Hear What Matters! Text-conditioned Selective Video-to-Audio Generation

Junwon Lee[1,†] Juhan Nam[1,2] Jiyoung Lee[3*]

[1]Graduate School of AI and [2]Graduate School of Cultural Technology, KAIST
[3]Division of AI and Software, Ewha Womans University

{james39,juhan.nam}@kaist.ac.kr, lee.jiyoung@ewha.ac.kr

https://jnwnlee.github.io/selva-demo/

Abstract

This work introduces a new task, text-conditioned selective video-to-audio (V2A) generation, which produces only the user-intended sound from a multi-object video. This capability is especially crucial in multimedia production, where audio tracks are handled individually for each sound source for precise editing, mixing, and creative control. We propose **SELVA***, a novel text-conditioned V2A model that treats the text prompt as an explicit selector to distinctly extract prompt-relevant sound-source visual features from the video encoder. To suppress text-irrelevant activations with efficient video encoder finetuning, the proposed supplementary tokens promote cross-attention to yield robust semantic and temporal grounding.* **SELVA** *further employs an autonomous video-mixing scheme in a self-supervised manner to overcome the lack of mono audio track supervision. We evaluate* **SELVA** *on* VGG-MONOAUDIO*, a curated benchmark of clean single-source videos for such a task. Extensive experiments and ablations consistently verify its effectiveness across audio quality, semantic alignment, and temporal synchronization.*

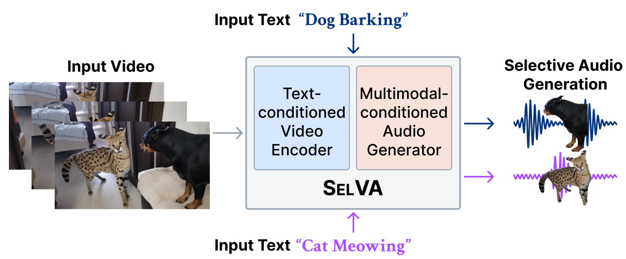


Figure 1. **SELVA** turns text prompts into precise selectors of sound sources within a video. The intent-focused video feature conditions the generator to synthesize only the user-specified sound source (*e.g.*, 'cat meowing' vs. 'dog barking').

1. Introduction

In a bustling café, you can effortlessly tune into a friend's laughter amid the chatter, or pick out the sound of a violin from an entire orchestra. This effortless segregation of sounds, achieved through *auditory scene analysis*, is a hallmark of human perception [3]. At the core of this process lies selective attention, which enables us to focus on a specific sound source while filtering out irrelevant noise. Such an attention-driven mechanism allows humans to extract what truly matters from a rich and noisy world.

Recent advances in neural audio generation have enabled realistic sound synthesis from text descriptions or visual scenes in film and game post-production, known as Foley [50]. Video-to-audio (V2A) models [7, 9, 32, 43, 49] now generate temporally coherent, context-aware audio directly from visual content. However, they typically produce a single holistic soundtrack at a time. This limitation stands in sharp contrast to the sound production [42, 50], where sound designers do not sonify every visible object. Instead, they build scenes by layering individually crafted tracks, selectively including or suppressing sound elements, and then mixing and mastering them, to achieve precise control. However, with current V2A models, even minor omissions in the output require re-synthesizing the entire audio, hindering practical usability.

In this work, we tackle the **selective sound generation** problem: generating only the target sound that aligns with a user's intention, given multimodal cues such as video and text. However, the presence of multiple co-occurring sound sources in real-world videos makes it inherently difficult to isolate the target sound without explicit supervision. A few works [45, 60] have attempted to localize sound sources using off-the-shelf segmentation models, where visual prompts such as points or masks are used to specify the target object. Yet, such approaches inherently operate on

† Work partly done during an internship at NAVER AI Lab.
* Corresponding author.

isolated spatial regions. Therefore, they struggle to handle broader visual context, including environmental or diffuse sounds (*e.g*., rainfall or wind), which cannot be localized to discrete visual boundaries. In addition, reliance on large segmentation networks increases substantial computational overhead. In contrast, we propose a new formulation of text-conditioned selective video-to-audio generation, where the target sound source is specified purely through a text prompt. This removes the dependency on explicit spatial prompts while enabling more flexible and semantically expressive control over sound generation.

We propose a novel **SELVA** method, consisting of two main modules: (1) a text-conditioned video encoder, and (2) a multimodal-conditioned selective audio generator. Unlike previous works [9, 32] that freeze visual-specific encoders, **SELVA** efficiently trains the video encoder to use text prompts as explicit selectors of audible semantics. To achieve a precise cross-modal grounding, we implement a cross-attention block in the video encoder with the proposed learnable token `[SUP]`. Inspired by the selective attention mechanism in human perception, `[SUP]` mitigates high-norm artifacts [1, 14] in which models tend to highlight irrelevant tokens in attention blocks due to the spurious correlation. To train **SELVA** without explicitly source-separated groundtruth audios, we further employ an autonomous video-mixing strategy in a self-supervised manner, where two videos are spatially concatenated as input and the audio from one of them is used as the target.

To assess performance on such a novel task, we introduce VGG-MONOAUDIO, a new evaluation benchmark comprising videos with a clear visual sound source corresponding to a mono audio track. Experimental results demonstrate that **SELVA** achieves state-of-the-art (SoTA) performance on VGG-MONOAUDIO, showing robustness in terms of audio quality, semantic alignment, and temporal alignment. Our contributions are summarized:

- **SELVA** is the first text-conditioned selective video-to-audio framework that relies solely on text prompts, without spatial guidance.
- A learnable supplementary token mitigates spurious cross-modal correlations while empowering intent-relevant cues, and the proposed video-mixing scheme enables selective learning without costly supervision.
- Experiments on new VGG-MONOAUDIO demonstrate the effectiveness of **SELVA**, showing superior audio quality, semantic fidelity, and temporal synchronization performance over existing SoTAs.

2. Related Work

Cross-modal neural audio generation has been widely explored due to its applicability in multimedia content production. Text-to-audio (T2A) aims to generate audio from an input text prompt, which usually describes the global semantics such as sound sources and their nuanced timbre (*e.g*., 'drill buzzing harshly'). The common baseline is first to extract a text embedding from a pretrained text encoder such as CLAP [16, 65] and T5 [10], then use it as a condition of generative models, including diffusion [19, 28, 47], auto-regressive modeling [38], and flow-matching [29]. While text prompts offer intuitive semantic control, they inherently lack the ability to convey temporal dynamics of intensity or harmonics in audio [11, 20, 24]. Meanwhile, video-to-audio (V2A) [7, 9, 15, 30, 49, 53] resolves this issue by generating audio in synchrony with video. Such synchronization entails two complementary goals: semantic and temporal alignment. As a spatiotemporal modality, video conveys rich cues about sounding objects, including appearance, spatial location, and dynamic motion. In practice, current V2A frameworks remain strongly dependent on pretrained visual encoders [2, 31, 49, 54, 62] to provide the conditioning representations.

Recently, some works [32, 43, 67] have leveraged the capability of pretrained T2A models for V2A generation to reduce the training cost and ensure controllability. Most approaches [9, 48, 53, 63] naively hypothesize that complementary relations of video and text conditions, producing high-fidelity audio. Text prompts complement video embeddings by supplying semantic cues that visual encoders often miss (*e.g*., visual ambiguity such as occlusion of sounding objects caused by camera work) [27, 48]. For example, ReWaS [32] and Video-Foley [43] rely on text to control the semantics of sound, while Multifoley [7] leveraged text to change the sound timbre. VinTAGe [39] generates both on-screen sound from visual cues and off-screen sound from textual cues. However, existing works do not use text to specify *which* sound sources should be heard. Instead, text serves merely as an auxiliary cue, not to modulate the given visual information. In this paper, we unlock the potential of text prompts by repositioning them as a direct modulator of video embeddings for controllable V2A.

Selective sound generation has only recently begun to emerge for professional multimedia production, where models synthesize audio exclusively for target sound sources. Hayakawa *et al*. [25] proposed an iterative, trackwise approach with sequential generation, where sounds produced in previous steps are excluded from the current one. This is accomplished using negative audio guidance that steers the flow-matching process to avoid regenerating audio from prior stages. While the motivation is related to ours, their method heavily relies on the limited separation capability of the pretrained V2A model, especially at the first generation stage. Otherwise, some works [45, 60] have utilized visual region-level cues, such as segmentation masks produced by pretrained models (*e.g*., SAM2 [55]), for object-focused sound generation. However, those approaches have notable limitations in that they necessi-

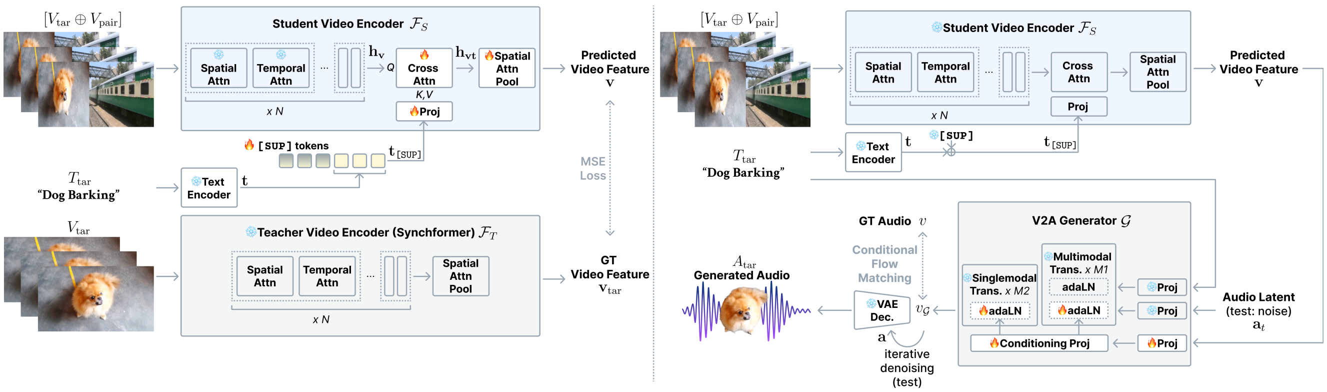


Figure 2. The overall training pipeline of SELVA. We learn a text-conditioned video encoder with a teacher-student distillation manner (left; first stage), and train an audio generator that conditions on text and isolated visual cues for the sound source (right; second stage). Learnable layers are marked with 🔥, while frozen layers are marked with ❄️.

tate the integration of computationally expensive pretrained segmentation models [8, 55], which have often struggled with occluded objects or incorrectly identifying non-object sounding sources (*e.g.*, rain drop, wind blowing). To overcome those limitations, our **SELVA** is the first to introduce a text prompt for describing the target sound source within the input video for robust selective sound generation.

3. Method

3.1. Motivation and problem statement

While existing works [9, 32, 49] have achieved promising results in generating holistic sounds aligned with the input video, they often suffer from low fidelity to the text prompts. Specifically, the model occasionally produces undesired outputs, *i.e.*, non-target objects' audio that appear in the video but are not specified by the text prompt. This limitation arises mainly because most approaches directly feed video features extracted from a frozen visual encoder—typically pretrained for general recognition tasks—into the generation pipeline. Such visual features tend to be noisy and entangled, containing both irrelevant visual cues alongside sound-related semantics. As a result, selectively generating only the intended sound remains challenging. This hinders users from creating harmonious audio in real-world scenarios [25, 42, 50]. For example, a professional audio creator often needs to synthesize a soundtrack with various elements such as speech, music, and sound effects under separate controllable conditions. At this juncture, we argue that the text-conditioned video feature grounding could make a huge room for the controllability of V2A.

Given a video V paired with an audio $\bar{A} = \sum_i A_i$ which is a mixture of multiple sound sources, and a text prompts $\{T_i\}$ that describes a i-th specific sound source, **SELVA** aims to generate audio exclusively that corresponds to the text:

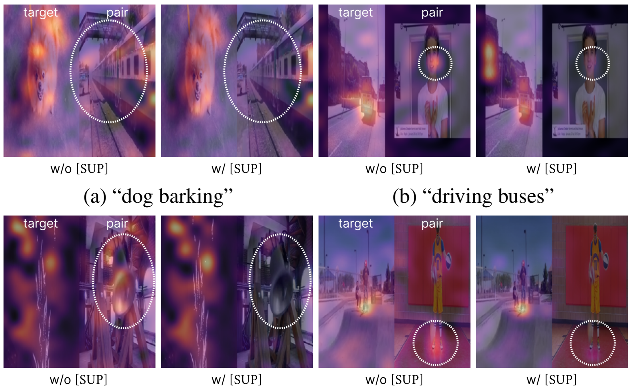
$$A_i = \mathcal{G}(\mathcal{F}(V, \mathbf{t}_i), \mathbf{t}_i) \tag{1}$$

where $\mathcal{F}$ is a visual encoder and $\mathcal{G}$ is a generative model, and a text feature $\mathbf{t}_i = \mathcal{E}(T_i)$ is obtained from a text encoder $\mathcal{E}$. Text prompts role as explicit selectors for video features to offer two main advantages over visual prompts. First, they clearly deliver the target sound source, while visual sound sources often fail to be segmented due to visual occlusions or camera movements. Second, text prompts offer flexible controllability, allowing users to modify the generated sound through simple language edits rather than complex visual manipulations. Such editability supports intuitive control, facilitating practical use in post-production workflows. Note that we employ a parameter-efficient tuning strategy, while most parameters are initialized from prior works and frozen. In what follows, learnable parameters appear in red, and frozen parameters appear in blue.

3.2. Text-guided visual feature generation

Cross-attention block. **SELVA** modulates visual features to encode sound-source relevant information that the text prompt describes. Most V2A models [7, 9, 15, 30, 61, 64, 69] rely on a pretrained vision encoder and, optionally, a text encoder to extract conditioning features. The vision encoders are generally frozen during the training process, serving as visual feature extractors for audio generation. The extracted visual features encompass the global scene context, yet they inherently carry noisy and excessive irrelevant information. Thereby, it impedes the generation of user-intended sound.

Our goal is to produce text-aligned video features by efficiently finetuning the video encoder $\mathcal{F}$. The base encoder is Synchformer [31], which is commonly used in recent V2A models [9, 32, 57, 61]. We introduce two key techniques: (1) A text-guided cross-attention block is inserted after the frozen spatiotemporal attention blocks to modify



(a) “dog barking” (b) “driving buses”

(c) “firework banging” (d) “skateboarding”

Figure 3. Attention visualization for `[eos]` token over auto-mixed frame in the last block without (left) / with (right) `[SUP]` tokens. Each subcaption denotes the corresponding target prompt.

intermediate visual features relevant to the text guidance, with only small extra parameters. Text features, obtained by pretrained text encoder $\mathcal{E}$ (*e.g.*, FLAN-T5-Base [10]), are employed as keys and values in the attention process. Formally, given a hidden video embedding after spatial and temporal attention blocks, $\mathbf{h_v}$, and a text embedding $\bar{\mathbf{t}} = \texttt{Proj}(\mathcal{E}(T))$ from the text encoder followed by a projection layer, the cross attention is performed:

$$\mathbf{h_{vt}} = \texttt{Cross-Attn}(Q = \mathbf{h_v}, K = \bar{\mathbf{t}}, V = \bar{\mathbf{t}}). \quad (2)$$

The predicted video feature $\mathbf{v}$ is obtained with a learnable spatial attention pooling layer (`Spatial-Attn-Pool`).

Learnable supplementary tokens. We expect hidden video embeddings to contain semantics that are exclusively aligned with text. However, a simple cross-attention mechanism yields a suboptimal result where the model still suffers from generating sounds corresponding to the motion dynamics of non-target instances. For example, when a dog is barking beside a cat, it incorrectly produces a meowing sound, reflecting semantic confusion between co-occurring sound sources. To mitigate this issue, we introduce a novel technique to attach learnable supplementary tokens `[SUP]` preceding text embeddings. As shown in Fig. 3 (‘w/o `[SUP]`’), such artifacts emerge as semantic patch tokens (particularly those associated with motion dynamics) become high-norm outliers in the attention space. However, as in prior approaches on vision transformers [1, 14], adding extra tokens to the sequence of visual embeddings increases the computational cost across all encoder blocks. Furthermore, we should design to learn intent-focused visual representations that emphasize regions and cues relevant to the user’s specified sound source. To this end, `[SUP]` are simply prepended to text features as follows:

$$\mathbf{t}_{\texttt{[SUP]}} = [\texttt{[SUP]} \oplus \mathbf{t}], \quad (3)$$

where $\oplus$ is a sequence-wise concatenation operation. The cross-attention block uses $\mathbf{t}_{\texttt{[SUP]}}$ to produce intent-modulated video feature $\mathbf{h_{vt}}$ as in Eq. (2). This design enables the model to suppress irrelevant or misleading visual activations while strengthening attention toward regions that correspond to the user-intended sound source, as demonstrated in Fig. 3 (‘w/ `[SUP]`’). As a result, the learned video representation leads to improved selectivity and audiovisual coherence. The detailed input data configuration of the video encoder is in Appendix A.1.

3.3. Selective sound generation

Our generator, $\mathcal{G}$, adopts a multimodal diffusion transformer (MM-DiT) architecture [18, 40]:

$$\mathbf{a} = \mathcal{G}(\mathbf{v}, \mathbf{t}_{\text{tar}}, \mathbf{a}_{\text{tar}}) \quad (4)$$

where $\mathbf{a}_{\text{tar}}$ is the audio latent extracted from a pretrained variational autoencoder (VAE) [35]. It closely follows the pipeline of MMAudio [9], with the only modification being the exclusion of CLIP [54] features from the conditioning inputs. It is worth noting that the generator already possesses sufficient capacity to synthesize audio conditioned on multimodal inputs. Therefore, our contribution does not lie in designing a specialized generator architecture, but rather in enabling selective sound generation through improved conditioning representations. Specifically, the initially sampled noise is transformed to target audio latent $\hat{\mathbf{a}}$ via a flow matching process [46, 59], jointly contextualizing video $\mathbf{v}$ and text semantics $\mathbf{t}$ in MM-DiT. MM-DiT consists of a stack of multimodal and single-modal transformer blocks. Given hidden audio features $\mathbf{h_a}$ and text features $\mathbf{h_t}$, the multimodal blocks compute $\texttt{Self-Attn}(Q, K, V = [\mathbf{h_t}, \mathbf{h_a}])$, while the single-modal blocks compute $\texttt{Self-Attn}(Q, K, V = \mathbf{h_a})$. The adaptive LayerNorm (`adaLN`) layers [52] condition the block-wise hidden state $\mathbf{h} \in \mathbb{R}^{L \times d}$ on the linear-projected video feature $\bar{\mathbf{v}} = W_{\mathbf{v}}\mathbf{v}$. Formally, this operation is defined as:

$$\texttt{adaLN}(\mathbf{h}, \bar{\mathbf{v}}) = \mathbf{1}W_\gamma(\bar{\mathbf{v}}) \cdot \texttt{LN}(\mathbf{h}) + \mathbf{1}W_\beta(\bar{\mathbf{v}}) \quad (5)$$

where W_γ and W_β are the conditioning projection layers and $\mathbf{1} \in \mathbb{R}^{L \times 1}$ is a matrix of ones for broadcasting.

3.4. Training

Video-mixing. As input videos usually comprise multiple sound sources, without explicit annotations for each sound source, it is nontrivial to isolate visual features. To address this, we introduce a self-supervised strategy, motivated by audiovisual separation works [17, 41], but reformulated for selective V2A generation. Concretely, two videos are randomly selected, and horizontally concatenated with a random ratio to make the desired {mixed-video, audio, text} pairs. One of the audio-text pairs is randomly chosen to serve as the target. Formally, an input video V in a mini-batch consists of randomly selected two videos $\{V_{\text{tar}}, V_{\text{pair}}\} \in \mathbb{R}^{H\times W}$:

$$V = [V_{\text{tar}} \in \mathbb{R}^{H\times \lambda W} \oplus V_{\text{pair}} \in \mathbb{R}^{H\times (1-\lambda) W}], \quad (6)$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ is a scaling factor for resizing the video sampled from a beta distribution, and $\oplus$ is a horizontal concatenation operation. Here, V_{tar} serves as a target video to semantically attend, while the paired video V_{pair} becomes a distractor. This scheme encourages robust cross-modal grounding by distinguishing the target visual region without explicit supervision.

Two-stage training. Learning *conditional feature extraction from mixed sources* and *multiple conditioned audio generation* simultaneously is inherently complex, as both modules depend on each other's evolving representations [26]. To ensure efficient and stable optimization, the joint training is organized into a two-stage learning scheme, allowing each module to converge toward a consistent representation before mutual conditioning. In the first stage, the video encoder learns to understand text prompts from output features of the teacher model [66, 68]. In the left part of Fig. 2, while the teacher model $\mathcal{F}_T$ (*i.e*., pretrained Synchformer [31]) takes a single source video V_{tar} to generate a pseudo feature $\mathbf{v}_{\text{tar}}$, the student model $\mathcal{F}_S$ produces a text-guided visual feature from the mixed source. Formally, the video feature $\mathbf{v}$ of student encoder $\mathcal{F}_S$ is learned to minimize the L2-norm regression loss:

$$||\mathcal{F}_S([V_{\text{tar}} \oplus V_{\text{pair}}], \mathbf{t}_{\text{tar}}) - \mathcal{F}_T(V_{\text{tar}})||^2 \quad (7)$$

where $\mathbf{t}_{\text{tar}}$ is the text embeddings corresponding to the target video V_{tar}. This stage updates the student network's parameter for cross-attention and spatial attention pooling layers exclusively. While the teacher model can only take visual inputs, the student extracts a specific representation from text guidance. In other words, our approach uses video features as teaching signals, allowing the model to learn how multimodal interactions can selectively emphasize informative cues while suppressing irrelevant sound sources as noise in the visual representation.

Next, we train a generator $\mathcal{G}$ while keeping the video encoder frozen in the second stage. We start from the MM-DiT in MMAudio [9] as the baseline of the generator. Rather than finetuning the whole parameters, we focus on specific modules that handle video features explicitly, as illustrated in the right part of Fig. 2. Specifically, we finetune two sub-modules only: (1) the initial projection layer of the video feature branch (*i.e*., $W_{\mathbf{v}}$) and (2) the adaptive LayerNorm (adaLN) module of the audio latent branch of multimodal and single-modal transformer blocks (*i.e*., W_γ, W_β). The model is trained with conditional flow-matching (CFM) [46, 59]. Given a noise distribution $q(\mathbf{a}_0) \sim \mathcal{N}(\mathbf{0}, I)$, training data distribution $q(\mathbf{a}_1 = \mathbf{a}_{\text{tar}}, \mathbf{c})$ with input condition features $\mathbf{c} = (\hat{\mathbf{v}}_{\text{tar}} = \mathcal{F}_S(V, \mathbf{t}_{\text{tar}}), \mathbf{t}_{\text{tar}})$, and timestep $t \in [0, 1]$, the CFM objective is formulated as:

$$\mathop{\mathbb{E}}_{t, q(x_0), q(\mathbf{a}_1, \mathbf{c})} ||v(t, \mathbf{c}, \mathbf{a}_t; \mathcal{G}) - v(\mathbf{a}_t|\mathbf{a}_0, \mathbf{a}_1)||^2, \quad (8)$$

where $\mathbf{a}_t = t\mathbf{a}_1 + (1-t)\mathbf{a}_0$ is the flow that generates a velocity v. Implementation details for optimizer settings are noted in Appendix A.3.

4. Experiments

4.1. Setup

Training dataset. SELVA is trained on VGGSound [5], which provides approximately 500 hours of video and 310 unique captions. We utilize these captions as the text prompts for our model. Following the experimental setup of our baseline model [9], we partition the official training data, setting aside 2k samples for validation. The training set includes 179k videos, and the test set is 15k. For both training and inference, all video clips are processed into 8-second segments.

Test benchmark. Evaluating selective sound generation requires clean, source-separated audio with corresponding text descriptions. However, existing in-the-wild datasets such as VGGSound [5] and AudioSet [21] typically provide only a single mixed track and video-level captions, often contaminated by recording noise or off-screen sounds [6, 13, 39]. To address these limitations, we introduce **VGG-MonoAudio**, an evaluation benchmark for selective V2A generation. We curate mono-source clips from UnAV-100 [22] overlapping with VGGSound test set, and filter them automatically and manually with three strict criteria: (1) a single source sounding with minimal background or off-screen noise, (2) the sounding object is clearly visible, (3) the text annotation precisely matches the auditory event. Finally, we obtain a final set of 67 clean, single-source videos spanning 39 unique events (*e.g*., 'dog barking', 'train wheels squealing') across 8 categories: *human, music, vehicle, tool, animal, nature, sport, other*. To construct test samples, we concatenate pairs of these videos side-by-side, each occupying half the frame width. The horizontally combined video $[V_1 \oplus V_2]$ serves as the model input, while the audio $A_{\mathbf{v}_1}$ and text $T_{\mathbf{v}_1}$ from one of the source videos are used as the target. From the 67 curated videos, we generate 1,071 mixed pairs in total, 560 inter-class (videos from different categories) and 511 intra-class (videos from the same category), for quantitative evaluation. Appendix B provides the full list of audio event categories and detailed benchmark statistics. We also include results on the original VGGSound test set for completeness; these are given in Appendix D.2, as this setting is not central to our evaluation.

Baselines. We establish four SoTA baselines for compar-

Model	Audio Quality			Semantic Alignment			Temporal Alignment
	FAD↓	KAD↓	IS↑	KL↓	CLAP↑	IB↑	DeSync↓
Inter-class							
ReWaS [32]	70.4	4.937	6.23	2.57	0.200	0.2454	1.364
VinTAGe [39]	**50.5**	1.309	11.51	**1.69**	0.283	0.2850	1.292
MMAudio-S-16k [9]	56.7	<u>0.874</u>	11.54	2.07	0.270	<u>0.3135</u>	<u>0.802</u>
VOS [8]+MMAudio [9]	60.0	0.878	<u>12.11</u>	1.91	<u>0.291</u>	0.3010	0.991
SELVA	<u>51.7</u>	**0.676**	**13.07**	<u>1.85</u>	**0.292**	**0.3251**	**0.721**
Intra-class							
ReWaS [32]	57.4	3.148	6.29	1.97	0.220	0.2569	1.377
VinTAGe [39]	**37.0**	0.690	<u>9.28</u>	**0.88**	0.277	0.2892	1.304
MMAudio-S-16k [9]	<u>41.5</u>	<u>0.654</u>	9.00	1.09	0.276	<u>0.3248</u>	<u>0.670</u>
VOS [8]+MMAudio [9]	43.4	0.656	8.91	1.11	**0.287**	0.3087	0.904
SELVA	**37.0**	**0.492**	**9.62**	<u>1.04</u>	<u>0.280</u>	**0.3262**	**0.639**

Table 1. Quantitative comparisons with state-of-the-art models on VGG-MONOAUDIO. All methods used text prompts corresponding to the target videos. The **best** scores are shown in bold, and the <u>second-best</u> scores are underlined.

ison. ReWaS [32], VinTAGe [39], and MMAudio [9] are text-conditioned V2A models where text semantically aids video. To implement concurrent segmentation-based approaches [45, 60], we leverage the pretrained video object segmentation (VOS) model [8] to build a 'VOS+MMAudio' system, in which we pass the text and video to obtain video-level object masks. The resulting masked video is then used as the conditional input to the MMAudio.

Metrics. Three main criteria matter for evaluating selective audio generation: audio quality, semantic alignment, and temporal alignment with the target.

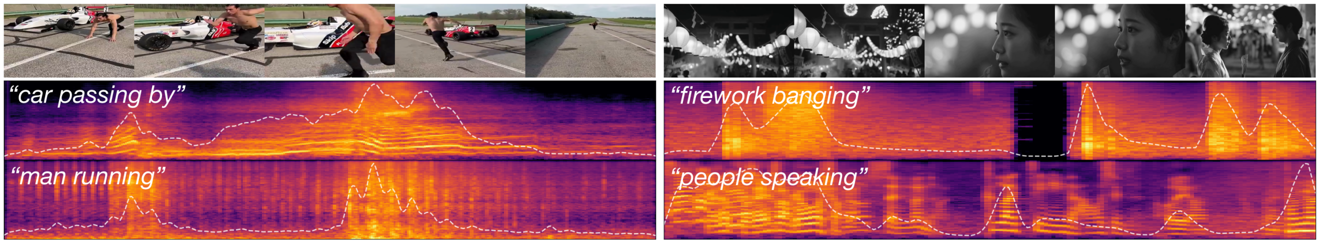
- <u>Audio quality</u> is assessed by Fréchet audio distance (FAD) [34], kernel audio distance (KAD) [12], and inception score (IS) [56] with PANNs[37, 58].
- <u>Semantic alignment</u> is assessed to evaluate prompt fidelity. While CLAP score (CLAP) [65] is used to measure how closely the generated audio aligns with the intended text, imagebind score (IB) [23] measures the alignment between audio and target video. In addition, Kullback-Leibler divergence (KL) with PANNs distribution is employed to evaluate semantic alignment between the generated and groundtruth audio tracks.
- <u>Temporal alignment</u> is assessed by DeSync [9], the average synchronized error (*i.e.*, predicted offset in seconds) between the audio and video. As temporal alignment is crucial for perceptual coherence in V2A, this metric serves as the primary reference in our ablation study.

4.2. Implementation details

During training of **SELVA**, mixing inputs are given within each minibatch with a probability of 0.75, while clipping the mixing ratio λ of the target video to be greater than 0.2. A total of 5 learnable `[SUP]` tokens are prepended to every input text prompt; this number was determined by Tab. D3. We initialize the video encoder $\mathcal{F}_S$ with pretrained Synchformer [31] and the generator $\mathcal{G}$ with MMAudio-small-16kHz weights. Note that we train 19M parameters in $\mathcal{F}_S$ and 22M in $\mathcal{G}$, corresponding to 14% of each model's total parameters, respectively. Following the original setup for classifier-free guidance (CFG) in MMAudio [9], we randomly substitute the video and text features with learned null video and text embedding ($\varnothing_\mathbf{v}$ and $\varnothing_\mathbf{t}$) with a probability of 0.1. In addition, we drop the text feature with an additional probability of 0.5 to enhance the visual fidelity. Inference on the flow matching model is performed using the Euler solver with 25 linear sampling steps. During inference, CFG is applied with a guidance strength of $\gamma = 4.5$.

4.3. Comparison with state-of-the-arts

Quantitative analysis. Table 1 summarizes the quantitative performance of V2A models on VGG-MONOAUDIO. **SELVA** outperforms baselines across all key aspects, including audio quality, semantic alignment, and temporal alignment. Notably, we achieve the best scores in both semantic and temporal audio-video alignment. MMAudio [9], which overlooks text modality, exhibits degraded CLAP scores than **SELVA**, whereas neglecting video modality often results in temporally misaligned results with poor DeSync scores, as seen in ReWaS [32] and VinTAGe [39]. This highlights that training a text-conditioned video encoder in **SELVA** is effective to achieve these dual goals. VOS baseline shows competitive semantic alignment but performs poorly on temporal synchronization. It is primarily due to the inherent limitations of VOS methods, which struggle to accurately localize fast-moving or motion-blurred objects, and vague or complex boundaries (*e.g.*, rain drop). Regarding the two subsets of VGG-MONOAUDIO, models generally achieve better objective scores in the intra-class subset. This happens because the paired non-target video is semantically similar to the target, leading objec-



(a) Spatially co-occurring events.

(b) Multiple events intertwined temporally.

Figure 4. Examples of selective generation with real-world videos. The white dotted curve is the root-mean-squared audio amplitude.

tive metrics to overestimate model performance. Therefore, human perceptual evaluation becomes essential.

Qualitative analysis. Fig. 4 demonstrates that **SELVA** selectively synthesizes target sounds even in complex real-world auditory scenes involving multiple simultaneous or temporally overlapping events. As shown in Fig. 4a, **SELVA** successfully generates distinct sounds for spatially co-occurring sound events such as 'car passing by' and 'man running', demonstrating spatial disentanglement within a shared visual context. In Fig. 4b, **SELVA** succeeds in temporally disentangling intertwined events, producing natural and temporally aligned sounds for 'firework banging' and 'people speaking'. These examples highlight our model's robustness to yield realistic and context-aware audio by capturing the user's intended focus. More qualitative examples are provided in Fig. D8 in Appendix.

Human study. Human listening test assesses the perceptual performance of the models, to complement our automatic metrics. A total of 26 participants rated three criteria scores: *overall audio quality* (AQ), *text-audio alignment* (TA) for semantic relevance, and *audiovisual temporal synchronization* (VA) using a 5-point Likert scale. The evaluation set consists of 16 unique videos: one from each of the 8 sound categories, selected from both the inter-class and intra-class VGG-MONOAUDIO benchmarks. Each video was presented with the corresponding audio by 4 different sources: ground-truth (GT), 'MMAudio-S-16k', 'VOS+MMAudio', and **SELVA**. Fig. 5 reports the mean opinion score (MOS), along with the corresponding 95% confidence interval. The subjective results show strong alignment with the objective evaluation. **SELVA** outperforms both MMAudio and the VOS baselines across all criteria. In terms of audio quality, **SELVA** achieves a comparable performance to GT, whereas the other baselines show noticeably lower scores. For video-audio alignment, **SELVA** also achieves the highest score among the comparable models, with the GT obviously achieving the best score. Notably, VOS baseline scored 3.78 (vs. 4.53 in **SELVA**) in text-audio alignment, even though its CLAP score in Tab. 1 is comparable to ours. This highlights a discrepancy between the objective metric and human perception. It indicates that human listeners are more sensitive to off-screen noises that are loosely aligned with the text prompt (see Appendix D.5).

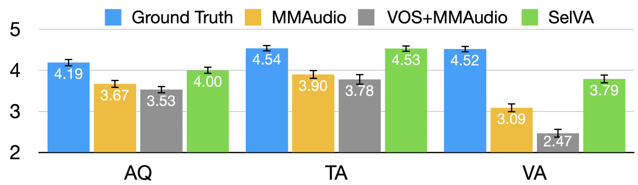


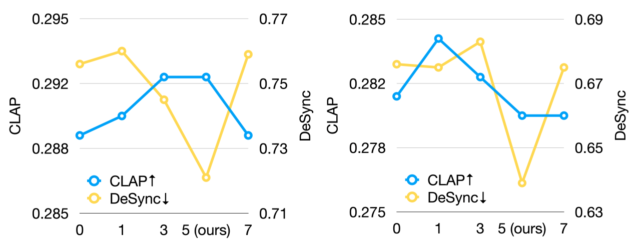
Figure 5. Human study results on VGG-MONOAUDIO. The GT results (*i.e.*, real sound) show oracle performance. **SELVA** outperforms state-of-the-art methods, including MMAudio and VOS baselines.

4.4. Ablation studies

Impact of each training component. Tab. 2 summarizes our ablation studies, which demonstrate the impact of removing each training component: (1) video encoder $\mathcal{F}_S$ finetuning (first stage), (2) V2A generator $\mathcal{G}$ finetuning (second stage), (3) prepended `[SUP]` tokens (used during the first stage), and (4) two-stage training. Finetuning only the V2A generator (while keeping the video encoder frozen) yields marginal gains in audio quality and semantic alignment, but causes a notable degradation in audiovisual temporal synchrony. We observe that the generator tends to develop an undesired shortcut behavior, producing sounds that align text semantics but drift from the actual video events. Conversely, excluding the V2A generator finetuning significantly reduces overall audio quality. The results indicate that finetuning the generator with auto-mixing samples is necessary to obtain optimal performance. The results reveal that removing `[SUP]` tokens especially deteriorates the temporal alignment score. This supports our hypothesis that `[SUP]` tokens facilitate selective generation by refining text-irrelevant spatial attention, while making a negligible sacrifice in audio quality and semantic alignment. Finally, joint training (*i.e.*, optimizing Eq. (7) and Eq. (8) simultaneously) shows notable drops in both semantic and temporal audiovisual alignment scores, indicating that the model fails to maintain coherent cross-modal correspondence. For instance, in the intra-class benchmark, IB (0.3229 vs. 0.3248)

Model	Audio Quality			Semantic Alignment			Temporal Alignment
	FAD↓	KAD↓	IS↑	KL↓	CLAP↑	IB↑	DeSync↓
Inter-class							
SELVA	51.7	0.676	13.07	1.85	0.292	0.3251	**0.721**
– Video Enc. FT	53.8	0.638	13.35	**1.75**	**0.300**	0.3303	0.868
– V2A Gen. FT	56.6	0.721	12.94	1.89	0.293	**0.3309**	0.736
– `[SUP]` tokens	51.4	**0.637**	12.95	1.79	0.289	0.3272	0.756
– two-stage training	**51.3**	0.707	**13.78**	1.81	0.299	0.3138	0.823
Intra-class							
SELVA	37.0	0.492	9.62	1.04	0.280	0.3262	**0.639**
– Video Enc. FT	38.2	**0.423**	10.15	**1.01**	**0.291**	0.3294	0.734
– V2A Gen. FT	39.4	0.553	9.35	1.06	0.281	**0.3300**	0.651
– `[SUP]` tokens	**36.3**	0.485	9.74	**1.01**	0.281	0.3277	0.676
– two-stage training	36.8	0.456	**10.18**	1.00	0.283	0.3229	0.777

Table 2. Ablation on model design variants: without video encoder training, generator training, `[SUP]` tokens, and two-stage training.



(a) Inter-class (b) Intra-class

Figure 6. Ablation on the number of `[SUP]` tokens, determined by balancing semantic and temporal alignment performance.

and DeSync (0.777 vs. 0.670) scores are even worse than the frozen MMAudio baseline. In particular, joint training often substitutes non-target sound events with text-aligned sounds, thereby deteriorating temporal synchronization.

The number of `[SUP]` tokens. Achieving the dual goal of semantic text-audio alignment and temporal video-audio alignment is crucial in selective sound generation. We therefore observe the change in both the CLAP and DeSync scores for different numbers of `[SUP]` tokens. As shown in Fig. 6, we identify a "sweet spot" at 5 tokens, which achieves the best DeSync score while maintaining a comparable CLAP score. It is a consistent observation from prefix-tuning [44], where performance typically degrades when too few tokens fail to convey sufficient conditioning information, or when too many tokens lead to redundancy and overfitting. Full results, including other metrics, are shown in Tab. D3 in Appendix.

5. Limitations and Future Work

We identify three primary directions for future work. First, the model performance is currently limited by the noisiness of the training data in VGGSound [5]. Therefore, more rigorous data filtering or refining the auto-mixing process with cleaner source data could improve the performance. Second, since text labels are typically simple noun-verb conjunctions and lack such descriptive detail, the model's complex text understanding capabilities could be enhanced. This includes fine-grained cross-modal distinction (*e.g.*, separating 'male singing' from 'male burping') and improved attribute controllability (*e.g.*, a dog barking 'aggressively'). Finally, while our method significantly alleviates the sound substitution issue, residual cases remain when the video encoder fails to track a target movement change consistently. We leave a comprehensive full training of the model as future work.

6. Conclusion

We present **SELVA**, text-conditioned V2A model tailored for audio production systems in the real world. **SELVA** efficiently modulates the video encoder to capture the user's textual intent, introducing a few learnable tokens and specialized training schemes. Experimental results show that **SELVA** delivers precise and controllable sound generation on our new benchmark, VGG-MONOAUDIO, significantly outperforming existing methods. These findings highlight **SELVA** as a strong step toward practical, reliable, and fully controllable selective video-to-audio generation.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST)), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00222383, RS-2025-16065706).

References

[1] Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R Glass, and Hilde Kuehne. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In *CVPR*, 2025. 2, 4

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 2

[3] A.S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, 1990. 1

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1

[5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 5, 8, 4

[6] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE TIP*, 2020. 5

[7] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. In *CVPR*, 2025. 1, 2, 3

[8] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023. 3, 6

[9] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, 2025. 1, 2, 3, 4, 5, 6

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 2024. 2, 4, 1

[11] Yoonjin Chung, Junwon Lee, and Juhan Nam. T-foley: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis. In *ICASSP*, 2024. 2

[12] Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon. Kad: No more fad! an effective and efficient evaluation metric for audio generation. *arXiv preprint arXiv:2502.15602*, 2025. 6, 3, 4

[13] Chenye Cui, Zhou Zhao, Yi Ren, Jinglin Liu, Rongjie Huang, Feiyang Chen, Zhefeng Wang, Baoxing Huai, and Fei Wu. Varietysound: Timbre-controllable video to sound generation via unsupervised information disentanglement. In *ICASSP*, 2023. 5

[14] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 2, 4

[15] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *CVPR*, 2023. 2, 3

[16] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP*, 2023. 2

[17] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM TOG*, 2018. 4

[18] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 4, 2

[19] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024. 2

[20] Hugo Flores García, Oriol Nieto, Justin Salamon, Bryan Pardo, and Prem Seetharaman. Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations. In *ICASSP*, 2025. 2

[21] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 5, 1

[22] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *CVPR*, 2023. 5, 2

[23] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 6, 3

[24] Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio generation with multiple conditional diffusion model. In *AAAI*, 2024. 2

[25] Akio Hayakawa, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. Step-by-step video-to-audio synthesis via negative audio guidance. *arXiv preprint arXiv:2506.20995*, 2025. 2, 3

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 5

[27] Feizhen Huang, Yu Wu, Yutian Lin, and Bo Du. Spotlighting partially visible cinematic language for video-to-audio generation via self-distillation. *arXiv preprint arXiv:2507.02271*, 2025. 2

[28] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *ICML*, 2023. 2

[29] Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024. 2

[30] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *BMCV*, 2021. 2, 3

[31] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP*, 2024. 2, 3, 5, 6, 1, 4

[32] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *AAAI*, 2025. 1, 2, 3, 6, 4, 5

[33] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *CVPR*, 2024. 2

[34] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019. 6, 3

[35] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 3

[37] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *TASLP*, 2020. 6, 4

[38] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022. 2

[39] Saksham Singh Kushwaha and Yapeng Tian. Vintage: Joint video and text conditioning for holistic audio generation. In *CVPR*, 2025. 2, 5, 6, 3, 4

[40] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux,2024`. 4, 2

[41] Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In *CVPR*, 2021. 4

[42] Junwon Lee, Modan Tailleur, Laurie M Heller, Keunwoo Choi, Mathieu Lagrange, Brian McFee, Keisuke Imoto, and Yuki Okamoto. Challenge on sound scene synthesis: Evaluating text-to-audio generation. In *NeurIPS Workshop*, 2024. 1, 3

[43] Junwon Lee, Jaekwon Im, Dabin Kim, and Juhan Nam. Video-foley: Two-stage video-to-sound generation via temporal event condition for foley sound. *TASLP*, 2025. 1, 2

[44] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL-IJCNLP*, 2021. 8

[45] Yingshan Liang, Keyu Fan, Zhicheng Du, Yiran Wang, Qingyang Shi, Xinyu Zhang, Jiasheng Lu, and Peiwu Qin. Hear-your-click: Interactive object-specific video-to-audio generation. *arXiv preprint arXiv:2507.04959*, 2025. 1, 2, 6, 3

[46] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 4, 5

[47] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *ICML*, 2023. 2

[48] Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see-video to audio generation through text. In *NeurIPS*, 2024. 2

[49] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2023. 1, 2, 3

[50] Sangshin Oh, Minsung Kang, Hyeongi Moon, Keunwoo Choi, and Ben Sangbae Chon. A demand-driven perspective on generative audio ai. *arXiv preprint arXiv:2307.04292*, 2023. 1, 3

[51] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *NeurIPS*, 2021. 1

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *CVPR*, 2023. 4

[53] A Polyak, A Zohar, A Brown, A Tjandra, A Sinha, A Lee, A Vyas, B Shi, CY Ma, CY Chuang, et al. Movie gen: A cast of media foundation models, 2025. *arXiv preprint arXiv:2410.13720*, 2024. 2

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 4, 1

[55] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2025. 2, 3

[56] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 6, 3

[57] Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025. 3

[58] Modan Tailleur, Junwon Lee, Mathieu Lagrange, Keunwoo Choi, Laurie M Heller, Keisuke Imoto, and Yuki Okamoto. Correlation of fréchet audio distance with human perception of environmental audio is embedding dependent. In *EUSIPCO*, 2024. 6, 4

[59] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *TMLR*, 2024. 4, 5

[60] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Saganet: Video object segmentation-aware audio generation. In *GCPR*, 2025. 1, 2, 6, 3

[61] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *ICASSP*, 2025. 3

[62] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal segment net-

works: Towards good practices for deep action recognition. In *ECCV*, 2016. 2

[63] Le Wang, Jun Wang, Chunyu Qiang, Feng Deng, Chen Zhang, Di Zhang, and Kun Gai. Audiogen-omni: A unified multimodal diffusion transformer for video-synchronized audio, speech, and song generation. *arXiv preprint arXiv:2508.00733*, 2025. 2

[64] Xihua Wang, Xin Cheng, Yuyue Wang, Ruihua Song, and Yunfeng Wang. Vaflow: Video-to-audio generation with cross-modality flow matching. In *ICCV*, 2025. 3

[65] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023. 2, 6, 3

[66] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 5

[67] Zhifeng Xie, Shengye Yu, Qile He, and Mengtian Li. Sonicvisionlm: Playing sound with vision language models. In *CVPR*, 2024. 2

[68] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *CVPR*, 2019. 5

[69] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024. 3