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

ECCV 2026 | MALMÖ | SEPT 8–12

Synchronisation belongs inside the attention mask, not in a bolted-on pretrained module.

flowley-v2a.github.io

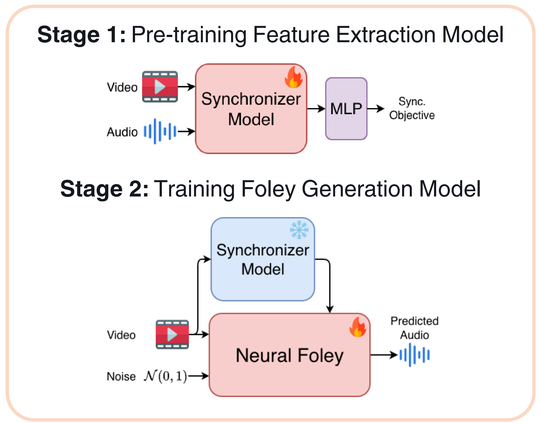
1 The Problem

V2A audio must be semantically right for the video and synchronised to the frame. Synchronisation is usually bought three ways:

(a) Multi-stage training pipelines — expensive and slow.

(b) Video → text → frozen T2A — discards fine-grained temporal cues.

(c) Bolt-on pretrained AV-alignment modules — extra parameters and compute.



Most studies adopted an external pretrained AV-alignment modules

2 Our Approach

Flowley: end-to-end, single-stage training, no pretrained audio-visual alignment module.

PSCA: synchronisation embedded in the attention mask at zero extra compute.

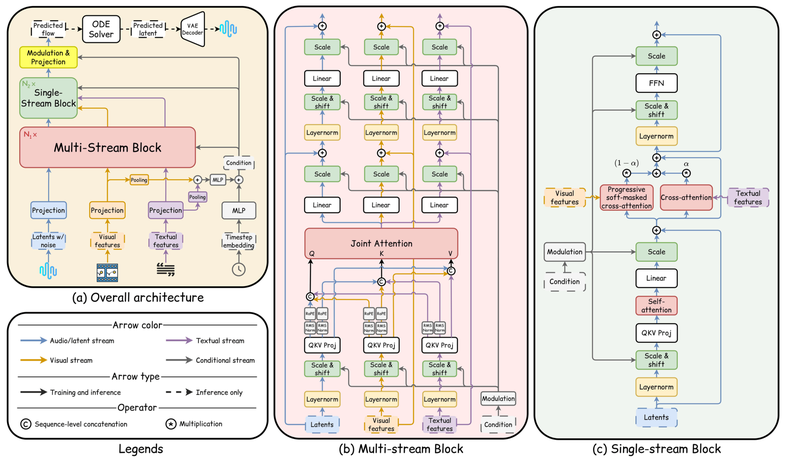
SoundCap: plug-and-play sound-aware recaptioning; optional, one-time, not a training stage.

5 SoundCap

Benchmark captions describe only what is visible; a frozen AV-LLM adds the **off-screen events.**

AV-LLM writes sound-oriented captions → Finetune a VLM on them → Reception any V2A dataset

3 Architecture



MMAudio-like architecture

4 Progressive Soft-masked Cross-Attention

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

hard window

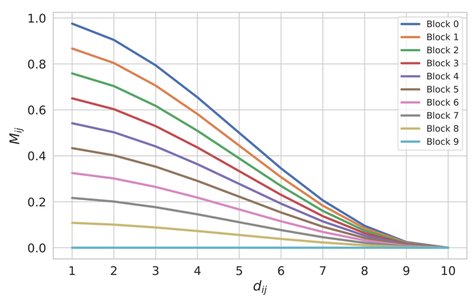
fading, scaled by depth

masked out

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

d_{ij} : distance to the aligned video frame

ω: hard-window radius · δ: fade width



Mask value vs. distance from the aligned frame, from block 0 to block 9

6 Results on VGGSound

Method	Params	KAD↓	IS↑	IB-Score↑	Align Acc↑
Frieren	159M	1.27	12.02	22.45	**97.13**
FoleyCrafter	1.22B	1.54	15.09	25.75	77.15
VinTAGe	1.32B	1.08	17.34	21.10	67.11
MMAudio	157M	0.57	12.68	28.09	89.73
+ SoundCap	157M	↑31.6% **0.39**	↑15.8% 14.68	↑2.7% 28.85	↑0.8% 90.53
Flowley	**169M**	0.42	18.25	29.32	89.37
+ SoundCap	**169M**	↑7.1% **0.39**	↑7.8% **19.68**	↑2.6% **30.07**	↑0.7% **90.02**

VGGSound test set. Bold = best, underline = runner-up, green = SoundCap gain.

7 Ablations

Single-stream	Align Acc↑
$CA_t + CA_v$	87.61
$CA_t + PSCA_v$	**89.37**

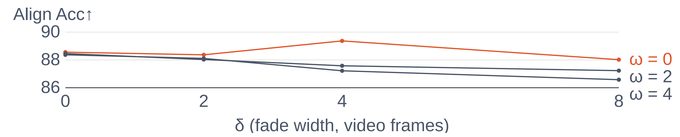
PSCA vs plain video CA: +1.76 Align Acc

Cross-modal	Align Acc↑
none	86.26
full design	**89.37**

Dual cross-attention: +3.11 Align Acc

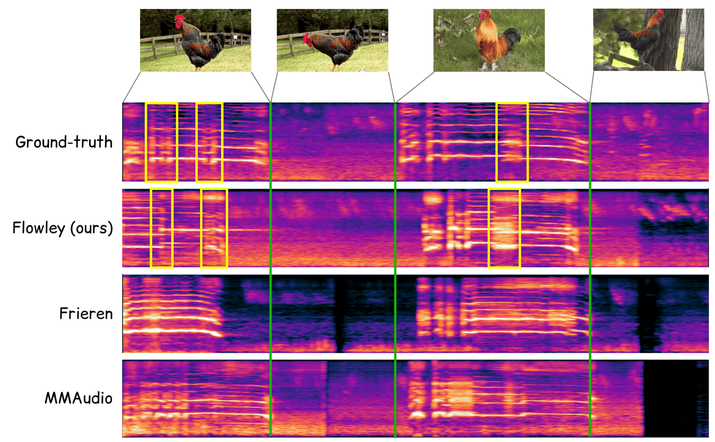
Progressive β	Align Acc↑
off	88.62
on	**89.37**

Progressive β: +0.75 Align Acc



Wider context windows cost synchronisation (89.37→ 86.58); IS and IB-Score stay flat.

8 Visualizations



Ground-truth vs. Flowley vs. Baselines: Only Flowley times both crows (yellow boxes)

Follow ①–⑧ through the poster.