Gotta Hear Them All: Towards Sound Source Aware Audio Generation

Wei Guo[1], **Heng Wang**[1], **Jianbo Ma**[2], **Weidong Cai**[1]
[1]University of Sydney
[2]Dolby Laboratories
wei.guo@sydney.edu.au, heng.wang@sydney.edu.au, jianbo.ma@dolby.com, tom.cai@sydney.edu.au

Abstract

Audio synthesis has broad applications in multimedia. Recent advancements have made it possible to generate relevant audios from inputs describing an audio scene, such as images or texts. However, the immersiveness and expressiveness of the generation are limited. One possible problem is that existing methods solely rely on the global scene and overlook details of local sounding objects (i.e., sound sources). To address this issue, we propose a Sound Source-Aware Audio (SS2A) generator. SS2A is able to locally perceive multimodal sound sources from a scene with visual detection and cross-modality translation. It then contrastively learns a Cross-Modal Sound Source (CMSS) Manifold to semantically disambiguate each source. Finally, we attentively mix their CMSS semantics into a rich audio representation, from which a pretrained audio generator outputs the sound. To model the CMSS manifold, we curate a novel single-sound-source visual-audio dataset VGGS3 from VGGSound. We also design a Sound Source Matching Score to clearly measure localized audio relevance. With the effectiveness of explicit sound source modeling, SS2A achieves state-of-the-art performance in extensive image-to-audio tasks. We also qualitatively demonstrate SS2A's ability to achieve intuitive synthesis control by compositing vision, text, and audio conditions. Furthermore, we show that our sound source modeling can achieve competitive video-to-audio performance with a straightforward temporal aggregation mechanism.

Demo Website — https://SSV2A.github.io/SSV2A-demo/

Introduction

As multimedia consumption surges, generating sound for a silent scene attracts high demands in various industries (Zhao, Xia, and Togneri 2019). The synthesized audio can complement a virtual reality scene (Kern and Ellermeier 2020), create Foley for films and games (Di Donato and McGregor 2024), and sonify visual contents for people with visual impairment (Zhou et al. 2018). By learning from text-audio or visual-audio pairs, recent methods can generate highly relevant audio clips given conditions as texts, images or videos. However, most existing methods (Božić and Horvat 2024) only model the mapping between global visual scene and sound while overlooking local details.

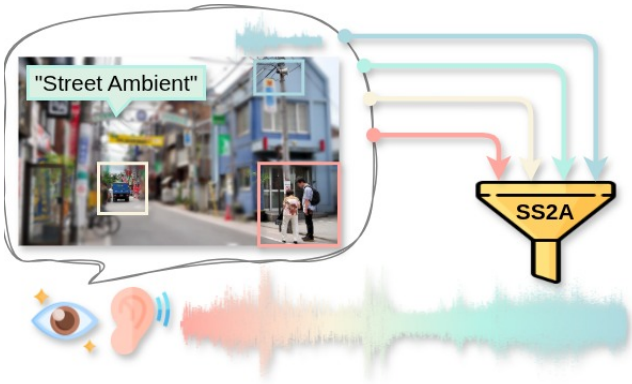

Figure 1: Our SS2A perceives multimodal sound sources in a scene for V2A immersiveness and expressiveness.

In reality, sound is produced and recognized from sounding objects, *i.e.*, **sound sources**, locally present in a soundscape (McAdams 1993). For instance, in a street the sound comes from individual vehicles and passengers as illustrated in Fig. 1. Humans also perceive audio immersiveness and expressiveness from sound source interactions (Gaver 1993). In practice, audio engineers leverage sound sources to intuitively synthesize sounds (Russ 2012).

Can an audio synthesizer utilize **sound source-aware** conditions to obtain better generation quality and control? To answer this question, we present a **S**ound **S**ource-Aware **A**udio (**SS2A**) generator. As image offers sound sources with straightforward curation and composition, we choose it as the primary modality to condition SS2A. We model our system in semantic spaces for learning efficiency and include multimodal conditions from text and audio to boost sound source control. As depicted in Fig. 1, the perception can also come from audio sound source as a loudspeaker and text source as "street ambient". We present SS2A's pipeline in Fig. 2. SS2A first **perceives** multimodal sound source conditions as CLIP (Radford et al. 2021) or CLAP (Elizalde et al. 2023) semantic embeddings with visual detection and cross-modal translation. We then project them to a Cross-Modal Sound Source (CMSS) Manifold to **disambiguate** each source. By disambiguation, we require the CMSS manifold to (1) contrast the source semantics and (2)

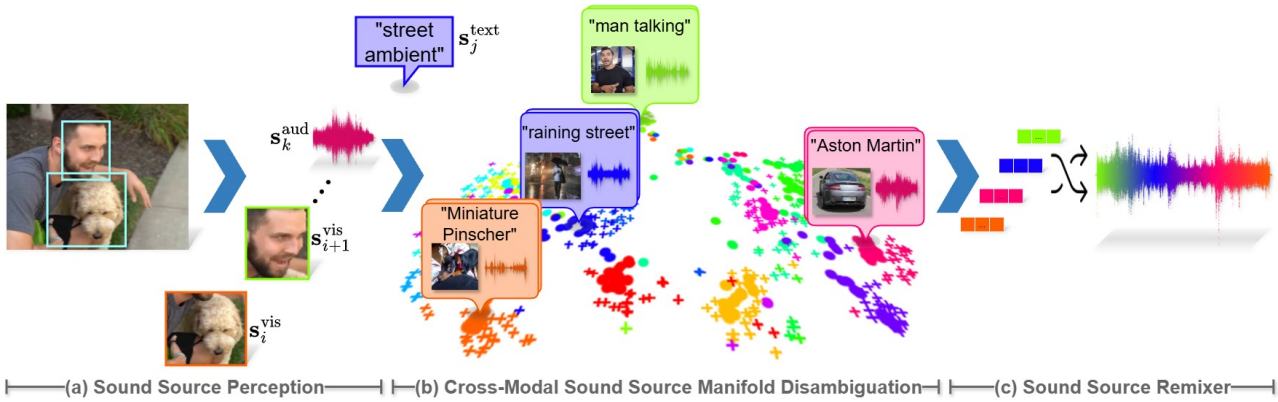

Figure 2: **Pipeline of SS2A.** We perceive sound sources prompted by vision, text, or audio and disambiguate them in the semantically learned CMSS Manifold, which are then mixed to generate an audio clip with immersiveness and expressiveness.

respect the audio characteristics of each sound source. After querying CMSS embeddings of individual sound sources, SS2A learns an attention-based Sound Source Remixer to **mix** them into a CLAP audio embedding with rich sound source information. This representation is passed to a pretrained audio generator, AudioLDM (Liu et al. 2023), to synthesize the output audio waveform.

As the CMSS manifold contrastively learns from single-sound-source (S3) image-audio pairs to disambiguate source semantics, we filter the VGGSound (Chen et al. 2020a) data with visual detection to form a novel dataset, VGGSound Single Source (VGGS3), that contains 106K high-quality S3 image-audio pairs. We also apply a novel Cross-Modal Contrastive Mask Regularization (CCMR) to retain rich CLIP-CLAP semantics by reducing CMSS contrastive influence on similar visual-audio sources with CLIP and CLAP priors. To effectively evaluate generation relevance, we introduce a Sound Source Matching Score (SSMS) to compute the F1 score of overlapping sound source labels on ground-truth and generated samples with an audio classifier.

Both objective and subjective results show that SS2A achieves state-of-the-art image-to-audio synthesis, indicating the benefits of our sound source modeling. We demonstrate SS2A's intuitive generation control by flexibly compositing multimodal sound source prompts from vision, text, and audio to synthesize immersive qualitative samples. We further showcase that our sound source modeling can be straightforwardly extended to competitive video-to-audio synthesis with a temporal aggregation mechanism.

In summary, our contributions are as follows:

- We present a novel framework, SS2A, addressing audio synthesis at the sound-source level. Extensive experiments show that our multimodal sound source modeling leads to state-of-the-art results in image-to-audio generation and competitive video-to-audio performance.
- We explore how sound-source disambiguation can enhance SS2A synthesis with the CMSS manifold, along with a novel CCMR mechanism to guide cross-modal contrastive learning with foundation model priors.
- During manifold training, we curate a high-quality single-sound-source image-audio dataset, VGGS3.
- In evaluating relevance between generated and ground-truth audio signals, we introduce a novel SSMS metric to explicitly match their localized sound sources, proposing a new objective for fine-grained audio generation.
- We showcase multimodal sound source composition, a fresh audio synthesis paradigm that offers intuitive generation control over a wide range of usage scenarios.

Related Works

Vision-to-Audio Generation

Early V2A methods (Owens et al. 2016; Chen et al. 2017; Zhou et al. 2018; Hao, Zhang, and Guan 2018; Chen et al. 2018, 2020b) train a source-specific V2A model on each audio class and cannot generalize to open-domain V2A synthesis. As a precursor, recent SpecVQGAN (Iashin and Rahtu 2021) learns a discrete neural codec (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021) of source-agnostic audio features and autoregressively generates audio codes with a Transformer (Vaswani et al. 2017). Following SpecVQGAN, Im2Wav (Sheffer and Adi 2023) further details its audio codec into low-level and high-level features. MaskVAT (Pascual et al. 2024) leverages a pretrained codec DAC (Kumar et al. 2023) and predicts audio tokens with a Masked Generative Transformer (Chang et al. 2022). Another line of methods employ Diffusion (Ho, Jain, and Abbeel 2020) models. CLIPSonic-IQ (Dong et al. 2023) queries CLIP (Radford et al. 2021) to condition its Diffusion process. Diff-Foley (Luo et al. 2023) contrastively learns a temporally-aligned visual-audio prior to guide video-audio synchronization. Draw-an-Audio (Yang et al. 2024) leverages loudness signal, text caption, and masked video conditions simultaneously. More recently, some methods bridge visual conditions to the prior of a pretrained audio generator for efficient V2A learning. V2A-Mapper (Wang et al. 2024a) maps CLIP embeddings to CLAP (Elizalde et al. 2023) space, from which a pretrained AudioLDM (Liu et al. 2023) model synthesizes the audio signal. V2A-SceneDetector (Yi

and Li 2024) extends V2A-Mapper to multi-scene video with a detection module. Seeing and Hearing (Xing et al. 2024) aligns ImageBind (Girdhar et al. 2023) visual embeddings to AudioLDM. FoleyCrafter (Zhang et al. 2024a) devises a timestamp predictor to enhance synchronization during bridging. Very recently, FRIEREN (Wang et al. 2024b) and MMAudio (Cheng et al. 2025) explore V2A generation with Rectified Flow Matching (Liu, Gong, and Liu 2023). MultiFoley (Chen et al. 2025) employs a diffusion transformer to jointly map multimodal conditions to audio. Most existing methods condition on global visual scenes for V2A synthesis. Some recent works (Li, Zhao, and Yuan 2024) (Li et al. 2024) leverage pixel-level conditions for V2A synthesis, partially describing visual sounding objects. In reality, human perceive object-level sound sources across modalities and time (McAdams 1993). Such a sound source-aware V2A generator remains uninvestigated.

Contrastive Cross-Modal Alignment

Contrastive representation learning (Hadsell, Chopra, and LeCun 2006) has significantly advanced cross-modal representation alignment. CLIP (Radford et al. 2021) aligns text and image modalities by learning from text-image pairs. Many aforementioned V2A methods (Sheffer and Adi 2023; Pascual et al. 2024; Dong et al. 2023; Zhang et al. 2024a; Wang et al. 2024a) benefit from its semantically rich visual representations. Similarly, CLAP (Elizalde et al. 2023) learns from text-audio pairs and is used extensively in V2A generation (Luo et al. 2023; Xing et al. 2024; Wang et al. 2024a; Yang et al. 2024; Liu et al. 2023, 2024). Aside from modality alignment, Diff-Foley (Luo et al. 2023) shows that it is possible to respect temporal alignment in a contrastive visual-audio representation to benefit synchronization. However, the semantic and temporal features in this representation are entangled. In this work, we focus on learning a contrastive manifold for static sound sources, decoupling their semantics from temporal interference.

Method

Approximating an audio distribution $Q(\mathbf{A}|\mathbf{a})$, the audio generator AudioLDM (Liu et al. 2023) generates audio signals $\mathbf{A}$ from CLAP (Elizalde et al. 2023) audio semantics $\mathbf{a}$. For learning efficiency, we employ a pretrained Q and synthesize $\mathbf{a}$ instead of $\mathbf{A}$. Conditioned on multimodal sound sources, our objective is to learn a conditional distribution:

$$P\left(\mathbf{a} \mid \left\{\mathbf{s}_i^{\text{vis}}\right\}, \left\{\mathbf{s}_j^{\text{text}}\right\}, \left\{\mathbf{s}_k^{\text{aud}}\right\}\right), \tag{1}$$

where $\left\{\mathbf{s}_i^{\text{vis}}\right\}$, $\left\{\mathbf{s}_j^{\text{text}}\right\}$, and $\left\{\mathbf{s}_k^{\text{aud}}\right\}$ denote respectively the semantic embedding sets of I visual sound sources, J text sources and K audio sources encoded with CLIP (Radford et al. 2021) or CLAP. We term the acquisition of these semantics as Sound Source Perception in Fig. 2 (a).

The most straightforward way to approximate Eq. (1) is to train a standalone model that maps the perceived CLIP-CLAP semantics directly to $\mathbf{a}$. However, two CLIP features **ambiguate** this direct learning: (1) the CLIP image space models global visual context rather than contrasting individual objects, and (2) CLIP learns only from text-image data, which lacks awareness of the sources' audio traits. As an efficient solution, we learn a Cross-Modal Sound Source (CMSS) manifold as illustrated in Fig. 2 (b) to project the CLIP-CLAP embeddings to a joint semantic space where the local sound sources are **disambiguated**.

Finally, we attentively mix the CMSS embeddings together in Fig. 2 (c) to generate $\mathbf{a}$. This stage involves an attention-based Sound Source Remixer module.

Sound Source Perception

Recall Eq. (1). To extract $\left\{\mathbf{s}_i^{\text{vis}}\right\}$ from a global visual cue when no manual sound-source annotation is available, we pass each image through a visual detector and crop out the detected regions with predicted bounding boxes. These image regions are then embedded by CLIP. To obtain $\left\{\mathbf{s}_j^{\text{text}}\right\}$, we translate the CLIP text embeddings of text prompts to CLIP image space with a pretrained DALL·E-2 Prior (Ramesh et al. 2022) model to mitigate the visual-text domain gap (Liang et al. 2022) and ease downstream disambiguation. For $\left\{\mathbf{s}_k^{\text{aud}}\right\}$, we pass the audio prompts through CLAP to get embeddings.

Cross-Modal Sound Source Manifold

We contrastively learn the CMSS manifold from single-sound-source visual-audio pairs to project the perceived sound source semantics to a joint semantic space for disambiguation, as shown in Fig. 3 (a). The CMSS manifold naturally accommodates the multimodality of our perceptions due to the bridging of CLIP and CLAP.

Manifold Learning. We formulate two CMSS manifold projections $\upsilon(\cdot)$ and $\phi(\cdot)$ as:

$$\mathbf{e}_{\text{CLIP}} = \upsilon(\mathbf{v}), \ \mathbf{e}_{\text{CLAP}} = \phi(\mathbf{a}), \tag{2}$$

given a single-source visual-audio pair as $(\mathbf{V}, \mathbf{A})$ and its CLIP-CLAP embeddings as $(\mathbf{v}, \mathbf{a})$. $\mathbf{e}$ denotes the CMSS embedding. The projectors optimize a contrastive loss to attract visual-audio embeddings from the same sound-source pair and repel those from different sources. Following the symmetric contrastive guidance of CLAP (Elizalde et al. 2023), this objective can be formulated for a batch of N pairs as:

$$\mathcal{L}_c = \frac{\ell_{\text{CLIP}}(\mathbf{C}) + \ell_{\text{CLAP}}(\mathbf{C})}{2}, \tag{3}$$

where $\ell_{\text{CLIP}}(\mathbf{C}) = -\frac{1}{N}\sum_{i=0}^{N} \log diag\left(softmax(\mathbf{C})\right)$ penalizes off-diagonal similarities in similarity entries $\mathbf{C}_{ij} = \tau * \left[\mathbf{e}^i_{\text{CLIP}} \cdot (\mathbf{e}^j_{\text{CLAP}})^\top\right]$. ℓ_{CLAP} follows ℓ_{CLIP} but swaps $\mathbf{e}_{\text{CLIP}}$ and $\mathbf{e}_{\text{CLAP}}$ in $\mathbf{C}_{ij}$. τ is a learned temperature parameter.

We define an auxiliary reconstruction $\chi(\cdot)$ to map the CMSS embeddings back to CLAP space, assisting their alignment with audio semantics. The reconstruction objective is designated for each visual-audio pair as:

$$\mathcal{L}_r = \frac{\|1 - sim(\mathbf{a},\ \chi(\mathbf{e}_{\text{CLAP}}))\| + \|1 - sim(\mathbf{a},\ \chi(\mathbf{e}_{\text{CLIP}}))\|}{2}, \tag{4}$$

where $sim(\cdot,\cdot)$ computes the cosine similarity.

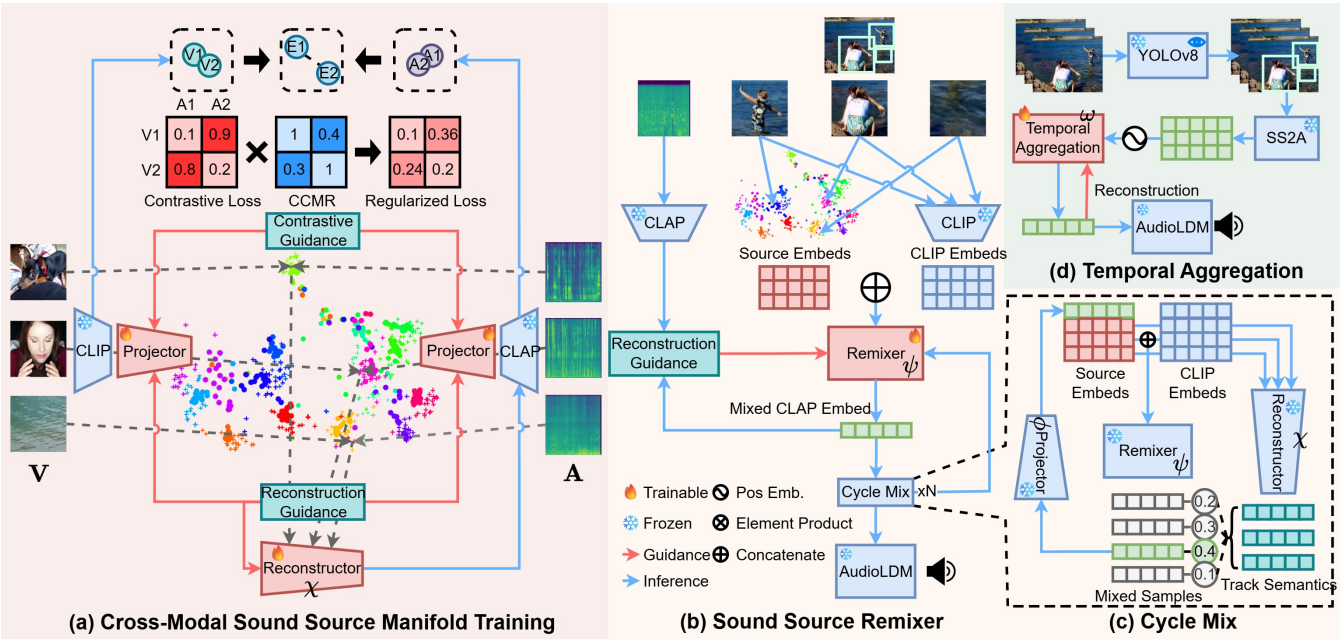

Figure 3: **Detailed Schematics of SS2A Modules. (a)** We learn two projectors to map the CLIP-CLAP embeddings of single-source visual-audio pairs to a joint semantic space with contrastive guidance, forming our CMSS manifold. An auxiliary CLAP reconstruction encodes audio semantics into this manifold. **(b)** The Sound Source Remixer attends to the CMSS embeddings concatenated with their CLIP semantics, generating a single CLAP audio representation which is passed to AudioLDM. **(c)** We reuse the CMSS reconstructor to generate source-wise "track semantics" in CLAP space and refine the Remixer samples iteratively. **(d)** We train an additional Temporal Aggregation (TA) module to attend to positionally embedded SS2A generations across video frames and enhance visual-audio synchronization.

We model $\upsilon(\cdot)$, $\phi(\cdot)$, and $\chi(\cdot)$ variationally with the reparameterization trick and add a Kullback-Leibler (K-L) divergence regularization term to each against the standard normal distribution as $\mathcal{L}_{kl}$. The final objective is then:

$$\mathcal{L}_{\text{fold}} = \mathcal{L}_c + \mathcal{L}_r + \lambda_1 \mathcal{L}_{kl}, \tag{5}$$

where λ_1 is a weight hyperparameter and $\mathcal{L}_{kl}$ is the summed K-L loss. During training, we model all three modules $\upsilon(\cdot)$, $\phi(\cdot)$, and $\chi(\cdot)$ with residually connected MLPs and alternatively optimize the projectors and generator with $\mathcal{L}_{\text{fold}}$.

Cross-Modal Contrastive Mask Regularization. To avoid the loss of rich semantics from CLIP and CLAP due to small training data, we employ a Cross-Modal Contrastive Mask Regularization (CCMR) mechanism to weaken the contrastive guidance $\mathcal{L}_c$ defined in Eq. (3) for similar cross-pair audio-visual samples. For each batch, we compute a CLIP-CLIP similarity matrix $\mathbf{M}_{\text{CLIP}}$ and a CLAP-CLAP similarity matrix $\mathbf{M}_{\text{CLAP}}$ per entry as:

$$\mathbf{M}^{ij}_{\text{CLIP}} = sim(\mathbf{v}_i, \mathbf{v}_j), \ \mathbf{M}^{ij}_{\text{CLAP}} = sim(\mathbf{a}_i, \mathbf{a}_j). \tag{6}$$

The CCMR mask $\mathbf{M}$ is then computed per entry as:

$$\mathbf{M}_{ij} = e^{-\alpha * \left(clamp\left(\mathbf{M}^{ij}_{\text{CLIP}} * \mathbf{M}^{ij}_{\text{CLAP}}\right)\right)^{\alpha}}, \tag{7}$$

where $clamp(\cdot)$ restricts the mask entry to be within $[0, 1]$. This is a stretched exponential decay that grows smaller when both $\mathbf{M}^{ij}_{\text{CLIP}}$ and $\mathbf{M}^{ij}_{\text{CLAP}}$ increase. The hyperparameter α controls the decay curvature and steepness. We apply $\mathbf{M}$ to the original contrastive similarity matrix $\mathbf{C}$ with an element-wise multiplication as $\mathbf{C}^*_{ij} = \mathbf{C}_{ij} * \mathbf{M}_{ij}$.

Data Curation and Training. We filter visual-audio pairs from VGGSound (Chen et al. 2020a) training set with a visual detection pipeline and obtain 106K single-sound-source visual-audio pairs as a novel dataset VGGSound Single Source (VGGS3). We term the VGGS3 pairs *curated pairs*. Additionally, we translate the single-source text-audio pairs from LAION-630K (Wu et al. 2023) to visual-audio pairs with a pretrained DALL·E-2 Prior (Ramesh et al. 2022) model. We term these pairs *translated pairs*. A Mean-Teacher (Tarvainen and Valpola 2017) paradigm trains the CMSS modules with these pairs. Please refer to Supplementary Section 2 for our data curation and training details.

Sound Source Remixer

We employ a Sound Source Remixer function $\psi(\cdot)$ to mix the embeddings $\{\mathbf{e}_m\}$ queried from the CMSS manifold in Fig. 3 (b), generating a CLAP audio representation with rich sound source semantics as $\mathbf{a}_{\text{mix}}$. To leverage all the semantical features helpful for this task, we concatenate each $\mathbf{e}$ with its CLIP embedding $\mathbf{v}$. Specifically, given a set of M sound sources, we formulate f_{mix} as:

$$\psi(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M) = \mathbf{a}_{\text{mix}}, \tag{8}$$

Method	VGGSound				Subjective Relevance		Subjective Fidelity	
	V-FAD↓	C-FAD↓	CS↑	SSMS↑	MOS↑	Std.	MOS↑	Std.
S&H	10.929	79.221	5.941	1.886	-	-	-	-
S&H-Text	5.087	31.461	8.690	2.646	2.358	0.845	2.460	0.748
Im2Wav	5.855	24.303	10.965	3.105	2.595	0.736	2.273	0.757
V2A-Mapper	**0.946**	**5.516**	<u>11.521</u>	<u>3.164</u>	<u>3.063</u>	1.024	<u>2.693</u>	1.210
SS2A(Ours)	<u>1.150</u>	<u>6.716</u>	**12.947**	**3.793**	**4.080**	0.527	**4.098**	0.459

Table 1: **General image to audio tests.** The VGGSound and subjective tests, without source annotation, generalize single-source and multi-source synthesis scenarios. The first and second places are **bolded** and <u>underlined</u>, respectively.

where $\mathbf{x}_i = concat(\mathbf{e}_m, \mathbf{v}_m)$ is the concatenated token for the m-th source. We model $\psi(\cdot)$ variationally to make it generative. The optimization objective is designed as:

$$\mathcal{L}_{\text{mix}} = \|1 - sim(\mathbf{a}, \mathbf{a}_{\text{mix}})\| + \lambda_2 \mathcal{L}_{kl}, \tag{9}$$

where $\mathcal{L}_{kl}$ is the KL divergence from standard normal distribution, and λ_2 is a weight hyperparameter.

We model $\psi(\cdot)$ with a stack of self-attention layers and learn it from visual-audio pairs in VGGSound. The visual sources are perceived from each video's central frame following our aforementioned perception method. Each token sequence $\{\mathbf{x}_m\}$ is padded to a fixed length of $M = 64$. To enhance generation diversity, a Classifier-free Guidance (Ho and Salimans 2021) is applied during training by randomly zeroing out tokens. We replace the classic attention with Efficient Attention (Shen et al. 2021) and detail this architecture in Supplementary Section 2.3. During inference, we set $\mathbf{v} = \mathbf{0}$ for sound source conditions from audio modality.

Cycle Mix. We can also obtain a CLAP embedding $\mathbf{a}_{\text{src}} = \chi(\mathbf{e})$ for each sound source through the CMSS manifold's reconstructor. $\mathbf{a}_{\text{src}}$ can be regarded as a set of source-wise audio semantics generated by our method. As one of our objectives for $\mathbf{a}_{\text{mix}}$ is to have high relevance to each sound source, $\{\mathbf{a}^m_{\text{src}}\}$ are recycled to iteratively guide the generation of $\mathbf{a}_{\text{mix}}$. This mechanism, termed Cycle Mix, is illustrated in Fig. 3 (c) and Algorithm 1 in Supplementary Section 2.3.

Temporal Aggregation. So far, the Sound Source Remixer learns an image-to-audio task. Following V2A-Mapper (Wang et al. 2024a), we adapt it to the video-to-audio task with a downstream Temporal Aggregation (TA) function $\omega(\cdot)$ depicted in Fig. 3 (d). Instead of averaging the frame-wise semantics, we learn a nonlinear $\omega(\cdot)$. We evenly extract 64 frames along time from one video and generate a CLAP embedding for each of them. Each embedding is then positionally embedded with its timestamp. $\omega(\cdot)$ learns to fuse these embeddings into a temporally-aligned CLAP audio representation $\mathbf{a}$ with the following loss:

$$\mathcal{L}_{\text{ta}} = \|1 - sim\left(\mathbf{a}, \omega\left(pos\left(\mathbf{a}^1_{\text{gen}}, \cdots, \mathbf{a}^{64}_{\text{gen}}\right), t\right)\right)\|, \tag{10}$$

where $\mathbf{a}_{\text{gen}}$ denotes the SS2A generated CLAP embeddings and $pos(\cdot, t)$ is the positional embedding function. The architecture of TA is a stack of self-attention layers.

Experiments and Results

Experimental Setup

Please see Supplementary Section 2 for SS2A's implementation details along with the architecture designs.

Datasets. We train our teacher CMMS manifold modules on the VGG Sound Source (VGG-SS) (Chen et al. 2021) dataset. The student modules learn from (1) VGG-SS and (2) curated and translated visual-audio pairs. Since VGG-SS does not have an official train-test split, we randomly sample 4.5K pairs from it for training and form a test set with the remaining 500 pairs. We train the Sound Source Remixer modules following the provided train-test split on VGGSound (Chen et al. 2020a), which contains 19K pairs across 310 audio categories. For image to audio tasks, we test on the VGGSound test set excluding VGG-SS entries, generating 10288 samples. This test does not differentiate single-source and multi-source generation scenarios as VGGSound has no source annotations. For source-annotated tests that clearly split these scenarios, we focus on VGG-SS which contains 38 multi-source pairs ($2\tilde{1}0$ sources each) and 455 single-source pairs. We also test on two out-of-distribution sets MUSIC (Zhao et al. 2018) and ImageHear (Sheffer and Adi 2023) to show SS2A's generalization capability. MUSIC contains 140 pairs with duet musical instrument performance, and 1034 pairs with solo instrument. ImageHear has 101 single-source images from 30 visual classes. We generate 10-second audio samples for all tests.

Objective Metrics. We measure generation quality objectively from two perspectives: fidelity and relevance. For generation fidelity, we adopt the Fréchet Audio Distance (FAD) (Roblek et al. 2019) with an open-source implementation (Tan 2024) to obtain two metrics, V-FAD and C-FAD, respectively from VGGish (Roblek et al. 2019) and CLAP (Elizalde et al. 2023) models. FAD measures the closeness of ground-truth and generated audio feature distributions. A low FAD score reflects high generation fidelity. For generation relevance, we adopt the CLIP-Score (CS) which maps an audio's CLAP embedding to the CLIP image space with a Wav2CLIP (Wu et al. 2022) model to compare its similarity with the paired image. For multi-source image-audio pairs in VGG-SS, we average CS between each sound source image and the paired audio. We compute CS on global images in other tests. A high CS represents high generation relevance.

Matching Score. We observe that the CS relevance comparison, by mapping audio features to image domain, causes

	Method	VGG-SS				MUSIC				ImageHear
		V-FAD↓	C-FAD↓	CS↑	SSMS↑	V-FAD↓	C-FAD↓	CS↑	SSMS↑	CS↑
Single-Source	GroundTruth	0	0.171	13.199	10	0	0	13.906	10	-
	Oracle	1.400	9.983	12.071	5.752	6.430	25.422	12.861	7.777	-
	S&H	16.015	90.656	5.901	1.903	49.045	156.898	4.126	1.421	3.417
	S&H-Text	7.118	37.899	9.761	3.685	25.081	77.218	10.259	5.635	7.401
	Im2Wav	7.573	29.213	11.011	4.451	26.344	57.596	8.374	6.214	10.758
	RAM+ALDM	6.532	30.461	9.199	2.714	23.681	63.810	7.795	3.421	8.765
	V2A-Mapper	**1.666**	**13.583**	11.842	4.488	**7.245**	27.657	12.901	6.288	12.689
	SS2A (Ours)	2.815	15.150	**12.215**	**4.936**	8.075	**25.390**	**13.859**	**7.330**	**13.930**
Multi-Source	GroundTruth	0	0.793	12.344	10	0	0	13.009	10	-
	Oracle	4.356	31.569	11.840	6.447	1.492	34.295	11.658	6.300	-
	S&H	21.447	121.371	6.594	2.568	27.661	175.708	3.979	0.986	-
	S&H-Text	12.678	81.944	9.573	4.026	9.887	105.529	9.149	5.223	-
	Im2Wav	12.915	64.648	11.309	5.132	12.055	81.321	6.426	5.357	-
	RAM+ALDM	14.820	76.406	9.009	3.026	12.985	92.316	8.892	4.261	-
	V2A-Mapper	10.228	59.660	11.331	4.684	4.490	48.665	11.126	4.907	-
	SS2A (Ours)	**6.810**	**46.933**	**11.744**	**5.973**	**3.387**	**31.115**	**12.951**	**6.000**	-

Table 2: **Source-annotated image to audio tests.** These datasets have source annotations to differentiate single-source and multi-source generation scenarios. Only CS is available for ImageHear as it lacks ground-truth pairing audio with each image.

loss of audio information. As a result, our method often outperforms Oracle AudioLDM generations in CS scoring from Tab. 2. We propose a novel metric, Sound Source Matching Score (SSMS), that adopts an audio classifier BEATs (Chen et al. 2023) to respectively predict N localized sound source labels for ground-truth and generated audios. We regard intersected labels from both sets as true positives, the difference of ground-truth against generation as false negatives, and the reverse difference as false positives. SSMS is computed as the F1 score of these statistics. We set $N = 10$ throughout experiments and show that SSMS distinguishes generation relevance more clearly than CS.

Subjective Metrics. Following recent works (Wang et al. 2024a; Zhang et al. 2024a), we conduct a subjective listening test with 20 human evaluators. We randomly sample 40 central video frames from AudioSet Strong (Hershey et al. 2021) and AVSBench (Zhou et al. 2022), generating 10-second audio clips with each image-to-audio method. The participants are asked to rate 20 of them for fidelity without visual cues. They then rate 20 samples for relevance given the visual conditions. We collect the ratings on a 5-point scale and compute the Mean Opinion Score (MOS) (Sector 1996) to measure generation fidelity and relevance. Please see Supplementary Section 6 for the evaluation setup.

Baseline Evaluations

We compare our generator with three image-to-audio methods: V2A-Mapper (Wang et al. 2024a), Seeing and Hearing (S&H) (Xing et al. 2024), and Im2Wav (Sheffer and Adi 2023). Additionally, we employ RAM (Zhang et al. 2024b) to generate image tags and pass them to GPT-4 (Achiam et al. 2023) for text captions, which are fed to AudioLDM to generate audio. We call this cascaded baseline RAM+ALDM. We qualitatively demonstrate how cascaded methods are inferior to SS2A in Supplementary Section 7.

For video-to-audio tasks, we compare with Diff-Foley (Luo et al. 2023), Frieren (Wang et al. 2024b), MultiFoley (Chen et al. 2025), and MMAudio (Cheng et al. 2025). Some baselines require different visual conditions. For fairness, we modify some methods following Supplementary Section 1 but still keep their original results.

Objective Results. As illustrated in Tab. 1 and Tab. 2, our method achieves superior performance in most objective metrics for both in-distribution and out-of-distribution tests. For single-source generation, we outperform baselines in generation relevance and stay in top 2 for generation fidelity. For multi-source generation, SS2A is superior in all metrics. Surprisingly, SS2A achieves a higher CS in generation relevance than the Oracle baseline, which is assumed to have optimal performance for V2A methods involving AudioLDM. This effect is no longer observed in SSMS, demonstrating our new metric's superiority in comparing audio generation relevance. Even S&H-Text has seen generated text captions, SS2A still surpasses it in both fidelity and relevance.

SS2A performs competitively in video-to-audio tasks with the TA extension as shown in Supplementary Section 5.1, showing that our sound source modeling can also benefit video-to-audio synthesis with a straightforward temporal feature integration.

Subjective Results. In Tab. 1, our method outperforms baselines significantly in human-evaluated generation fidelity and relevance. We choose to test S&H-Text instead of S&H to obtain the best generation performance Seeing and Hearing can achieve, even though it sees extra text captions.

Ablation Study

We conduct several ablation experiments to consolidate our claims in the Method section. We also provide an analysis on the learned CMSS manifold space and more ablations in Supplementary sections 3 and 4.

CMSS	CLIP	Single-Source Generation				Multi-Source Generation			
		V-FAD↓	C-FAD↓	CS↑	SSMS↑	V-FAD↓	C-FAD↓	CS↑	SSMS↑
w/o	w/	39.622	122.127	3.987	1.385	34.378	119.692	5.574	1.579
w/	w/o	17.949	96.045	6.049	1.213	7.689	48.776	11.156	5.553
w/	w/	**2.815**	**15.150**	**12.215**	**4.936**	**6.810**	**46.933**	**11.744**	**5.973**

Table 3: **Ablation of Sound Source Remixer conditions.** We achieve best performance with both CMSS and CLIP semantics.

α	Single-Source Generation				Multi-Source Generation			
	V-FAD↓	C-FAD↓	CS↑	SSMS↑	V-FAD↓	C-FAD↓	CS↑	SSMS↑
0	13.612	73.849	5.838	1.149	17.053	98.747	5.580	1.342
0.35	**2.815**	**15.150**	**12.215**	**4.936**	**6.810**	**46.933**	**11.744**	**5.973**
0.65	2.877	16.194	11.860	4.356	9.788	61.565	11.397	4.658
1	3.323	16.740	11.299	4.075	10.098	60.810	11.585	4.237

Table 4: **Ablation of CCMR.** We achieve the best performance with $\alpha = 0.35$, which is used throughout other experiments.

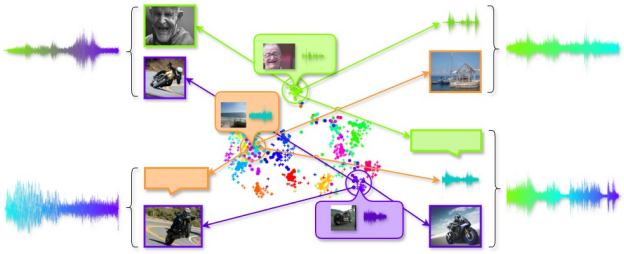

Figure 4: **Multimodal Sound Source Composition scenarios.** Our method can flexibly composite sound sources across visual, text, and audio modalities to guide V2A generation.

Effect of CMSS Manifold. SS2A could learn to perform the V2A task without CMSS disambiguation. In order to prove the benefits of this disambiguation, we perturb the same Sound Source Remixer model with three different generation conditions: without CLIP embeddings, without CMSS embeddings, and with both embeddings. We train them on the same VGGSound data and evaluate the results with VGG-SS tests in Tab. 3. A significant performance drop is observed in both generation fidelity and relevance when the CMSS conditioning is suppressed. This ablation confirms that CMSS disambiguation benefits our V2A task.

Effect of CCMR. Recall that α controls CCMR's behavior in Eq. (7). When $\alpha = 0$, the mask becomes identity and CCMR is stifled. We train the same CMSS manifold modules under four settings of α and conduct VGG-SS tests. Tab. 4 shows that with CCMR, we can enrich the CMSS semantics to benefit downstream generation. However, setting α to higher values degrades generation quality.

Multimodal Sound Source Composition

Since SS2A accepts sound source prompts as vision, text, and audio, we can intuitively control its generation by (1) editing specific sound sources and (2) compositing sources across modalities. We term this novel generation control scheme Multimodal Sound Source Composition. We show four visually-related composition scenarios in Fig. 4. The composition results are best experienced via our website.

Visual Composition. SS2A can generate realistic audio by composing visual sound sources. The result respects the supplied sources to render a convincing audio scene. For instance, we can synthesize a "motorbike riders laughing" audio from pictures of a motorbike and a laughing man.

Visual-Text Composition. SS2A can further control the V2A generation with textual semantics. For example, we can supply a "motorbike" image and obtain a seaside riding audio with the text prompt "seaside".

Visual-Audio Composition. We can achieve a similar style control with audio semantics. For example, we can accompany a "boat pier" image with a "talking" audio to synthesize audio of a busy pier.

Visual-Text-Audio Composition. We can synthesize audio with all three modalities involved. We have successfully produced a "coastline motorcycle racing" audio with a motorcycle image, a "crowd cheering" text, and a "beach" audio.

Conclusion

In this work, we explore learning a sound source-aware audio generator, SS2A, that supports multimodal conditioning. By explicitly modeling the source disambiguation process with a contrastive cross-modal manifold on single-source visual-audio pairs, we are able to significantly boost our method's generation fidelity and relevance. Consequently, SS2A achieves state-of-the-art image-to-audio performance in both objective and subjective evaluations. With a simple temporal aggregation mechanism, SS2A also achieves competitive performance in video-to-audio synthesis. Moreover, we demonstrate the intuitive control of our generator in composition experiments of vision, text, and audio sound sources. During the learning of our manifold, we curate a new single-sound-source visual-audio dataset VGGS3. Additionally, we contribute a novel Sound Source Matching Score that measures fine-grained audio-audio relevance with sound source detection. As SS2A is a fresh exploration, we discuss its limitations in Supplementary Section 8.

References

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

Božić, M.; and Horvat, M. 2024. A survey of deep learning audio generation methods. *arXiv preprint arXiv:2406.00146*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. MaskGIT: Masked generative image transformer. In *CVPR*, 11315–11325.

Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021. Localizing Visual Sounds the Hard Way. In *CVPR*, 16867–16876.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020a. VGGSound: A Large-Scale Audio-Visual Dataset. In *ICASSP*, 721–725.

Chen, K.; Zhang, C.; Fang, C.; Wang, Z.; Bui, T.; and Nevatia, R. 2018. Visually indicated sound generation by perceptually optimized classification. In *ECCV Workshop*.

Chen, L.; Srivastava, S.; Duan, Z.; and Xu, C. 2017. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 349–357.

Chen, P.; Zhang, Y.; Tan, M.; Xiao, H.; Huang, D.; and Gan, C. 2020b. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 29: 8292–8302.

Chen, S.; Wu, Y.; Wang, C.; Liu, S.; Tompkins, D.; Chen, Z.; Che, W.; Yu, X.; and Wei, F. 2023. BEATs: audio pre-training with acoustic tokenizers. In *ICML*, 5178–5193.

Chen, Z.; Seetharaman, P.; Russell, B.; Nieto, O.; Bourgin, D.; Owens, A.; and Salamon, J. 2025. Video-guided foley sound generation with multimodal controls. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18770–18781.

Cheng, H. K.; Ishii, M.; Hayakawa, A.; Shibuya, T.; Schwing, A.; and Mitsufuji, Y. 2025. Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*.

Di Donato, B.; and McGregor, I. 2024. The digital Foley: what Foley artists say about using audio synthesis. In *Audio Engineering Society Conference: AES 2024 International Audio for Games Conference*. Audio Engineering Society.

Dong, H.-W.; Liu, X.; Pons, J.; Bhattacharya, G.; Pascual, S.; Serrà, J.; Berg-Kirkpatrick, T.; and McAuley, J. 2023. CLIPSonic: Text-to-audio synthesis with unlabeled videos and pretrained language-vision models. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1–5.

Elizalde, B.; Deshmukh, S.; Al Ismail, M.; and Wang, H. 2023. CLAP Learning Audio Concepts from Natural Language Supervision. In *ICASSP*, 1–5.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *CVPR*, 12873–12883.

Gaver, W. W. 1993. An Ecological Approach to Auditory Event Perception. *Ecological Psychology*, 5(1): 1–29.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. ImageBind: One embedding space to bind them all. In *CVPR*, 15180–15190.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 1735–1742.

Hao, W.; Zhang, Z.; and Guan, H. 2018. CMCGAN: A uniform framework for cross-modal visual-audio mutual generation. In *AAAI*, volume 32.

Hershey, S.; Ellis, D. P.; Fonseca, E.; Jansen, A.; Liu, C.; Moore, R. C.; and Plakal, M. 2021. The benefit of temporally-strong labels in audio event classification. In *ICASSP*, 366–370.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*, 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS Workshop*.

Iashin, V.; and Rahtu, E. 2021. Taming Visually Guided Sound Generation. In *BMVC*.

Kern, A. C.; and Ellermeier, W. 2020. Audio in VR: Effects of a soundscape and movement-triggered step sounds on presence. *Frontiers in Robotics and AI*, 7: 20.

Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2023. High-fidelity audio compression with improved RVQGAN. In *NeurIPS*, 27980–27993.

Li, T.; Huang, B.; Zhuang, X.; Jia, D.; Chen, J.; Wang, Y.; Anumanchipalli, G.; Chen, Z.; and Wang, Y. 2024. Object-Aware Audio-Visual Sound Generation. *OpenReview*.

Li, Z.; Zhao, B.; and Yuan, Y. 2024. Cyclic Learning for Binaural Audio Generation and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26669–26678.

Liang, W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. 2022. Mind the gap: understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 17612–17625.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *ICML*, 21450–21474.

Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2024. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2871–2883.

Liu, X.; Gong, C.; and Liu, Q. 2023. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*.

Luo, S.; Yan, C.; Hu, C.; and Zhao, H. 2023. Diff-Foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 48855–48876.

McAdams, S. 1993. Recognition of sound sources and events. *Thinking in sound: The cognitive psychology of human audition*, 146–198.

Owens, A.; Isola, P.; McDermott, J.; Torralba, A.; Adelson, E. H.; and Freeman, W. T. 2016. Visually indicated sounds. In *CVPR*, 2405–2413.

Pascual, S.; Yeh, C.; Tsiamas, I.; and Serrà, J. 2024. Masked Generative Video-to-Audio Transformers with Enhanced Synchronicity. *arXiv preprint arXiv:2407.10387*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 8748–8763.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Roblek, D.; Kilgour, K.; Sharifi, M.; and Zuluaga, M. 2019. Fréchet Audio Distance: A Reference-free Metric for Evaluating Music Enhancement Algorithms. In *Proc. Interspeech*, 2350–2354.

Russ, M. 2012. *Sound synthesis and sampling*. Routledge.

Sector, I. T. U. T. S. 1996. *Methods for subjective determination of transmission quality*. International Telecommunication Union.

Sheffer, R.; and Adi, Y. 2023. I hear your true colors: Image guided audio generation. In *ICASSP*, 1–5.

Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; and Li, H. 2021. Efficient attention: Attention with linear complexities. In *WACV*, 3531–3539.

Tan, H. 2024. https://github.com/gudgud96/frechet-audio-distance. Website.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 1195–1204.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *NeurIPS*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 6000–6010.

Wang, H.; Ma, J.; Pascual, S.; Cartwright, R.; and Cai, W. 2024a. V2A-Mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *AAAI*, volume 38, 15492–15501.

Wang, Y.; Guo, W.; Huang, R.; Huang, J.; Wang, Z.; You, F.; Li, R.; and Zhao, Z. 2024b. FRIEREN: Efficient Video-to-Audio Generation with Rectified Flow Matching. In *NeurIPS*.

Wu, H.-H.; Seetharaman, P.; Kumar, K.; and Bello, J. P. 2022. Wav2CLIP: Learning Robust Audio Representations From CLIP. In *ICASSP*, 4563–4567.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 1–5.

Xing, Y.; He, Y.; Tian, Z.; Wang, X.; and Chen, Q. 2024. Seeing and Hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 7151–7161.

Yang, Q.; Mao, B.; Wang, Z.; Nie, X.; Gao, P.; Guo, Y.; Zhen, C.; Yan, P.; and Xiang, S. 2024. Draw an Audio: Leveraging Multi-Instruction for Video-to-Audio Synthesis. *arXiv preprint arXiv:2409.06135*.

Yi, M.; and Li, M. 2024. Efficient video to audio mapper with visual scene detection. *arXiv preprint arXiv:2409.09823*.

Zhang, Y.; Gu, Y.; Zeng, Y.; Xing, Z.; Wang, Y.; Wu, Z.; and Chen, K. 2024a. FoleyCrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*.

Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; et al. 2024b. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1724–1732.

Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; and Torralba, A. 2018. The sound of pixels. In *ECCV*, 570–586.

Zhao, Y.; Xia, X.; and Togneri, R. 2019. Applications of deep learning to audio generation. *IEEE Circuits and Systems Magazine*, 19(4): 19–38.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio–visual segmentation. In *ECCV*, 386–403.

Zhou, Y.; Wang, Z.; Fang, C.; Bui, T.; and Berg, T. L. 2018. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 3550–3558.