	0.73	1.46
(iii) Sfx-only Cond.	0.65	2.42	6.83	2.96	0.90	1.45
(iv) Random Audio Cond.	0.71	2.20	9.51	4.92	1.05	1.44
(v) Random Image Cond.	0.68	2.28	9.33	4.84	0.97	1.46
(vi) No CFG	0.53	2.33	7.44	3.37	0.79	1.28
(vii) No Cond.	0.98	1.98	16.77	7.59	1.27	1.45
(viii) From Scratch	0.44	2.50	6.18	2.41	0.74	1.51
Ours-full	**0.20**	**2.83**	**5.13**	**1.64**	**0.59**	**2.03**

Table 4: Quantitative ablation studies on the *CityWalk* dataset.

clearly demonstrates that our approach outperforms the model trained on clean speech by a large margin, providing empirical evidence to support our hypothesis.

Acoustic properties play an important role in soundscape stylization. We investigate whether our method can resemble plausible acoustic properties to the conditional examples. To examine this, we first train a model with speech extracted from a separation model, matching the acoustic properties of the target and thus excluding the acoustic factor in stylization. Moreover, we train another model conditioned solely on the separated sound effects and their corresponding images. In this scenario, since the conditional audio does not contain speech, it is not sufficiently informative about acoustic variations, leading to arbitrary acoustic changes at test time. As depicted in (ii) and (iii) of Table 4, the performances of these variants drastically decline, indicating the critical role of acoustic properties when it comes to soundscape stylization.

Visual conditioning complements audio conditioning. We explore the impact of visual conditioning on performance. To this end, we create misaligned audio-visual pairs by substituting either the conditional audio or its corresponding image with a random one, and then assess our model's performance with these pairs. As illustrated in (iv) and (v) of Table 4, our model exhibits similar resistance against such perturbations, regardless of whether it is conditioned on the original audio with random images or their inverted versions. This implies that the visual-only model can capture and interpret scene properties, and that visual conditioning delivers complementary information for soundscape stylization than audio conditioning alone.

Pre-training, CFG, and conditioning enhance stylization. We assess the impact of pre-training, CFG, and conditioning on enhancing output relevance. Table 4 demonstrates that our approach outperforms variants that lack these components, demonstrating their effectiveness in improving output relevance.

4.4 Cross-Modal and Cross-Domain Evaluation

Comparison to uni-modal models. We explore the performance of CLAP audio and CLIP image encoders under various conditional settings. Table 5 presents a comparison between the fine-tuned audio encoder and its non-fine-tuned counterpart. Notably, fine-tuning significantly enhances performance, suggesting that

A	V	FT-A	FT-V	RTE* (↓)	PESQ* (↑)	FD (↓)	FAD (↓)	KL (↓)	IS (↑)	IB (↑)	L3 (↑)
✗	✓	✗	✗	0.47	2.45	6.40	2.28	0.91	1.41	0.125	0.882
✗	✓	✗	✓	0.44	2.39	6.39	2.29	0.91	1.40	0.123	0.884
✓	✗	✗	✗	0.61	1.22	10.22	6.12	0.88	1.53	0.111	0.817
✓	✗	✓	✗	0.38	2.44	5.94	2.08	0.71	1.74	0.137	0.892
✓	✓	✓	✗	**0.20**	**2.83**	**5.13**	**1.64**	**0.59**	**2.03**	**0.172**	**0.915**

Table 5: Comparison of our uni-modal (A or V) and audio-visual models. A: Audio; V: Visual; FT-A: Fine-tuning for audio; FT-V: Fine-tuning for visual.

the original CLAP model has limited generalization capabilities for our dataset. Conversely, fine-tuning the image encoder has little effect on performance gain, indicating its inherent strong generalization abilities. Furthermore, we observe that the fine-tuned audio conditioning model surpasses the image conditioning one, which suggests that audio is a more informative modality for representing soundscapes than visual. Based on these findings, we adopt a late fusion approach [91], combining a fine-tuned CLAP audio encoder with a fixed CLIP image encoder for our audio-visual model. This configuration achieves the best performance among all the models, demonstrating the cross-modal information from the visual modality can help craft more comprehensive soundscapes than the audio modality alone.

Generalization to other datasets. We evaluate the generalization capabilities of our model using out-of-distribution data. Specifically, we explore the model's proficiency in restyling speech from the LRS dataset [82] conditioned on clips from the AVSpeech dataset [18]. As illustrated in Figure 5, we use the clips captured in the indoor room, lecture, coffee shop, and church. We show that our model, whether conditioned on audio-visual or visual-only clips, exhibits robust in-context learning capabilities, effectively adjusting acoustic properties to suit far-field conditions, generating or eliminating ambient sounds, and reducing reverb to enhance speech clarity in alignment with the conditional clips.

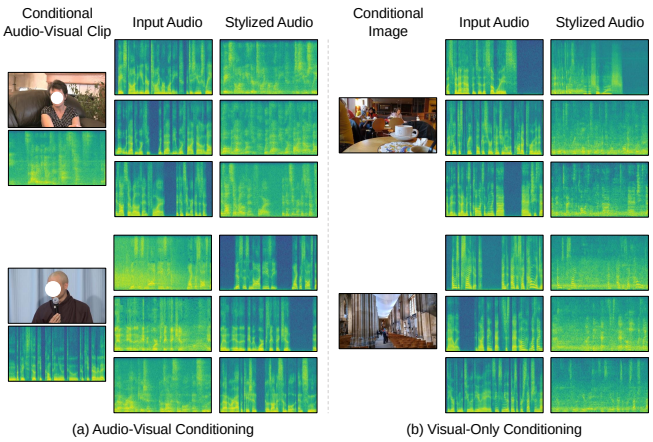


Fig. 5: Qualitative generalization results. We restyle audio from LRS [82] conditioned on audio-visual (or visual-only) clips taken from AVSpeech [18].

Generalization to non-speech sounds. We examine the adaptability of our model to non-speech sounds. To test this, we introduce sounds such as dog barks and train chimes to the model. As shown in Figure 6, we find that our model is

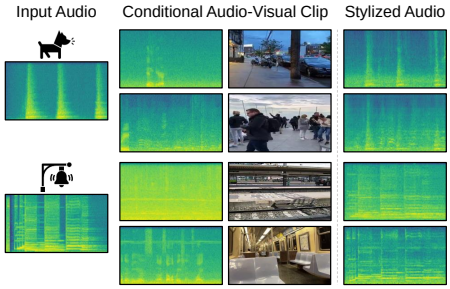


Fig. 6: Generalization to non-speech sounds. We restyle the sounds of dog barks and train chimes like they are in the conditional scenes.

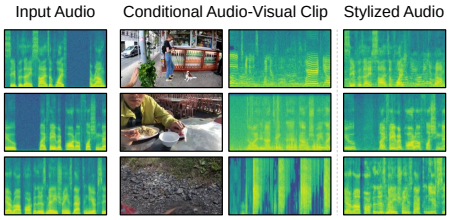


Fig. 7: Failure cases. Our model fails to resemble the soundscapes of the scenes, perhaps due to vocal effort or invisible sounding objects. It also fails to generate impact sound synchronized with the conditional example.

capable of stylizing them to emulate soundscapes in streets, viewing platforms, and the interior or exterior of a train, even though it is originally trained solely on speech.

5 Conclusion

In this paper, we proposed the *audio-visual soundscape stylization* task, aiming to restyle speech to resemble the rich soundscapes from unlabeled in-the-wild audio-visual data. We also constructed an egocentric video dataset and proposed a diffusion model-based method for solving this task in a self-supervised manner. Objective and subjective evaluations demonstrate our model's capability to capture the acoustic properties and ambient sounds of conditional examples. We also show the adaptability of our model to other datasets, visual-only conditioning, and non-speech audio. We hope that our work not only contributes to the task itself but also encourages further exploration into how soundscapes shape our perception of the world. We release the code on our project webpage.

Limitations and broader impacts. While our model demonstrates promising results across various scenarios, its performance can be inconsistent. As shown in Figure 7, our model struggles with vocal effort challenges [41], affecting its ability to capture nuances like pitch variations due to speaker-listener distance. Furthermore, when sounding objects are not visually apparent in the conditional clip (e.g., wind sounds), our model may not replicate these sounds accurately. The model also faces difficulties in maintaining audio-visual consistency [6], particularly with nonstationary audio like impact sounds. This highlights the need for model and dataset expansion to enhance scalability. Lastly, while soundscape stylization is useful for content creation such as movie dubbing, it poses a potential risk for creating disinformation videos.

Acknowledgements. We thank Alexei A. Efros, Justin Salamon, Bryan Russell, Hao-Wen Dong, and Ziyang Chen for their helpful discussions and Baihe Huang for proofreading the paper. This work was funded in part by the Society of Hellman Fellowship and Sony Research Award.

References

1. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence (2018) 3
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 609–617 (2017) 3, 8
3. Bau, D., Andonian, A., Cui, A., Park, Y., Jahanian, A., Oliva, A., Torralba, A.: Paint by word. arXiv preprint arXiv:2103.10951 (2021) 3
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 3
5. Chen, C., Gao, R., Calamia, P., Grauman, K.: Visual acoustic matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18858–18868 (2022) 2, 3, 7, 8, 9, 10
6. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Audio-visual synchronisation in the wild. arXiv preprint arXiv:2112.04432 (2021) 14
7. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 3
8. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020) 3
9. Chen, Z., Qian, S., Owens, A.: Sound localization from motion: Jointly learning sound direction and camera rotation. arXiv preprint arXiv:2303.11329 (2023) 3
10. Corporation, B.B.: BBC Sound Effects (2017), available: https://sound-effects.bbcrewind.co.uk/search 7
11. Cramer, A.L., Wu, H.H., Salamon, J., Bello, J.P.: Look, listen, and learn more: Design choices for deep audio embeddings. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3852–3856. IEEE (2019) 8
12. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: Proceedings of the IEEE international conference on computer vision. pp. 2051–2060 (2017) 4
13. Donahue, C., Caillon, A., Roberts, A., Manilow, E., Esling, P., Agostinelli, A., Verzetti, M., Simon, I., Pietquin, O., Zeghidour, N., et al.: Singsong: Generating musical accompaniments from singing. arXiv preprint arXiv:2301.12662 (2023) 6
14. Dong, H., Yu, S., Wu, C., Guo, Y.: Semantic image synthesis via adversarial learning. In: Proceedings of the IEEE international conference on computer vision. pp. 5706–5714 (2017) 3
15. Du, C., Teng, J., Li, T., Liu, Y., Yuan, T., Wang, Y., Yuan, Y., Zhao, H.: On uni-modal feature learning in supervised multi-modal learning. In: International Conference on Machine Learning. pp. 8632–8656. PMLR (2023) 3
16. Du, Y., Chen, Z., Salamon, J., Russell, B., Owens, A.: Conditional generation of audio from video via foley analogies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2426–2436 (2023) 4
17. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: Clap learning audio concepts from natural language supervision. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 6, 7

18. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics (TOG) **37**(4) (2016) 3, 7, 13
19. Ephrat, A., Peleg, S.: Vid2speech: speech reconstruction from silent video. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5095–5099. IEEE (2017) 3
20. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 3
21. Feng, C., Chen, Z., Owens, A.: Self-supervised video forensics by audio-visual anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10491–10503 (2023) 3
22. Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X.: Fsd50k: an open dataset of human-labeled sound events. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 829–852 (2021) 7
23. Gan, C., Huang, D., Chen, P., Tenenbaum, J.B., Torralba, A.: Foley music: Learning to generate music from videos. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 758–775. Springer (2020) 3
24. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 35–53 (2018) 3
25. Gao, R., Grauman, K.: 2.5 d visual sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 324–333 (2019) 3
26. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017) 3, 7
27. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023) 8
28. Grinfeder, E., Lorenzi, C., Haupert, S., Sueur, J.: What do we mean by "soundscape"? a functional description. Frontiers in Ecology and Evolution **10**, 894232 (2022) 1
29. Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., Glass, J.: Jointly discovering visual objects and spoken words from raw sensory input. In: Proceedings of the European conference on computer vision (ECCV). pp. 649–665 (2018) 3
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 6, 11
31. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp). pp. 131–135. IEEE (2017) 8
32. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2. pp. 557–570 (2023) 3

33. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 5
34. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 6
35. Hu, C., Tian, Q., Li, T., Yuping, W., Wang, Y., Zhao, H.: Neural dubber: Dubbing for videos according to scripts. Advances in neural information processing systems **34**, 16582–16595 (2021) 3
36. Huang, P.Y., Sharma, V., Xu, H., Ryali, C., Fan, H., Li, Y., Li, S.W., Ghosh, G., Malik, J., Feichtenhofer, C.: Mavil: Masked audio-video learners (2023) 3
37. Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., Zhao, Z.: Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In: International Conference on Machine Learning (ICML) (2023) 3
38. Huang, S., Li, Q., Anil, C., Bao, X., Oore, S., Grosse, R.B.: Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer. arXiv preprint arXiv:1811.09620 (2018) 3
39. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017) 3
40. Huh, J., Chalk, J., Kazakos, E., Damen, D., Zisserman, A.: Epic-sounds: A large-scale dataset of actions that sound. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 4
41. Hunter, E.J., Cantor-Cutiva, L.C., van Leer, E., Van Mersbergen, M., Nanjundeswaran, C.D., Bottalico, P., Sandage, M.J., Whitling, S.: Toward a consensus description of vocal effort, vocal load, vocal loading, and vocal fatigue. Journal of Speech, Language, and Hearing Research **63**(2), 509–532 (2020) 14
42. Iashin, V., Rahtu, E.: Taming visually guided sound generation. In: The British Machine Vision Conference (BMVC) (2021) 3
43. Inc., A.: Enhance Speech: Remove noise and echo from voice recordings (2023), available: https://podcast.adobe.com/enhance 4
44. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 3
45. Kaneko, T., Kameoka, H.: Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In: 2018 26th European Signal Processing Conference (EUSIPCO). pp. 2100–2104. IEEE (2018) 3
46. Kilgour, K., Zuluaga, M., Roblek, D., Sharifi, M.: Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In: INTERSPEECH. pp. 2350–2354 (2019) 8
47. Kim, C.D., Kim, B., Lee, H., Kim, G.: 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). pp. 119–132 (2019) 7
48. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 6
49. Koepke, A.S., Wiles, O., Moses, Y., Zisserman, A.: Sight to sound: An end-to-end approach for visual piano transcription. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1838–1842. IEEE (2020) 3

50. Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems **33**, 17022–17033 (2020) 7
51. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 2880–2894 (2020) 8
52. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Proceedings of the Advances in Neural Information Processing Systems (2018) 3
53. Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., Adi, Y.: Audiogen: Textually guided audio generation. In: International Conference on Learning Representations (ICLR) (2023) 3
54. Lee, S.H., Roh, W., Byeon, W., Yoon, S.H., Kim, C., Kim, J., Kim, S.: Sound-guided semantic image manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3377–3386 (2022) 3
55. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022) 9, 10
56. Li, T., Lin, Q., Bao, Y., Li, M.: Atss-net: Target speaker separation via attention-based neural network. In: Interspeech. pp. 1411–1415 (2020) 3
57. Li, T., Liu, Y., Hu, C., Zhao, H.: Cvc: Contrastive learning for non-parallel voice conversion. In: Interspeech (2021) 3
58. Li, T., Liu, Y., Owens, A., Zhao, H.: Learning visual styles from audio-visual associations. In: European Conference on Computer Vision. pp. 235–252. Springer (2022) 3
59. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. In: International Conference on Machine Learning (ICML) (2023) 3, 6, 7, 8, 9, 10
60. Lo, C.C., Fu, S.W., Huang, W.C., Wang, X., Yamagishi, J., Tsao, Y., Wang, H.M.: Mosnet: Deep learning based objective assessment for voice conversion. arXiv preprint arXiv:1904.08352 (2019) 8
61. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 8
62. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. arXiv preprint arXiv:2306.17203 (2023) 3
63. McDermott, J.H., Simoncelli, E.P.: Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron **71**(5), 926–940 (2011) 2
64. Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M.D., Zou, Y., Wang, W.: Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. arXiv preprint arXiv:2303.17395 (2023) 9, 10
65. Morgado, P., Vasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: Advances in Neural Information Processing Systems (2018) 3
66. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12475–12486 (2021) 3

67. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (2018) 3
68. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2405–2413 (2016) 3
69. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 801–816. Springer (2016) 3, 4
70. Patrick, M., Huang, P.Y., Misra, I., Metze, F., Vedaldi, A., Asano, Y.M., Henriques, J.F.: Space-time crop & attend: Improving cross-modal video representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10560–10572 (2021) 3
71. Petermann, D., Wichern, G., Wang, Z.Q., Le Roux, J.: The cocktail fork problem: Three-stem audio separation for real-world soundtracks. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 526–530. IEEE (2022) 4, 9, 10
72. Pijanowski, B.C., Villanueva-Rivera, L.J., Dumyahn, S.L., Farina, A., Krause, B.L., Napoletano, B.M., Gage, S.H., Pieretti, N.: Soundscape ecology: the science of sound in the landscape. BioScience **61**(3), 203–216 (2011) 1
73. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: Learning individual speaking styles for accurate lip to speech synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13796–13805 (2020) 3
74. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 6, 7
75. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. pp. 28492–28518. PMLR (2023) 8
76. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221). vol. 2, pp. 749–752. IEEE (2001) 8
77. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 5
78. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016) 8
79. Sheffer, R., Adi, Y.: I hear your true colors: Image guided audio generation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 3
80. Singh, N., Mentch, J., Ng, J., Beveridge, M., Drori, I.: Image2reverb: Cross-modal reverb impulse response synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 286–295 (2021) 2

81. Somayazulu, A., Chen, C., Grauman, K.: Self-supervised visual acoustic matching. Advances in Neural Information Processing Systems **36** (2024) 2, 3, 7
82. Son Chung, J., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6447–6456 (2017) 13
83. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 6, 8
84. Steinmetz, C.J., Bryan, N.J., Reiss, J.D.: Style transfer of audio effects with differentiable signal processing. arXiv preprint arXiv:2207.08759 (2022) 3
85. Su, K., Liu, X., Shlizerman, E.: How does it sound? Advances in Neural Information Processing Systems **34**, 29258–29273 (2021) 3
86. Team, S.: Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad (2021) 5, 8
87. Ulyanov, D.: Audio texture synthesis and style transfer (2016), available: https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/ 3
88. Välimäki, V., Parker, J., Savioja, L., Smith, J.O., Abel, J.: More than 50 years of artificial reverberation. In: Audio engineering society conference: 60th international conference: dreams (dereverberation and reverberation of audio, music, and speech). Audio Engineering Society (2016) 2
89. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 6, 7
90. Verma, P., Smith, J.O.: Neural style transfer for audio spectograms. arXiv preprint arXiv:1801.01589 (2018) 3
91. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12695–12705 (2020) 7, 13
92. Wang, Y., Ju, Z., Tan, X., He, L., Wu, Z., Bian, J., Zhao, S.: Audit: Audio editing by following instructions with latent diffusion models. arXiv preprint arXiv:2304.00830 (2023) 3
93. Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., Yu, D.: Diffsound: Discrete diffusion model for text-to-sound generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023) 3
94. Yang, F., Ma, C., Zhang, J., Zhu, J., Yuan, W., Owens, A.: Touch and go: Learning from human-collected vision and touch. In: Neural Information Processing Systems (NeurIPS) - Datasets and Benchmarks Track (2022) 3
95. Yang, F., Zhang, J., Owens, A.: Generating visual scenes from touch. International Conference on Computer Vision (ICCV) (2023) 3
96. Yang, K., Russell, B., Salamon, J.: Telling left from right: Learning spatial correspondence of sight and sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9932–9941 (2020) 3
97. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1735–1744 (2019) 3
98. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 570–586 (2018) 3

99. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Visual to sound: Generating natural sound for videos in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3550–3558 (2018) 3
100. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017) 3