Precise Video-to-Audio Generation with Cross-Modal Alignment in Latent Space

Thanh V. T. Tran[1], Ngoc-Son Nguyen[1], Luong Tran[1], Long-Khanh Pham[1], Paarth Neekhara[2], Shehzeen Hussain[2], and Van Nguyen[1]

[1] FPT Software AI Center, Vietnam
[2] NVIDIA Corporation, USA
https://flowley-v2a.github.io
{thanhtvt1,sonnn45,luongtk,khanhpl2,vannth19}@fpt.com
{pneekhara,shehzeenh}@nvidia.com

Abstract. Video-to-audio (V2A) generation aims to synthesize realistic audio that is both semantically consistent with and temporally synchronized to a silent video. Despite recent progress, many methods still rely on multi-stage training, resulting in high computational costs and long runtimes, or transform visual input into text to leverage pretrained text-to-audio models, sacrificing fine-grained temporal cues. To overcome these limitations, we propose Flowley, an end-to-end, single-stage training architecture that produces soundtracks by combining visual features with textual prompts. Crucially, we introduce Progressive Soft-masked Cross-Attention, which embeds audio-visual synchronization directly within its attention mechanism, adding zero additional computational cost compared to standard attention layers. We further observe that existing V2A benchmarks lack sound-oriented descriptive captions, which can potentially degrade the quality of the synthesized audio. To remedy this, we propose SoundCap, a plug-and-play pipeline for creating detailed, sound-aware captions that guide the model. Remarkably, without integrating any pretrained audio-visual alignment modules, Flowley achieves state-of-the-art performance on VGGSound across multiple metrics. Moreover, by incorporating SoundCap, we further exceed the performance of the strongest existing close-sourced methods in terms of audio quality in the zero-shot setting.

Keywords: Video-to-Audio · Cross-attention · Video captioning

1 Introduction

Sound design is the craft of storytelling through sonic composition. A key branch of this field is Foley [46], where sound effects are created in precise synchronization with onscreen action during post-production. These soundscapes are brought to life by a skilled Foley artist working on a purpose-built stage stocked with a wide array of props and materials for generating the required sounds[3].

[3] See how a foley artist performing a scene in our project page.

Recent advances in generative audio modeling have facilitated text-to-audio (T2A) systems [11, 28, 32], allowing designers to generate sound effects directly from written descriptions. Despite accelerating the search for appropriate sounds, they must still manually adjust timing to align audio with on-screen action. This contrasts sharply with traditional Foley work, where artists naturally craft and sync sound effects in real time by interacting physically with props.

To overcome this limitation, video-to-audio (V2A) generation has gained significant attention, dividing into two major directions. The first converts visual features from silent video frames into text, leveraging pretrained T2A models to synthesize sound [51, 57]. While this approach leverages the strengths of pretrained T2A systems, it inevitably discards fine-grained temporal details, which are crucial for film-production sound effects. The second line of work uses multi-stage training: early phases learn to extract or align acoustic cues from video frames, either via dedicated regression networks [18, 47, 52] or contrastive objectives [31], and subsequent stages build on these pretrained modules. While effective, these pipelines incur substantial delays and computational overhead. Furthermore, integrating textual descriptions into the generation workflow, beyond their use in T2A systems, remains under-explored, despite the demonstrated benefits of text guidance in other domains [20, 48, 50]. This gap is partly due to existing V2A datasets [6, 34] lacking rich, sound-focused text annotations.

In this work, we introduce *Flowley*, a flow-based architecture that synthesizes high-fidelity audio from silent video by harnessing multimodal cues. Our framework consists of two interconnected modules: **(1)** multi-stream blocks, which jointly fuse audio-latent, visual, and textual embeddings into rich, cross-modal representations; and **(2)** single-stream blocks, which further refine the audio pathway. To enhance multimodal conditioning within each single-stream block, we incorporate a cross-attention mechanism with textual features and propose a novel ***P**rogressive **S**oft-masked **C**ross-**A**ttention* (PSCA) module that effectively aligns acoustic and visual information. Empirically, we show that this design improves the temporal alignment of synthesized audio, eliminating the need for pretrained synchronization modules or additional pretraining stages. In practice, however, most V2A datasets either lack descriptive annotations or contain low-quality captions, which limits a model's ability to maintain semantic consistency in generated audio. To overcome this, we introduce ***Sound**-aware **Cap**tioner* (SoundCap), a plug-and-play pipeline that leverages pretrained audio-visual large language models (AV-LLMs) to generate detailed, sound-oriented captions as ground truth for training vision-language models (VLMs), thereby enabling robust inference of both on- and off-screen acoustic events. Comprehensive experiments on VGGSound show that Flowley exceeds state-of-the-art (SOTA) methods across a range of metrics. Notably, when augmented with SoundCap, Flowley surpasses Movie Gen Audio in zero-shot audio quality, underscoring the added effectiveness of SoundCap.

2 Related Works

2.1 Video-to-Audio Generation

V2A generation has garnered significant interest because of its potential utility in multimedia production. Early efforts predominantly employed autoregressive models: Zhou *et al.* [58] first demonstrated SampleRNN [33] for direct waveform synthesis from video frames, while SpecVQGAN [17] used a transformer-based autoregressive architecture conditioned on RGB inputs and optical flow. Im2Wav [42] further advanced this line by adopting a dual-resolution transformer framework and leveraging CLIP [39] embeddings to predict VQVAE code indices. However, these autoregressive methods are hampered by slow inference speed, as they must generate audio samples sequentially. To overcome these limitations, diffusion-based approaches have emerged, demonstrating high-quality soundtrack synthesis [37]. Luo *et al.* [31], Wang *et al.* [53], and Zhang *et al.* [56] designed an audio-visual contrastive feature and utilized latent diffusion and flow-based model to predict mel-spectrograms. V2A-Mapper [51] projects CLIP visual embeddings into CLAP space for conditioning AudioLDM [27]. Temporal conditioning has also been explored in SonicVisionLM [54], ReWaS [18], Foley-Crafter [57], STA-V2A [40], and Mel-QCD [52]. However, these frameworks typically presume perfect alignment between text and video and depend on frozen T2A models, which can cause textual cues to dominate visual information. Recently, VinTAGe [23] and MMAudio [7] have investigated joint multimodal conditioning but heavily rely on external pretrained audio-visual alignment modules to ensure synchronization and semantic consistency. In contrast, Flowley achieves SOTA results with a streamlined, single-stage training pipeline that forgoes any external pretrained audio-visual alignment modules.

2.2 Visual-guided Audio Generation with Acoustic-enhanced Captions

Although simple captions can effectively guide audio synthesis, many models struggle with complex, detail-rich prompts, a challenge commonly referred to as "prompt following" [3]. While prompt engineering has been explored in vision generation [3, 5], it remains under-investigated in audio generation, and major audio–visual datasets like AudioSet [13] and VGGSound offer at best minimal textual descriptions. Few works have attempted to bridge this gap: VATT [30] fine-tunes VLMs to predict audio events from video alone using LTU-generated captions as supervision [15], and another approach employs a multi-model pipeline to extract visual and acoustic cues separately before feeding them into an LLM for caption synthesis [45, 55]. However, both strategies share the same flaw of cross-modal inconsistency. For instance, when we replicated the setup of Sound-VECaps [55], the acoustic model frequently confused the rumble of a car engine with a lion's roar. To overcome these limitations, we leverage AV-LLMs to produce detailed, sound-focused captions as unified ground truth, thereby eliminating modality mismatch and effectively capturing both on- and off-screen audio events.

3 Methodology

3.1 Preliminaries

Flow Matching. We use the flow matching (FM) [26, 29] framework to train our model. FM enables sampling from the target data distribution by progressively transforming a sample drawn from a prior distribution, such as a standard Gaussian distribution. During training, given an audio latent x_1, we randomly sample a timestep $t \in [0, 1]$ and a noise vector $x_0 \sim \mathcal{N}(0, 1)$, which are then combined to form a training point x_t. The objective is to train the model to estimate the velocity vector $v_t = \frac{\mathrm{d}}{\mathrm{d}t}x_t$ which teaches it to "move" the sample x_t, guiding the transformation of x_t toward the target x_1. While there are numerous ways to construct x_t, we use a simple linear interpolation:

$$x_t = tx_1 + (1 - t)x_0.$$

This leads to a constant ground-truth velocity:

$$v_t = \frac{\mathrm{d}}{\mathrm{d}t}x_t = x_1 - x_0.$$

Letting θ denote the model parameters and c the multimodal conditioning inputs, the predicted velocity is expressed as $v_\theta(x_t, t, c)$. We train the model by minimizing the MSE between the predicted and ground-truth velocities:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,x_0,x_1,c}||v_\theta(x_t, t, c) - v_t||^2. \tag{1}$$

At inference, we begin by sampling $x_0 \sim \mathcal{N}(0, 1)$, and then solve an ordinary differential equation (ODE) using the model's estimated velocity field to obtain the final output x_1.

Audio encoding. Following previous works [27, 53], for computational efficiency, we downsample audio to 16 kHz and convert it to a mel-spectrogram, which is then encoded into a latent variable $x_1 \in \mathbb{R}^{L_{\mathrm{aud}} \times D_{\mathrm{aud}}}$ by a pretrained variational autoencoder (VAE) [19], where L_{aud} and D_{aud} represent the sequence length and VAE bottleneck dimension, respectively. During inference, the predicted latent is decoded back into a mel-spectrogram, and the waveform is finally reconstructed using BigVGAN [24].

3.2 Flowley's Architecture Overview

To predict the flow field v_θ for a given latent state x_t, Flowley incorporates both visual and textual conditioning. As depicted in Figure 1a, our overall architecture adopts a multi-to-single stream design paradigm [4, 7]. After obtaining representations from pretrained encoders, Flowley first employs N_1 multi-stream blocks [10], which jointly process video, text, and audio latents, and then uses N_2 single-stream blocks to refine the audio latent stream, which serves as the primary stream of the block. Crucially, Flowley introduces an audio-visual alignment-injected mechanism that removes the need for any external dedicated modules, which is an aspect that existing approaches have not yet resolved. We describe each of these components below.

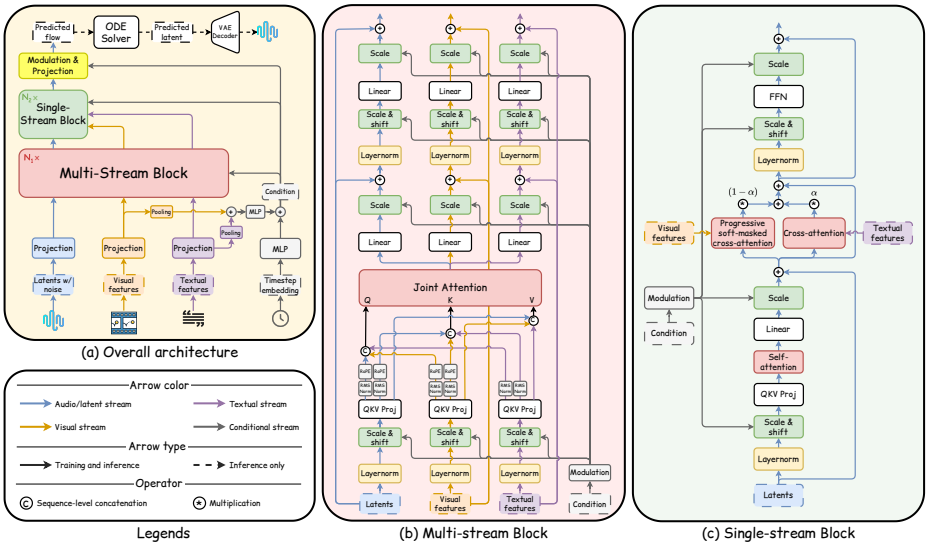


Fig. 1: **(a)** The proposed Flowley framework consists of two core modules. **(b)** First, visual, textual, and audio latent representations are processed together through the multi-stream block. **(c)** Latent features are then passed into the single-stream block, where they undergo weighted cross-attention with the visual and textual streams to estimate the flow field. At inference time, we integrate this learned flow using standard ODE solvers to generate the compressed mel-spectrogram, which is subsequently decoded and vocoded into the final audio waveform.

Representations. We adopt FLAN-T5 [8] as our text encoder to produce embeddings $f_{\text{txt}} \in \mathbb{R}^{L_{\text{txt}} \times D_{\text{txt}}}$, where L_{txt} is the number of tokens and D_{txt} the dimensionality of the textual features. For visual cues, we sample the video at 8 FPS and pass each frame through CLIP [12, 39] to obtain frame-wise representations. The resulting video embedding is denoted $f_{\text{vis}} \in \mathbb{R}^{L_{\text{vis}} \times D_{\text{vis}}}$, where L_{vis} and D_{vis} represent the number of frames and visual feature dimension, respectively. Notably, unlike prior methods that rely on pretrained temporal modules, such as onset detectors [57], energy extractors [18], or audio-visual alignment networks [7, 53], we integrate temporal alignment directly within our model, as described in the Section 3.3. Finally, we combine visual and textual representations with a learned timestep embedding via average pooling and addition to form the global condition $c \in \mathbb{R}^{1 \times D}$, where D is the hidden dimension of Flowley. This condition is injected into the network via adaptive layer normalization (adaLN) layers [36], which apply learned scales and biases. Specifically, given an input $y \in \mathbb{R}^{L \times h}$, each adaLN layer computes:

$$\text{adaLN}(y, c) = \text{LayerNorm}(y) \cdot \mathbf{1}\mathbf{W}_1(c) + \mathbf{1}\mathbf{W}_2(c),$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are two MLPs, and $\mathbf{1} \in \mathbb{R}^{L \times 1}$ is an all-ones vector that broadcasts the scale and bias across the sequence length L.

Multi-stream Block. Our multi-stream block builds on the MM-DiT image generation architecture [10]. As shown in Figure 1b, it enables cross-modal interaction by applying a joint attention mechanism over visual, textual, and audio latents. To ensure stable training, we follow established best practices [60] by integrating QK-Norm and Rotary Positional Embeddings (RoPE) [43] directly into the key-query dot-product attention computation.

Single-stream Block. After passing through N_1 multi-stream blocks, the audio latent is routed into a single-stream network comprising of N_2 blocks. This two-stage design helps decouple multimodal fusion from modality-specific refinement, improving scalability and parameter efficiency. For these blocks, we extend the DiT architecture [35] by augmenting its cross-attention mechanism to incorporate both learned visual and textual embeddings (see Figure 1c). By jointly attending to aligned video and audio frames, we enhance temporal synchronization, while the inclusion of detailed, sound-oriented text improves semantic fidelity within the acoustic latent representation. Notably, we introduce PSCA, a novel mechanism that directly aligns visual and acoustic features on a frame-wise basis within the flow model's latent space. A detailed explanation of PSCA is presented in the next section.

3.3 Progressive Soft-masked Cross-Attention

The original cross-attention mechanism, introduced by Vaswani *et al.* [49], enables the decoder to focus on the most relevant portions of the encoder's outputs:

$$\mathrm{CA}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where d_k is the hidden dimension of each attention head. In our setting, this translates to acoustic features attending to various frames of the visual input. However, this setup introduces a risk: an audio segment may mistakenly attend to unrelated visual frames, leading to temporal misalignment in the generated audio. To mitigate this, we introduce a soft mask M with entries in $[0, 1]$, where values approaching zero impose stronger masking. Specifically, denote the audio query sequence $Q_{\text{aud}} \in \mathbb{R}^{L_{\text{aud}} \times D}$ and the video key/value sequences $K_{\text{vis}}, V_{\text{vis}} \in \mathbb{R}^{L_{\text{vis}} \times D}$, sampled at rates r_a and r_v (with $r_a > r_v$). Because each video frame spans multiple audio frames, we define the video frame index aligned to audio position i as:

$$j_c(i) = \min\left(\left\lfloor \frac{r_v}{r_a} i \right\rfloor, L_{\text{vis}} - 1\right).$$

Let $d_{ij} = |j - j_c(i)|$ be the absolute distance (in video-frame indices) between audio token i and video frame j, we define our base soft-mask $M_{ij}^{(0)}$ as:

$$M_{ij}^{(0)} = \begin{cases} 1, & d_{ij} \le \omega, \\ \mathcal{F}(d_{ij} - \omega), & \omega < d_{ij} \le \omega + \delta, \\ 0, & d_{ij} > \omega + \delta, \end{cases}$$

where ω defines a hard attention window of full weight, and δ indicates the additional number of video frames beyond the hard window over which attention weights are softly decayed from 1 down to 0, since adjacent frames often carry useful contextual or motion cues. Therefore, instead of discarding them entirely, we apply a decay kernel $\mathcal{F} : [0, \delta] \to [0, 1]$ that satisfies $\mathcal{F}(0) = 1$ and $\mathcal{F}(\delta) = 0$. The mathematical expression of $\mathcal{F}$ is as follows:

$$\mathcal{F}(m) = \frac{1}{2}\left[\cos\left(\frac{\pi m}{\delta}\right) + 1\right].$$

Furthermore, to encourage early layers to explore a wider context while later layers focus on precise local evidence, we scale the fade zone by a layer-dependent progressive parameter β:

$$\beta_\ell = 1 - \frac{\ell}{N_2 - 1}, \quad \ell \in \{0, \ldots, N_2 - 1\}.$$

Therefore, the final soft-mask for layer ℓ is:

$$M_{ij}^{(\ell)} = \begin{cases} 1, & d_{ij} \le \omega, \\ \beta_\ell \mathcal{F}(d_{ij} - \omega), & \omega < d_{ij} \le w + \delta, \\ 0, & d_{ij} > \omega + \delta. \end{cases}$$

Figure 2 visualizes the mask values $M_{ij}^{(\ell)}$ as a function of the audio-video offset d_{ij} and network layer ℓ. Within any given block, entries closest to the aligned center frame $j_c(i)$ (i.e., smallest d_{ij}) exhibit the highest mask values, facilitating stronger cross-modal attention, while more distant frames are progressively attenuated. Moreover, as the layer index increases, the overall mask values diminish, converging to zero in the final layer and effectively reducing PSCA to a hard sliding-window. Given the mask formulation, we formally express the PSCA mechanism at layer ℓ as:

$$\mathrm{PSCA}_\ell(Q_{\mathrm{aud}}, K_{\mathrm{vis}}, V_{\mathrm{vis}}) = \mathrm{softmax}\left(\frac{Q_{\mathrm{aud}} K_{\mathrm{vis}}^T}{\sqrt{d_k}} + \log(M^{(\ell)} + \epsilon)\right) V_{\mathrm{vis}},$$

where ϵ is a very small number to avoid $\log(0)$. PSCA unifies *hard* locality, *soft* contextual fading, and a *depth-aware* progression in a single, differentiable mask. Early layers benefit from broader, but still weighted, video context, reducing alignment errors, while deeper layers concentrate capacity on the most tightly aligned frames, improving fine-grained synchronization. Finally, the updated representation $\tilde{x}^{(\ell)}$ of the audio latent stream $x^{(\ell)}$ in ℓ-th block combines text- and vision-conditioned attention via:

$$\tilde{x}^{(\ell)} = x^{(\ell)} + \alpha^{(\ell)} \cdot \mathrm{CA}(Q_{\mathrm{aud}}, K_{\mathrm{txt}}, V_{\mathrm{txt}}) + (1 - \alpha^{(\ell)}) \cdot \mathrm{PSCA}_\ell(Q_{\mathrm{aud}}, K_{\mathrm{vis}}, V_{\mathrm{vis}}),$$

where $\alpha^{(\ell)} \in [0, 1]$ is a learnable parameter determining the contributions from the two cross-attention branches at ℓ-th block.

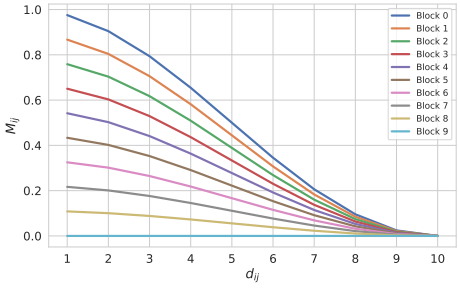


Fig. 2: Mask values within the fade region of the progressive soft-mask M across all $N_2 = 10$ single-stream blocks, using $\omega = 0$ and $\delta = 10$.

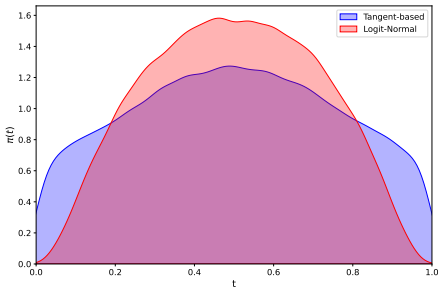


Fig. 3: Distribution of tangent-based and logit-normal schedulers.

3.4 Training and Inference

The noise scheduler plays a critical role in training both flow-based and diffusion-based models. Although the logit-normal distribution [1] is often preferred for its emphasis on intermediate timesteps, it collapses to zero density at the endpoints $t = 0$ and $t = 1$ (Figure 3). To address this, we employ a tangent-based sampling schedule defined by:

$$t = 1 - \frac{1}{\tan\left(\frac{\pi}{2}u\right) + 1},$$

where $u \sim \mathcal{U}[0, 1]$. The resulting distribution is illustrated in Figure 3 and the mathematical derivation is provided in the Supplementary Material. To accelerate the training process, we leverage data-noise alignment [25] which determines in advance the target noise level for each example's latent before noise injection. Furthermore, in addition to the FM loss (Equation (1)), we introduce a velocity direction supervision term using cosine distance:

$$\mathcal{L}_{\text{vel}} = 1 - \text{sim}\big(v_\theta(x_t, t, c), (x_1 - x_0)\big),$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity. Our overall training objective thus becomes $\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda\mathcal{L}_{\text{vel}}$, where λ is the hyperparameter. To support classifier-free guidance (CFG) [16] at inference, we randomly mask either the visual tokens or the textual prompt with a 10% probability during training. With that, our velocity prediction becomes:

$$\tilde{v}_\theta(x_t, t, c) = v_\theta(x_t, t, \varnothing) + s\Big(v_\theta(x_t, t, c) - v_\theta(x_t, t, \varnothing)\Big),$$

where $\varnothing$ denotes empty token and s is the guidance weight.

3.5 Sound-aware Captioner

Having detailed the network design of Flowley, we next address a complementary challenge: data expressiveness. Existing V2A datasets provide only short

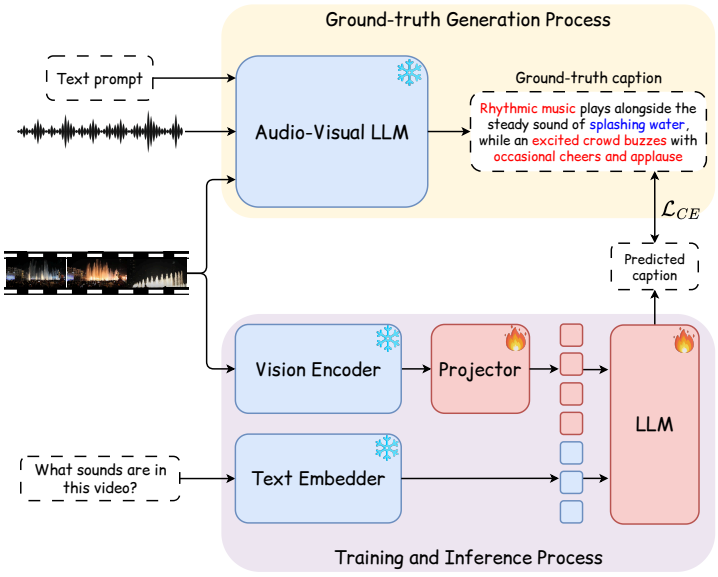


Fig. 4: Overview of SoundCap's ground-truth generation, training, and inference pipeline. Blue labels denote audio events that a standard VLM can detect, while red labels highlight events that are challenging to infer from visuals alone.

labels (most under 6-10 words), limiting the semantic granularity available to the model. SoundCap remedies this limitation by: (1) leveraging an AV-LLM to craft fine-grained, sound-oriented captions and (2) employing these captions to supervise a VLM that in turn enhances Flowley's conditioning signals. As shown in Figure 4, for each video-audio pair (V, A) and instruction prompt $T_p = \{t_p^1, \ldots, t_p^K\}$, the AV-LLM $P_\Theta(T_a|T_p, V, A)$ produces a detailed audio description $T_a = \{t_a^1, \ldots, t_a^N\}$. This generated description not only captures both on- and off-screen sounds, but also remains consistent with the video's visual and acoustic cues, owing to the unified strength of the multimodal model. Next, we treat T_a as ground truth to fine-tune a VLM: feeding it the visual input and a prompt T_i, we train it to reproduce the sound-oriented description by minimizing the negative log-likelihood of the target tokens conditioned on the video and instruction:

$$\mathcal{L}_{\text{vlm}}(\hat{T}_a|T_i, V) = -\sum_{l=1}^{N} \log[P_\Phi(\hat{t}_a^l = t_a^l|T_i, V)],$$

where Φ is the set of trainable weights of VLM. At inference time, only the VLM is executed, enabling sound-oriented captioning without access to audio. This design explicitly teaches the VLM to recover auditory events that are hard to tell from visuals alone (red) while retaining its ability to describe visually evident (blue) cues (Figure 4), thus enriching the conditioning signal later employed by

Flowley. Furthermore, as VGGSound is an in-the-wild dataset characterized by inherent noise (e.g., irrelevant background audio or speech), we insert specific instructional "warnings" into the prompts to guide the AV-LLM generation. This strategy prevents the VLM from learning from misaligned samples, ensuring more accurate captioning. Further details of the text prompts and examples of audio captions are described in Supplementary Material.

4 Experimental Results

4.1 Experimental Setup

Implementation Details. Flowley employs $N_1 = 5$ multi-stream blocks followed by $N_2 = 10$ single-stream blocks, with the hidden dimension $D = 512$. Each Flowley block uses 8 attention heads and a feed-forward network multiplier of four. We apply a dropout rate of 0.1 and limit input text sequences to a maximum length of 77 tokens. We use $\lambda = 0.5$ and $\omega = 0$ for all experiments, meaning that only exact video-audio frames pair are allowed to fully attend to each other, while they can attend to other frames in fade zone within the range $\delta = 4$. At inference, we integrate the ODE using Euler's method with 25 steps and apply CFG with a strength of $s = 7.5$.

For SoundCap training, we use video-SALMONN [44] to generate ground-truth text, which is subsequently employed to finetune Qwen2.5-VL [2] for one epoch. We further emphasize that SoundCap is independent of Flowley and serves as an *optional*, one-time recaptioning step to enrich the captions. Consequently, it is not considered an additional stage in Flowley's training pipeline. Additional details are provided in the Supplementary Material.

Datasets and Baselines. For a fair comparison with existing baselines, we train Flowley on VGGSound, a standard and large-scale dataset containing approximately 200k 10-second video clips, accompanied by corresponding audio tracks. Following previous baselines, we only use the first 8 seconds of each video for training and evaluation. To provide a comprehensive performance analysis, we benchmark Flowley against seven recent SOTA methods under same standardized conditions, all of which are described in the Supplementary Material.

Evaluation Metrics. We follow previous works to evaluate our method along four dimensions. First, **distribution matching** measures how closely the statistical properties of generated audio align with those of real recordings. To this end, we calculate the Fréchet Audio Distance (FAD) [21] and Kernel Audio Distance (KAD) [9], both using PANNs [21] as the embedding network, and employ PaSST [22] to compute KL divergence in accordance with AudioLDM. Second, **audio quality** is quantified via the Inception Score (IS) [41], also leveraging PANNs as the classifier per Frieren. Third, **semantic alignment** assesses the consistency between the input video content and the synthesized audio by extracting visual and audio embeddings with ImageBind [14]

Table 1: Results on the VGGSound test split. Reported parameter counts omit feature encoders, decoders, and vocoders. The [†] denotes results we reproduced using the authors' official checkpoints and inference scripts; [‡] indicates evaluation on generated samples obtained from the authors; and [⋄] marks models that were retrained on the VGGSound dataset using the official implementation. **Best results** are written in bold, and the runner-up scores are underlined. Tiny green numbers indicate the performance boost from incorporating SoundCap.

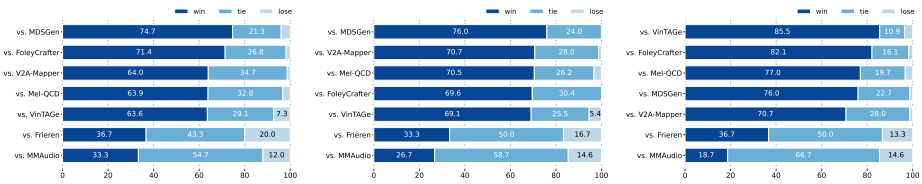
Method	Params	Distribution Matching			Quality	Semantic Alignment(×100)		Sync.
		KAD↓	FAD↓	KL↓	IS↑	IB-Score↑	LB-Score↑	Align Acc↑
Frieren [53]†	159M	1.27	12.8	2.82	12.02	22.45	19.09	**97.13**
FoleyCrafter [57]†	1.22B	1.54	19.17	2.19	15.09	25.75	24.66	77.15
V2A-Mapper [51]‡	229M	1.34	11.73	2.50	12.43	22.38	22.32	79.08
MDSGen [37]†	131M	5.33	39.68	2.85	6.87	17.75	19.05	<u>91.70</u>
Mel-QCD [52]†	859M	1.53	19.17	2.09	10.32	23.79	23.80	73.85
VinTAGe [23]†	1.32B	1.08	17.88	2.15	17.34	21.10	21.51	67.11
MMAudio [7]⋄	157M	0.57	7.89	1.91	12.68	28.09	21.98	89.73
+ **SoundCap**	157M	(↑31.6%) **0.39**	(↑10.1%) **7.09**	(↑18.3%) **1.56**	(↑15.8%) 14.68	(↑2.7%) 28.85	(↑3.0%) 22.64	(↑0.8%) 90.53
Flowley	169M	<u>0.42</u>	7.65	<u>1.57</u>	<u>18.25</u>	<u>29.32</u>	<u>24.87</u>	89.37
+ **SoundCap**	169M	(↑7.14%) **0.39**	(↑1.7%) <u>7.52</u>	(↑0.6%) **1.56**	(↑7.8%) **19.68**	(↑2.6%) **30.07**	(↑1.8%) **25.33**	(↑0.7%) 90.02

and LanguageBind [59] and reporting their average cosine similarity, denoted as the IB-Score and LB-Score, respectively. Finally, **sync.** evaluates temporal coherence by adopting Alignment Accuracy (Align Acc) metric [31].

4.2 Results

Objective Evaluation. As shown in Table 1, Flowley outperforms existing baselines across all metrics, with the sole exception of the Align Acc metric. Regarding distribution matching, it leads all methods by significant margins. Specifically, its KAD score of 0.42 represents a 26.3% gain over the next best baseline, MMAudio (0.57) and nearly triples the performance of the third-ranked method, VinTAGe (1.08). In terms of generation quality, Flowley attains an Inception Score of 18.25, exceeding VinTAGe despite the latter having roughly eight times more parameters. Our synthesized audio further demonstrates exceptional audio-video correspondence, achieving top results on semantic alignment benchmarks. Finally, on synchronization metric, Flowley surpasses approaches that rely on explicit time-conditioned modules (e.g., onset and energy predictors) such as Mel-QCD and FoleyCrafter. It remains competitive with MMAudio and MDSGen, while only trailing Frieren, all of which utilize a pretrained audio-visual alignment encoder. Notably, *MDSGen and Frieren employ the same pretrained encoder for video feature extraction that serves as the backbone for the Align Acc metric*. This overlap may introduce evaluation bias, potentially inflating the scores of these specific models. To provide a more comprehensive assessment, we conduct a subjective evaluation, detailed in the following section.

The four bottom rows of Table 1 present the results and performance gains obtained by incorporating SoundCap-generated detailed text descriptions into the training and inference pipelines of MMAudio and Flowley. SoundCap acts



(a) Audio Quality. **(b)** Semantic Alignment. **(c)** Temporal Alignment.

Fig. 5: Human preference comparison between Flowley and competing methods.

as a powerful performance multiplier, most notably for MMAudio in the distribution matching category. By incorporating these rich descriptions, MMAudio achieves the best FAD score of 7.09. Furthermore, SoundCap yields a significant 31.6% improvement in KAD for MMAudio. For Flowley, SoundCap also improves overall performance across several metrics, yielding gains of up to 7.8%. These findings demonstrate the effectiveness of SoundCap across different frameworks and highlight the importance of rich text descriptions, which are largely absent in current datasets and can be addressed through SoundCap. We provide qualititative examples and analyses in the Supplementary Material.

Subjective Evaluation. For a more comprehensive assessment, we complemented our quantitative findings with subjective human evaluations. To mitigate the inherent noise of the VGGSound test set, we rigorously filtered the data by excluding samples with low ImageBind alignment scores, low resolution ($\leq$360p), or human speech. From this subset, we randomly selected 30 videos to construct 180 distinct A/B comparison pairs (Flowley vs. each baseline). Twenty participants were recruited, each rating 36 pairs to ensure every pair was independently evaluated four times, yielding 720 total responses.

As illustrated in Figure 5, Flowley demonstrates a clear and consistent advantage across all categories. Notably, although Flowley trailed MMAudio, MDSGen, and Frieren in the Align Acc metric (Table 1), it was preferred over all three in human subjective testing for temporal synchronization. This preference is particularly pronounced against MDSGen, where Flowley achieved a 76% win rate. These results highlight the effectiveness of our approach and suggest that model-based metrics may not fully capture the perceptual nuances of audio-visual alignment.

4.3 Ablation Studies

Sensitivity of PSCA. We evaluate the sensitivity of PSCA and its influence on overall system performance by varying the hard-attention window size ω and soft-attention zone size δ, as shown in Table 2. Increasing ω and δ generally leads to lower Align Acc metrics, which can be attributed to the expansion of the context window that introduces a greater risk of audio-video misalignment. In contrast, IS and IB-Score remain relatively stable across these hyperparameter

Table 2: Results obtained by varying the hyperparameters of the PSCA module. Text in blue indicates our default settings.

ω	δ	IS↑	IB-Score↑	Align Acc↑
0	0	17.85	28.84	88.56
	2	18.18	29.21	88.37
	4	18.25	29.32	**89.37**
	8	18.27	29.34	88.02
2	0	18.15	**29.48**	88.46
	2	**18.57**	29.35	88.03
	4	18.2	29.44	87.58
	8	18.22	29.21	87.23
4	0	18.33	29.39	88.37
	2	18.54	29.26	88.11
	4	18.48	29.36	87.22
	8	18.52	29.18	86.58

Table 3: Results obtained by altering single-modal stream block.

#	CA_t	CA_v	$PSCA_v$	KAD↓	IS↑	IB-Score↑	Align Acc↑
1	✗	✗	✗	0.44	16.99	27.38	86.26
2	✓	✗	✗	0.44	17.29	<u>28.93</u>	87.43
3	✗	✓	✗	**0.40**	17.58	28.02	87.04
4	✗	✗	✓	**0.40**	17.69	28.51	<u>88.72</u>
5	✓	✓	✗	**0.40**	<u>18.15</u>	28.74	87.61
6	✓	✗	✓	<u>0.42</u>	**18.25**	**29.32**	**89.37**

Table 4: Results obtained by removing layer-dependent progressive parameter.

β	KAD↓	IS↑	IB-Score↑	Align Acc↑
✗	**0.42**	17.78	**29.53**	88.62
✓	**0.42**	**18.25**	29.32	**89.37**

Table 5: Ablation study on the impact of noise-aware conditioning in SoundCap.

Method	KAD↓	IS↑	IB-Score↑	Align Acc↑
Flowley	<u>0.42</u>	<u>18.25</u>	29.32	<u>89.37</u>
+ SoundCap w/o noise conditions	0.45	15.16	<u>29.61</u>	87.13
+ SoundCap w/ noise conditions	**0.39**	**19.68**	**30.07**	**90.02**

settings, suggesting that PSCA maintains robustness in terms of audio quality and semantic consistency. These findings are consistent with our design intuition that PSCA primarily affects audio-video synchronization while leaving overall semantic alignment largely unchanged.

Modules in Single-stream Blocks. We investigate the impact of different configurations of text-audio and visual-audio cross-attention (CA), as well as the effectiveness of the proposed PSCA mechanism. As presented in Table 3, we conduct an ablation study by independently removing the cross-attention to either the textual or visual features (or both) and by replacing the PSCA mechanism with a standard cross-attention module. Our key findings are as follows: (1) Reducing the single-stream block to a vanilla DiT by removing both CA streams results in a significant performance drop across all metrics, highlighting the importance of dual cross-attention (Row 1 *vs.* 6); (2) Incorporating textual features consistently improves the semantic relevance of the generated audio (Rows 2, 5, 6); and (3) Employing the progressive soft mask M improves the temporal synchronization (Rows 4 and 6).

Impact of layer-dependent progressive parameter. To understand the contribution of β to the overall performance, we perform the ablation study

Table 6: Zero-shot performance on MovieGen Audio Bench.

Method	Params	Data(h)	IS↑	IB-Score↑	Align Acc↑
Movie Gen Audio	13B	$\mathcal{O}$(1M)	8.01	**35.86**	**64.33**
Flowley	169M	$\sim$ 400	7.74	23.65	59.28
+ SoundCap	169M	$\sim$ 400	**8.18**	25.78	61.07

presented in Table 4, where the depth-aware parameter is removed from the cross-attention masking calculation. The results show that this omission leads to a notable decline in both IS and Align Acc. While the IB-Score remains relatively stable, the degradation in audio quality and temporal synchronization metrics underscores the critical role of β in achieving precise audio-visual alignment.

Handling noise in SoundCap. Following the discussion in Section 3.5, we mitigate the inherent noise of the VGGSound dataset by inserting specific noise conditions into the text prompts. The effectiveness of this approach is evaluated in Table 5, where we compare Flowley's performance using two distinct prompt configurations (detailed in the Supplementary Material). The results demonstrate that our noise-handling strategy yields consistent improvements across all metrics. Conversely, when SoundCap is unaware of these conditions, audio quality significantly degrades: IS drops by 16.9%, Align Acc decreases to 87.13%, and KAD rises marginally. These findings underscore that noise-robust conditioning is vital for effectively leveraging in-the-wild datasets, as it prevents the model from converging on misaligned or low-quality audio-visual pairs.

Zero-shot Performance. Finally, we explore the boundaries of our method by testing Flowley on the Movie Gen Audio Bench [38], a synthetic benchmark dataset whose distribution deviates significantly from that of VGGSound. Table 6 summarizes our results. Surprisingly, despite being 77$\times$ smaller and trained on just 1/2500 of the data, Flowley performs comparably to Movie Gen Audio [38] in terms of audio quality (0.27 difference) and outperforms it when SoundCap is incorporated into the inference pipeline, while marginally falling behind in Align Acc by 3.26%. Regarding semantic alignment, Movie Gen Audio outperforms both versions by a notable margin. We believe this is due to the limited scope of our training data, which restricts our model's ability to generalize to the wider and more varied content present in the benchmark. We present additional details and results in the Supplementary Material.

5 Conclusion

We present Flowley, an end-to-end, flow-based multimodal framework for generating audio that is both semantically meaningful and temporally synchronized

with silent video content. Unlike prior works that rely on multi-stage training pipelines or pretrained audio-visual alignment modules, Flowley directly learns both semantic and temporal alignment through the integration of dual cross-attention pathways and a task-specific progressive masking mechanism tailored for video-to-audio synthesis. To further enhance semantic fidelity, we propose SoundCap, a dedicated sound-aware captioning pipeline that provides enriched textual conditioning for guiding audio generation. Empirical evaluations show that Flowley achieves state-of-the-art results in VGGSound across multiple metrics, outperforming models up to 8× larger in parameter size. Moreover, when SoundCap is incorporated into the inference pipeline, Flowley delivers the highest audio quality in a zero-shot setting, outperforming a model 77× larger and trained on 2500× more data on a synthetic evaluation set. Comprehensive ablation studies confirm the effectiveness of each architectural component and highlight the robustness of the overall approach.

References

1. ATCHISON, J., SHEN, S.: Logistic-normal distributions:some properties and uses. Biometrika **67**(2), 261–272 (08 1980). https://doi.org/10.1093/biomet/67.2.261, https://doi.org/10.1093/biomet/67.2.261
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
4. BFLabs: Black-forest-labs/flux.1-dev (2024), https://huggingface.co/black-forest-labs/FLUX.1-dev
5. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators
6. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053174
7. Cheng, H.K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., Mitsufuji, Y.: Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 28901–28911 (June 2025)
8. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models. Journal of Machine Learning Research **25**(70), 1–53 (2024), http://jmlr.org/papers/v25/23-0870.html

9. Chung, Y., Eu, P., Lee, J., Choi, K., Nam, J., Chon, B.S.: Kad: No more fad! an effective and efficient evaluation metric for audio generation. arXiv preprint arXiv:2502.15602 (2025)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 12606–12633. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/esser24a.html
11. Evans, Z., Parker, J.D., Carr, C., Zukowski, Z., Taylor, J., Pons, J.: Stable audio open. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10888461
12. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A.T., Shankar, V.: Data filtering networks. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=KAk6ngZ09F
13. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780 (2017). https://doi.org/10.1109/ICASSP.2017.7952261
14. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15180–15190 (June 2023)
15. Gong, Y., Luo, H., Liu, A.H., Karlinsky, L., Glass, J.R.: Listen, think, and understand. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=nBZBPXdJlC
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
17. Iashin, V., Rahtu, E.: Taming visually guided sound generation. In: British Machine Vision Conference (BMVC) (2021)
18. Jeong, Y., Kim, Y., Chun, S., Lee, J.: Read, watch and scream! sound generation from text and video. Proceedings of the AAAI Conference on Artificial Intelligence **39**(17), 17590–17598 (Apr 2025). https://doi.org/10.1609/aaai.v39i17.33934, https://ojs.aaai.org/index.php/AAAI/article/view/33934
19. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: The Second International Conference on Learning Representations (2014), https://openreview.net/forum?id=33X9fd2-9FyZd
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
21. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio, Speech and Lang. Proc. **28**, 2880–2894 (Nov 2020). https://doi.org/10.1109/TASLP.2020.3030497, https://doi.org/10.1109/TASLP.2020.3030497

22. Koutini, Khaled and Schlüter, Jan and Eghbal-Zadeh, Hamid and Widmer, Gerhard: Efficient Training of Audio Transformers with Patchout. In: Interspeech 2022. pp. 2753–2757 (2022). https://doi.org/{10.21437/Interspeech.2022-227}
23. Kushwaha, S.S., Tian, Y.: Vintage: Joint video and text conditioning for holistic audio generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 13529–13539 (June 2025)
24. gil Lee, S., Ping, W., Ginsburg, B., Catanzaro, B., Yoon, S.: BigVGAN: A universal neural vocoder with large-scale training. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=iTtGCMDEzS_
25. Li, Y., Jiang, H., Kodaira, A., Tomizuka, M., Keutzer, K., Xu, C.: Immiscible diffusion: Accelerating diffusion training with noise assignment. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 90198–90225. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/a422a2f016c14406a01ddba731c0969a-Paper-Conference.pdf
26. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t
27. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: AudioLDM: Text-to-audio generation with latent diffusion models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 21450–21474. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/liu23f.html
28. Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., Plumbley, M.D.: Audioldm 2: Learning holistic audio generation with self-supervised pretraining. IEEE/ACM Trans. Audio, Speech and Lang. Proc. **32**, 2871–2883 (May 2024). https://doi.org/10.1109/TASLP.2024.3399607, https://doi.org/10.1109/TASLP.2024.3399607
29. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=XVjTT1nw5z
30. Liu, X., Su, K., Shlizerman, E.: Tell what you hear from what you see - video to audio generation through text. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 101337–101366. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/b782a3462ee9d566291cff148333ea9b-Paper-Conference.pdf
31. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 48855–48876. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/98c50f47a37f63477c01558600dd225a-Paper-Conference.pdf
32. Majumder, N., Hung, C.Y., Ghosal, D., Hsu, W.N., Mihalcea, R., Poria, S.: Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 564–572. MM '24, Association for Computing Machin-

ery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3681688, https://doi.org/10.1145/3664647.3681688
33. Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., Bengio, Y.: SampleRNN: An unconditional end-to-end neural audio generation model. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=SkxKPDv5xl
34. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
35. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (October 2023)
36. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018). https://doi.org/10.1609/aaai.v32i1.11671, https://ojs.aaai.org/index.php/AAAI/article/view/11671
37. Pham, T.X., Ton, T., Yoo, C.D.: MDSGen: Fast and efficient masked diffusion temporal-aware transformers for open-domain sound generation. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=yFEqYwgttJ
38. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
40. Ren, Y., Li, C., Xu, M., Liang, W., Gu, Y., Chen, R., Yu, D.: Sta-v2a: Video-to-audio generation with semantic and temporal alignment. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10890132
41. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016), https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf
42. Sheffer, R., Adi, Y.: I hear your true colors: Image guided audio generation. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10096023
43. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (Feb 2024). https://doi.org/10.1016/j.neucom.2023.127063, http://dx.doi.org/10.1016/j.neucom.2023.127063
44. Sun, G., Yu, W., Tang, C., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Wang, Y., Zhang, C.: video-SALMONN: Speech-enhanced audio-visual large language

models. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 47198–47217. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/sun24l.html

45. Sun, L., Xu, X., Wu, M., Xie, W.: Auto-acd: A large-scale dataset for audio-language representation learning. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 5025–5034. MM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3681472, https://doi.org/10.1145/3664647.3681472
46. Theme Ament, V.: The Foley Grail. Routledge (Apr 2014). https://doi.org/10.4324/9780203766880, http://dx.doi.org/10.4324/9780203766880
47. Ton, T., Hong, J.W., Yoo, C.D.: Taro: Timestep-adaptive representation alignment with onset-aware conditioning for synchronized video-to-audio synthesis. arXiv preprint arXiv:2504.05684 (2025)
48. Van, C.T., Tran, T.V.T., Nguyen, V., Hy, T.S.: Effective context modeling framework for emotion recognition in conversations. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10888112
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
50. Wang, C., Qi, Q., Wang, J., Sun, H., Zhuang, Z., Wu, J., Zhang, L., Liao, J.: Chattime: A unified multimodal time series foundation model bridging numerical and textual data. Proceedings of the AAAI Conference on Artificial Intelligence **39**(12), 12694–12702 (Apr 2025). https://doi.org/10.1609/aaai.v39i12.33384, https://ojs.aaai.org/index.php/AAAI/article/view/33384
51. Wang, H., Ma, J., Pascual, S., Cartwright, R., Cai, W.: V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. Proceedings of the AAAI Conference on Artificial Intelligence **38**(14), 15492–15501 (Mar 2024). https://doi.org/10.1609/aaai.v38i14.29475, https://ojs.aaai.org/index.php/AAAI/article/view/29475
52. Wang, J., Xu, C., Yu, C., Shang, L., Hu, Z., Wang, S., Bo, L.: Synchronized video-to-audio generation via mel quantization-continuum decomposition. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 3111–3120 (June 2025)
53. Wang, Y., Guo, W., Huang, R., Huang, J., Wang, Z., You, F., Li, R., Zhao, Z.: Frieren: Efficient video-to-audio generation network with rectified flow matching. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 128118–128138. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/e7384de302bef52319b5067c3115bbfb-Paper-Conference.pdf
54. Xie, Z., Yu, S., He, Q., Li, M.: Sonicvisionlm: Playing sound with vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26866–26875 (June 2024)

55. Yuan, Y., Jia, D., Zhuang, X., Chen, Y., Liu, Z., Chen, Z., Wang, Y., Wang, Y., Liu, X., Kang, X., Plumbley, M.D., Wang, W.: Sound-VECaps: Improving audio generation with visual enhanced captions. In: Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation (2024), https://openreview.net/forum?id=mPmOxpTivB
56. Zhang, K., Pham, T.X., Lee, S., Niu, A., Senocak, A., Chung, J.S.: Model-guided dual-role alignment for high-fidelity open-domain video-to-audio generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=02qdHz1LU0
57. Zhang, Y., Gu, Y., Zeng, Y., Xing, Z., Wang, Y., Wu, Z., Chen, K.: Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds (2024), https://arxiv.org/abs/2407.01494
58. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Visual to sound: Generating natural sound for videos in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
59. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., HongFa, W., Pang, Y., Jiang, W., Zhang, J., Li, Z., Zhang, C.W., Li, Z., Liu, W., Yuan, L.: Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=QmZKc7UZCy
60. Zhuo, L., Du, R., Xiao, H., Li, Y., Liu, D., Huang, R., Liu, W., Zhu, X., Wang, F.Y., Ma, Z., Luo, X., Wang, Z., Zhang, K., Zhao, L., Liu, S., Yue, X., Ouyang, W., Qiao, Y., Li, H., Gao, P.: Lumina-next : Making lumina-t2x stronger and faster with next-dit. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 131278–131315. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/ed2dad593d87ca474a636cba610a29d3-Paper-Conference.pdf