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] Van Nguyen[1]
[1]FPT Software AI Center, Vietnam [2]NVIDIA Corporation, USA

Abstract

Video-to-audio (V2A) generation aims to create realistic, synchronized sound for silent videos. However, current methods struggle with either expensive multi-stage training or poor temporal alignment when relying on text-based models. To address this, we introduce Flowley, an end-to-end, single-stage architecture that synthesizes audio using visual features and text prompts. Its core innovation, Progressive Soft-masked Cross-Attention, embeds audio-visual synchronization directly into the attention mechanism at zero extra cost. 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. Flowley achieves state-of-the-art performance on VGGSound without pretrained alignment modules, and adding SoundCap further enhances results across architecture designs.

1. Introduction

Video-to-audio (V2A) generation automates Foley-style sound synthesis from silent video. Existing methods follow two directions. The first converts visual frames to text and leverages pretrained text-to-audio models [14, 17], discarding fine-grained temporal details. The second employs multi-stage training [15, 10], suffering from high computational overhead. Furthermore, integrating textual descriptions into the generation workflow, beyond their use in T2A systems, remains under-explored, partly due to existing V2A datasets [2] lacking rich, sound-focused text annotations.

We introduce *Flowley*, a flow-based architecture synthesizing high-fidelity audio from silent video using multimodal cues. Our framework consists of: **(1)** multi-stream blocks that jointly fuse audio-latent, visual, and textual embeddings, and **(2)** single-stream blocks that refine the audio pathway. Within the single-stream blocks, we propose ***P**rogressive **S**oft-masked **C**ross-**A**ttention* (PSCA), a progressive masking mechanism that directly aligns acoustic and visual features on a frame-wise basis, eliminating the need for pretrained synchronization modules without introducing additional computation. In practice, 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. Experiments on VGGSound show that Flowley exceeds state-of-the-art (SOTA) methods across multiple metrics, while the addition of SoundCap further enhances performance across various architectural designs.

2. Methodology

2.1. 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 [3]. After obtaining representations from pretrained encoders, Flowley first employs N_1 multi-stream blocks [5], 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.

Multi-stream Block. As shown in Figure 1b, the multi-stream block 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 by integrating QK-Norm and Rotary Positional Embeddings (RoPE) 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 [11] 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.

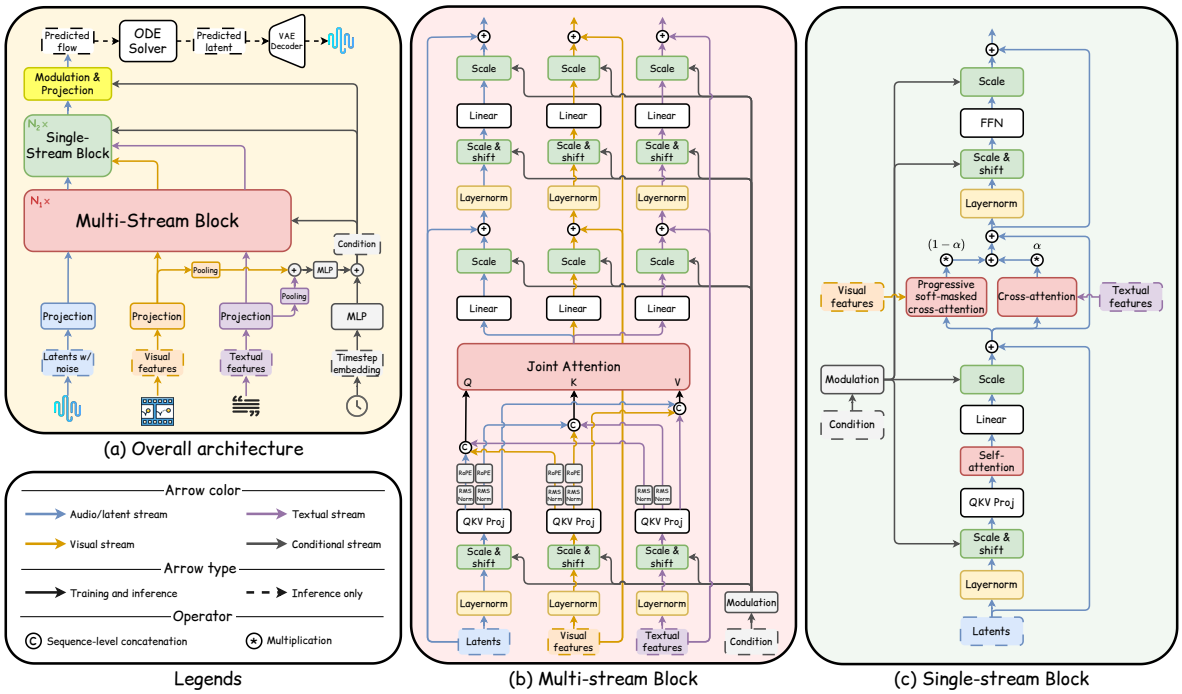


Figure 1: **(a)** Overview of the Flowley framework. **(b)** The multi-stream block jointly processes visual, textual, and audio latent representations. **(c)** The single-stream block refines the audio latents via weighted cross-attention with the visual and textual streams to estimate the flow field.

2.2. Progressive Soft-masked Cross-Attention

The original cross-attention mechanism enables the decoder to focus on the most relevant portions of the encoder's outputs. 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 $M^{(\ell)}$ with entries in the range of $[0, 1]$, where values approaching zero impose stronger masking. Specifically, we denote the audio query Q_{aud} and the video key/value sequences $K_{\text{vis}}, V_{\text{vis}}$, 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 a mask that applies a full weight of 1 within a hard window ω, and a cosine decay $\mathcal{F}$ over an extended window δ to retain adjacent contextual motion cues:

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

where $\mathcal{F}(m) = \frac{1}{2}[\cos(\frac{\pi m}{\delta}) + 1]$. In our experiments, we set $\omega = 0$ and $\delta = 4$. To encourage early layers to explore broader contexts while forcing deeper layers to focus on precise alignment, we scale the fade zone using a progressive parameter $\beta_\ell = 1 - \frac{\ell}{N_2 - 1}$. As layer depth ℓ increases, β_ℓ approaches zero, effectively reducing the mechanism to a hard sliding window. Given the mask formulation, we formally express the PSCA mechanism at layer ℓ as:

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

where ϵ prevents $\log(0)$ and d_k is the hidden dimension per head. PSCA unifies *hard* locality, *soft* fading, and *depth-aware* progression into a single differentiable mask. Early layers leverage broader video context to reduce alignment errors, while deeper layers focus on strictly aligned frames for fine-grained synchronization. Finally, the updated audio latent representation $\tilde{x}^{(\ell)}$ dynamically balances text and vision conditioning via a learnable parameter $\alpha^{(\ell)} \in [0, 1]$:

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

2.3. Sound-aware Captioner

To address the lack of descriptive labels in existing V2A datasets, we introduce SoundCap (Figure 2). First, we use a

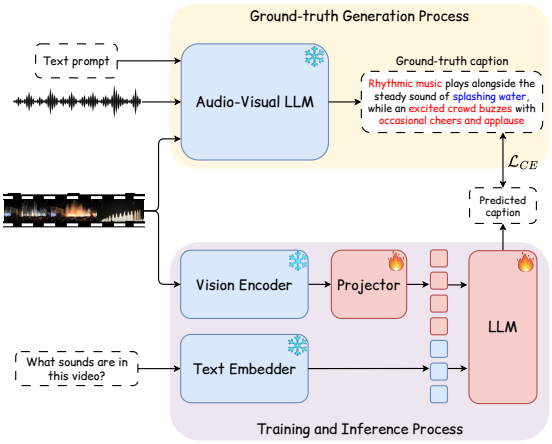


Figure 2: Overview of the SoundCap pipeline. Blue labels denote visually evident audio events, while red labels indicate sounds difficult to infer from video alone.

pretrained AV-LLM [13] to generate highly detailed, sound-oriented captions $T_a = \{t_a^1, \ldots, t_a^N\}$ from video-audio pairs (V, A) and prompts T_p. Next, we use T_a as ground truth to fine-tune a VLM [1], teaching it to predict these audio descriptions from silent video V alone by minimizing the negative log-likelihood:

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

where Φ is the set of trainable weights of VLM. During inference, the VLM operates without audio to generate rich conditioning descriptions for Flowley, predicting both visually evident (blue) and off-screen (red) sounds (Figure 2). 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.

3. Experimental Results

3.1. Experimental Setup

Datasets. 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.

Evaluation Metrics. Following prior work, we evaluate our method across four dimensions. For **distribution matching**, we calculate Fréchet Audio Distance (FAD) and Kernel Audio Distance (KAD) [4] using PANNs [7], alongside PaSST-based [8] KL divergence. **Audio quality** is measured via the Inception Score (IS). To evaluate **semantic alignment**, we report the average cosine similarity of audio-visual embeddings extracted via ImageBind [6] (IB-Score) and LanguageBind [18] (LB-Score). Finally, **sync.** is assessed using the Alignment Accuracy (Align Acc) metric [10].

3.2. Results

Objective Evaluation. As shown in Table 1, Flowley outperforms existing baselines across almost all metrics. It leads significantly in distribution matching with a KAD of 0.42 (a 26.3% improvement over the second-best, MMAudio) and achieves an Inception Score of 18.25, surpassing much larger models like VinTAGe. Flowley also attains top semantic alignment results. While it trails Frieren, MDSGen, and MMAudio in the Align Acc synchronization metric, both Frieren and MDSGen utilize the same pretrained encoder as the evaluation metric itself, which potentially inflates their scores. To provide a more comprehensive assessment, we conduct a subjective evaluation in the following section.

Furthermore, integrating SoundCap-generated descriptions significantly boosts performance across frameworks (Table 1, bottom rows). SoundCap acts as a powerful multiplier: it enables MMAudio to achieve the best overall FAD score (7.09) alongside a 31.6% KAD improvement, while Flowley yields metric gains of up to 7.8%. These findings underscore the critical role of rich textual conditioning, which often absent in existing datasets, and validate SoundCap's cross-framework applicability.

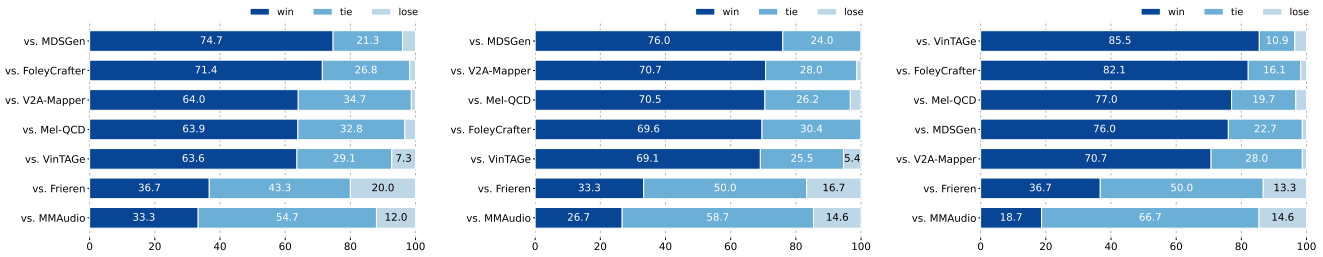
Subjective Evaluation. To complement our quantitative findings, we conducted human A/B evaluations on an ImageBind-pruned subset (to mitigate inherent dataset noise). Evaluators compared Flowley against baselines based on audio quality, semantic alignment, and temporal synchronization. As illustrated in Figure 3, Flowley was consistently preferred across all categories. Crucially, despite trailing MMAudio, MDSGen, and Frieren in the objective Align Acc metric, human raters favored Flowley's temporal synchronization over all three, including a notable 76% win rate against MDSGen. This suggests that current model-based metrics may not fully capture the perceptual nuances of true audio-visual alignment.

4. Conclusion

We present Flowley, an end-to-end, flow-based multimodal framework for generating semantically and temporally synchronized audio from silent video. Unlike prior works requiring multi-stage pipelines or pretrained alignment modules, Flowley directly learns synchronization via Progressive Soft-masked Cross-Attention (PSCA) and dual cross-attention pathways. To improve semantic fidelity, we propose SoundCap, a sound-aware captioning pipeline that

Table 1: Results on the VGGSound test split. 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.

Method	Params	Distribution Matching			Quality	Semantic Alignment(×100)		Sync.
		KAD↓	FAD↓	KL↓	IS↑	IB-Score↑	LB-Score↑	Align Acc↑
Frieren [16]†	159M	1.27	12.8	2.82	12.02	22.45	19.09	**97.13**
FoleyCrafter [17]†	1.22B	1.54	19.17	2.19	15.09	25.75	24.66	77.15
V2A-Mapper [14]‡	229M	1.34	11.73	2.50	12.43	22.38	22.32	79.08
MDSGen [12]†	131M	5.33	39.68	2.85	6.87	17.75	19.05	<u>91.70</u>
Mel-QCD [15]†	859M	1.53	19.17	2.09	10.32	23.79	23.80	73.85
VinTAGe [9]†	1.32B	1.08	17.88	2.15	17.34	21.10	21.51	67.11
MMAudio [3]◇	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



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

Figure 3: Human preference comparison between Flowley and competing methods on the VGGSound test split.

enriches text conditioning. Experiments show that Flowley achieves state-of-the-art results on VGGSound, outperforming models up to 8× larger. Moreover, adding SoundCap further enhances results across architecture designs.

References

[1] S. Bai et al. Qwen2. 5-vl technical report. *arXiv:2502.13923*, 2025. 3

[2] H. Chen et al. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 1

[3] H. K. Cheng et al. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, 2025. 1, 4

[4] Y. Chung et al. Kad: No more fad! an effective and efficient evaluation metric for audio generation. *arXiv:2502.15602*, 2025. 3

[5] P. Esser et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1

[6] R. Girdhar et al. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 3

[7] Q. Kong et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *TASLP*, 2020. 3

[8] Koutini, Khaled and Schlüter, Jan and Eghbal-Zadeh, Hamid and Widmer, Gerhard. Efficient Training of Audio Transformers with Patchout. In *Interspeech*, 2022. 3

[9] S. S. Kushwaha and Y. Tian. Vintage: Joint video and text conditioning for holistic audio generation. In *CVPR*, 2025. 4

[10] S. Luo et al. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2023. 1, 3

[11] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1

[12] T. X. Pham et al. MDSGen: Fast and efficient masked diffusion temporal-aware transformers for open-domain sound generation. In *ICLR*, 2025. 4

[13] G. Sun et al. video-SALMONN: Speech-enhanced audio-visual large language models. In *ICML*, 2024. 3

[14] H. Wang et al. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. *AAAI*, 2024. 1, 4

[15] J. Wang et al. Synchronized video-to-audio generation via mel quantization-continuum decomposition. In *CVPR*, 2025. 1, 4

[16] Y. Wang et al. Frieren: Efficient video-to-audio generation network with rectified flow matching. In *NeurIPS*, 2024. 4

[17] Y. Zhang et al. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds, 2024. 1, 4

[18] B. Zhu et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2024. 3