SIGHT AND SOUND WORKSHOP

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

by:

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

2 NVIDIA

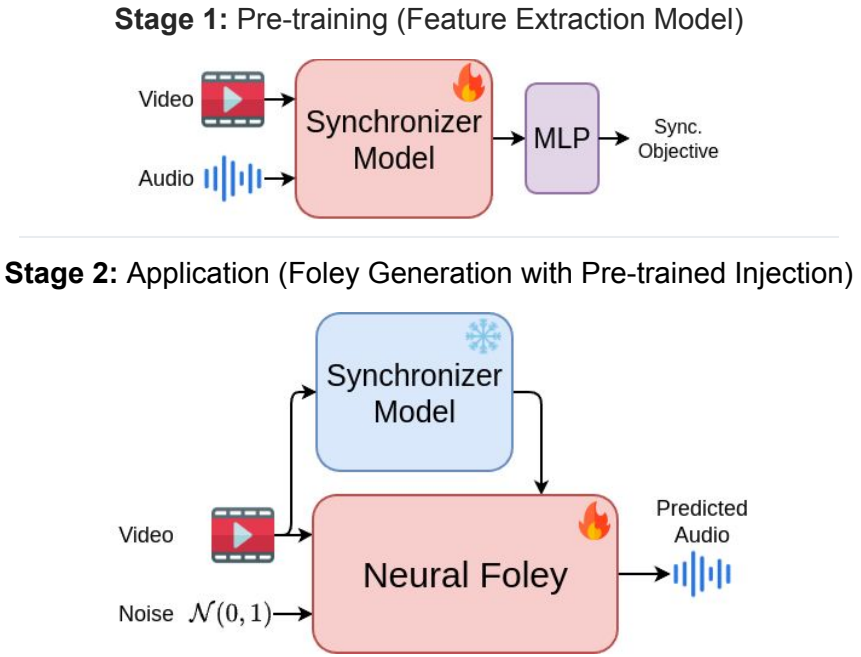
CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

CVPR JUNE 3-7, 2026 DENVER COLORADO

Structural Limitations

Current V2A methods (right) rely on multi-stage training pipelines or external synchronization modules, leading to high computational overhead.



Flowley: Overview

CVPR JUNE 3-7, 2026 DENVER COLORADO

Architecture Highlights

- **Multi-to-Single:** Adopts an efficient structural transition.
- **Single-Stream:** DiT with dual cross-attention layers.
- **Multi-Stream:** Adapted from MM-DiT architecture.
- **No External Modules:** Removes the need for auxiliary synchronizers.

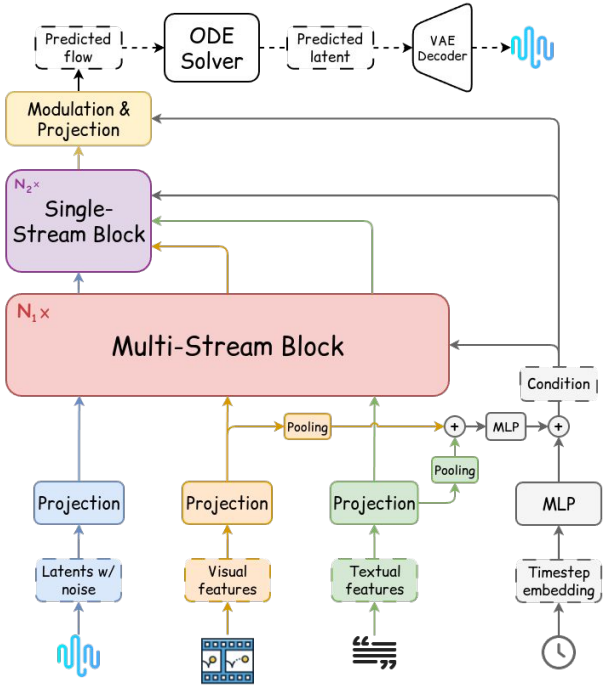


Fig. 1: Architecture overview

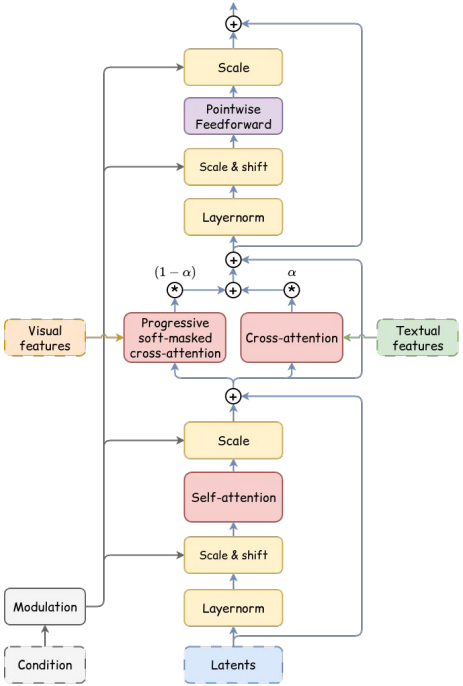


Fig. 2: Single-Stream Block

Flowley: Overview

CVPR JUNE 3-7, 2026 DENVER COLORADO

Architecture Highlights

- **Multi-to-Single:** Adopts an efficient structural transition.
- **Single-Stream:** DiT with dual cross-attention layers.
- **Multi-Stream:** Adapted from MM-DiT architecture.
- **No External Modules:** Removes the need for auxiliary synchronizers.

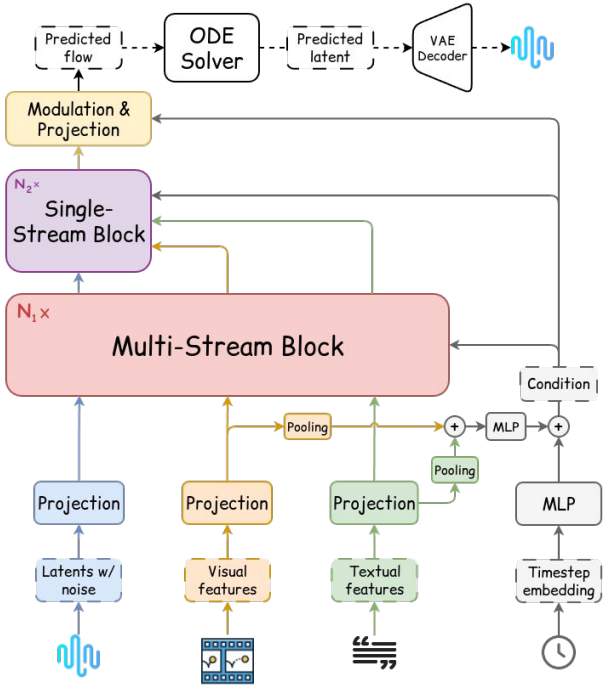


Fig. 1: Architecture overview

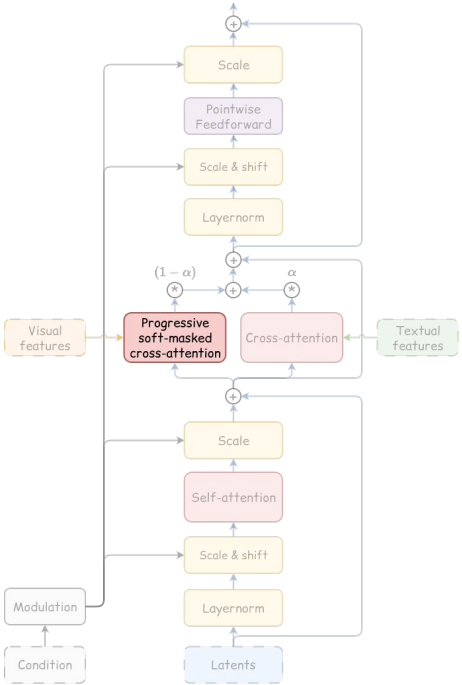


Fig. 2: Single-Stream Block

Flowley: Synchronizer

CVPR JUNE 3-7, 2026 DENVER COLORADO

Core Contribution

To remove the dependency on external modules

⇒ Introduce **Progressive Soft-masked Cross-Attention (PSCA)**

Built directly into the attention mechanism to ensure high-fidelity synchronization in the latent space.

Mathematical Framework

Cross-attention mechanism

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

Progressive Masking

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

Decay Equation

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

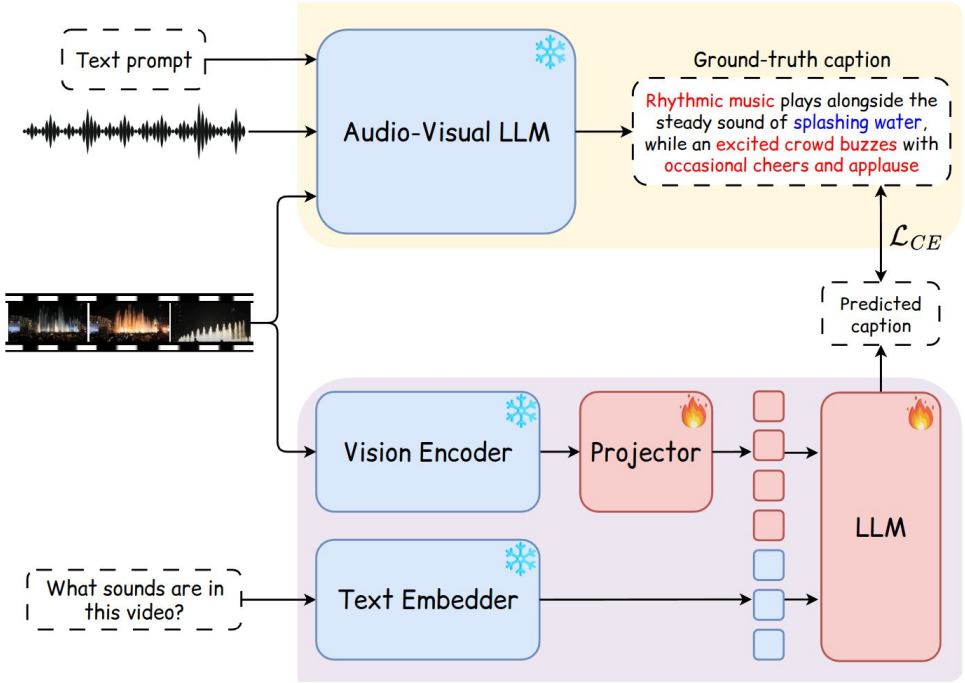
Semantic Bottleneck

Data Limitations

Existing V2A datasets provide only short labels, limiting the semantic granularity available to the model.

SoundCap uses:

- **AV-LLM Teacher:** Generates detailed audio descriptions
- **VLM Student:** Learns to infer sounds from visual alone for inference
- **Noise robustness:** Instructional warnings filter irrelevant background noises (e.g., speech)



SoundCap: Training and Inference Process

Experimental Results

CVPR JUNE 3-7, 2026 DENVER COLORADO

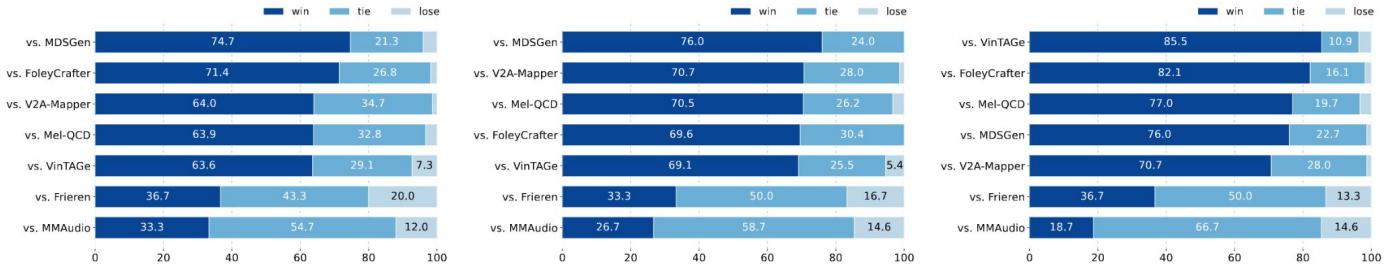
Objective Benchmarking Results

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

Key Takeaways:

- **Flowley** demonstrates superior on both subjective and objective metrics.
- **SoundCap** enhances overall performance by addressing semantic gaps.

Subjective Human Preference Results



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

Ablation Study

CVPR JUNE 3-7, 2026 DENVER COLORADO

Ablation on Single-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**

Ablation on SoundCap's noise handling

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**

Key Takeaways:

- **Significant Impact:** Removing both cross-attention streams results in a significant performance drop (#1 vs. #6).
- **Superiority:** PSCA is better than standard cross-attention (#4 vs. #6).
- **SoundCap's Robustness:** Noise awareness is important for in-the-wild datasets.

THANK YOU!

CVPR
JUNE 3-7, 2026

DENVER
COLORADO