Omni2Sound:

Towards Unified Video-Text-to-Audio Generation

CVPR 2026 Highlight ☆

Yusheng Dai[2,3], Zehua Chen[1,3†], Yuxuan Jiang[1,3]
Qiuhong Ke[2], Jianfei Cai[2], Jun Zhu[1,3†]

[1]Tsinghua University, Beijing, China [2]Monash University, Melbourne, Australia [3]Shengshu AI, Beijing, China

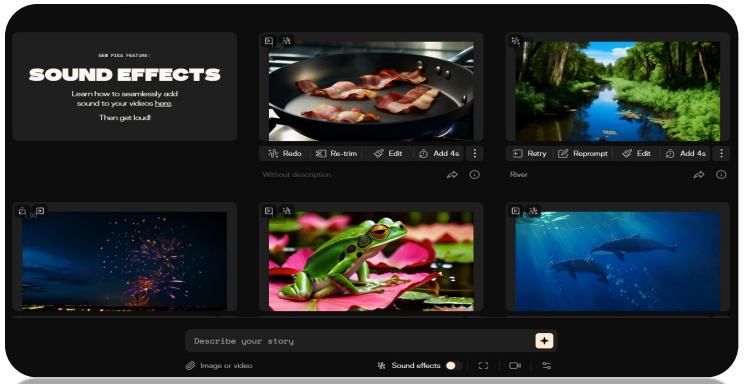
What & Why —— Audio Generation

Definition

Synthesizing non-speech audio (e.g. event-synchronized action audio, sound effects, background music)

Application

Audiobook



Movie Product

World Model

Scene-aware soundscapes that follow the story

Generative Foley, post-production sound effects

Diegetic, reactive audio for immersive environments

Generative audio is increasingly replacing handcrafted Foley, becoming a practical, low-cost boost to immersion.

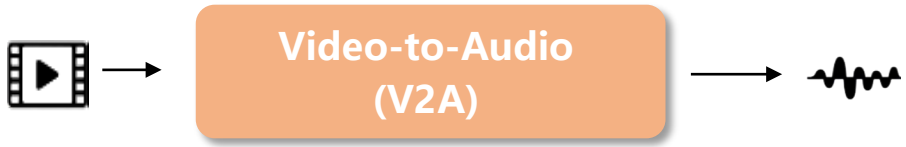
Technical evolution — T2A → V2A → VT2A

From unimodal to multimodal conditional control

2018

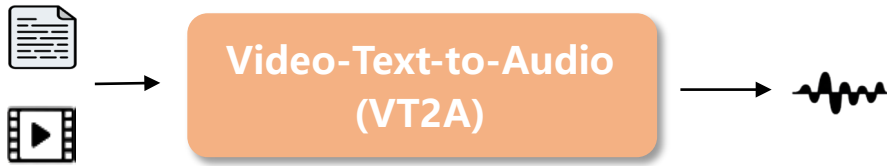
e.g. Tango, AudioLDM, AudioGen, Make-An-Audio

- ✓ Strong semantics
- ✓ High-Quality
- ✗ No temporal control



e.g. Diff-Foley, SpecVQGAN, FoleyGen, V2A-Mapper

- ✓ Tight temporal sync
- ✗ Weak reasoning
- ✗ hard to control; variable quality



e.g. HunyuanVideo-Foley , ThinkSound, ReasonAudio

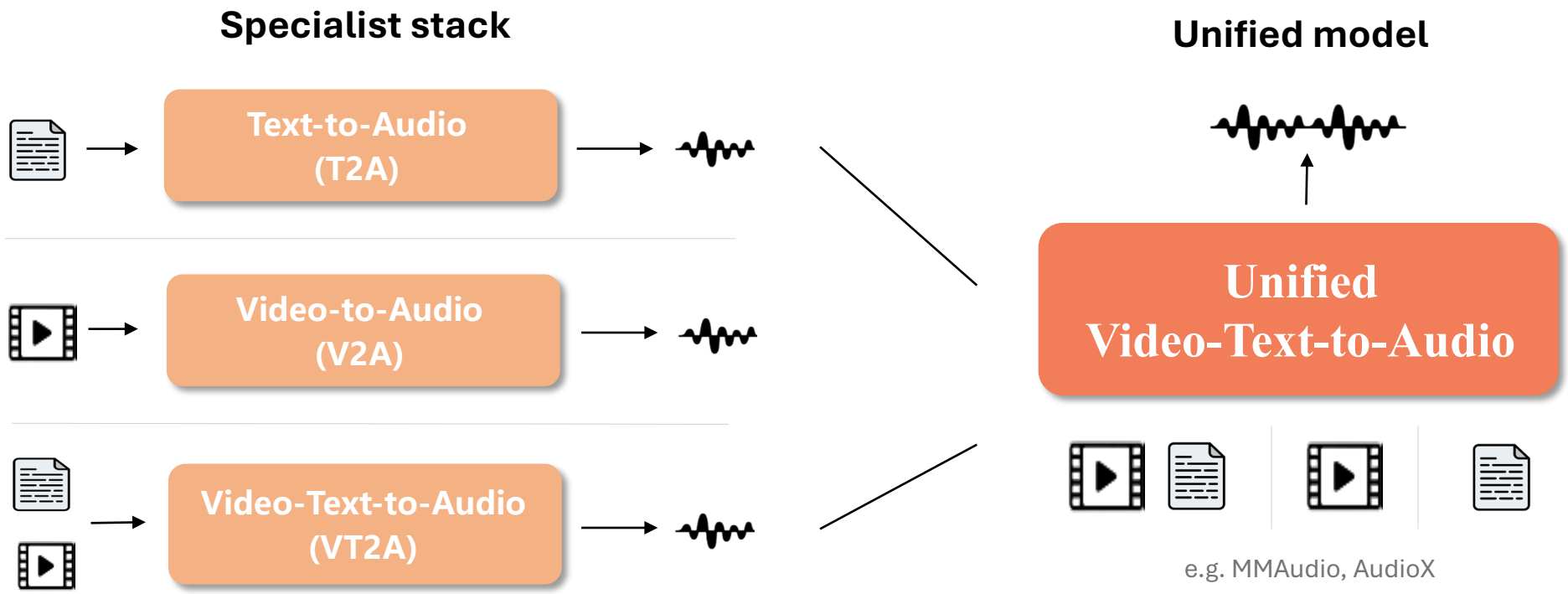
- ✓ Best of both
- ✗ Fragile to missing modality

Now

Technical evolution — Specialist to Generalist

From specialists toward a unified model

- *Unified VT2A Model: One backbone, all three input regimes — flexible inputs, no hard-switching.*



- ❑ Separate T2A · V2A · VT2A models
- ❑ Redundant architectures
- ❑ Hard-switching at deployment

- ❑ Single diffusion backbone
- ❑ Modality-flexible inputs (T / V / V+T)
- ❑ Aligned with the broader AIGC shift

Two fundamental challenges in Unified VT2A Model

➢ Data: **Lack of high-quality V-A-T aligned audio captions**

Audio captioning is a long-standing hard problem — far harder than image captioning

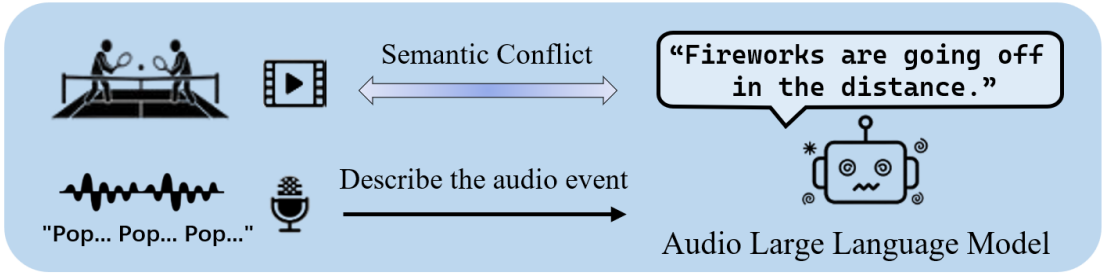
- *Audio events are inherently ambiguous*
- *Non-intuitive to describe in language*

➢ Pipeline ①: Audio-only captioning (ALM)

- **ALM hallucination + Audio ambiguity**:

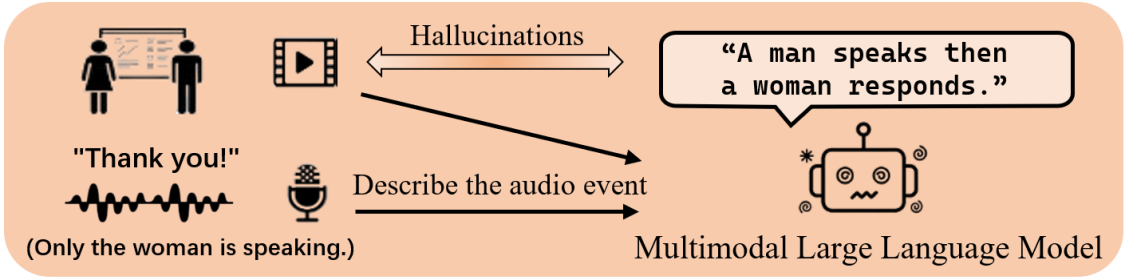
distant fireworks ≈ tennis hits; car engine ≈ electric drill

- **Severe semantic conflict between audio caption and video**



➢ Pipeline ②: Audio + Video captioning (MLLM)

- MLLMs carry strong visual bias *cause* **vision-induced hallucination**
- **Silent objects captioned as sound-emitting;**
- **Off-screen sources dropped**



Two fundamental challenges in Unified VT2A Model

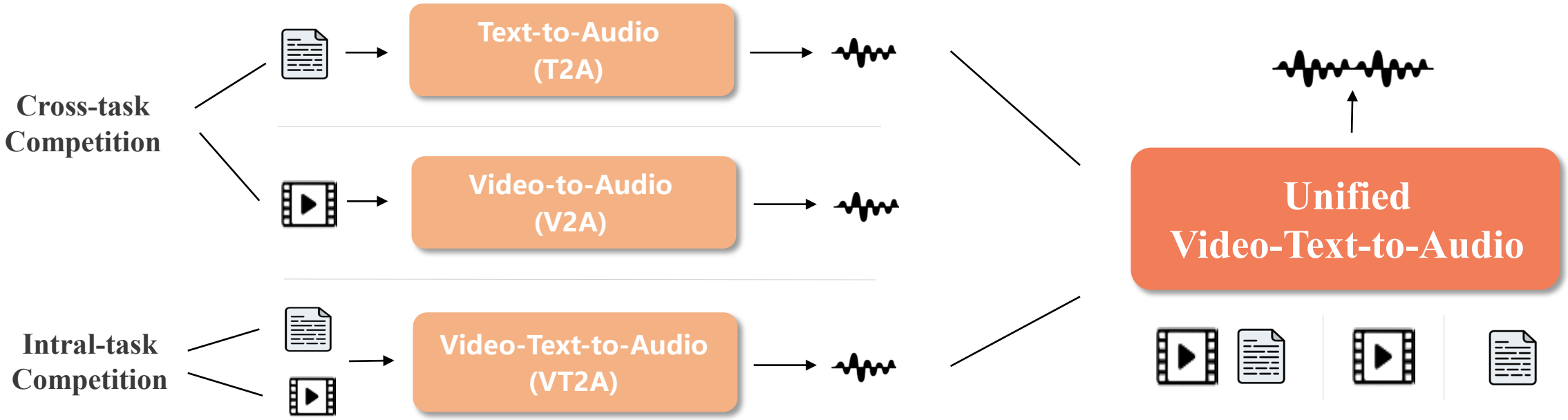
➢ Training: **Cross-task & intra-task competition under joint training**

➢ **Cross-task competition**

- Severe **V2A ↔ T2A adverse trade-off** under joint training
- A zero-sum dynamic rooted in text/video heterogeneity

➢ **Intra-task modality bias** (inside VT2A)

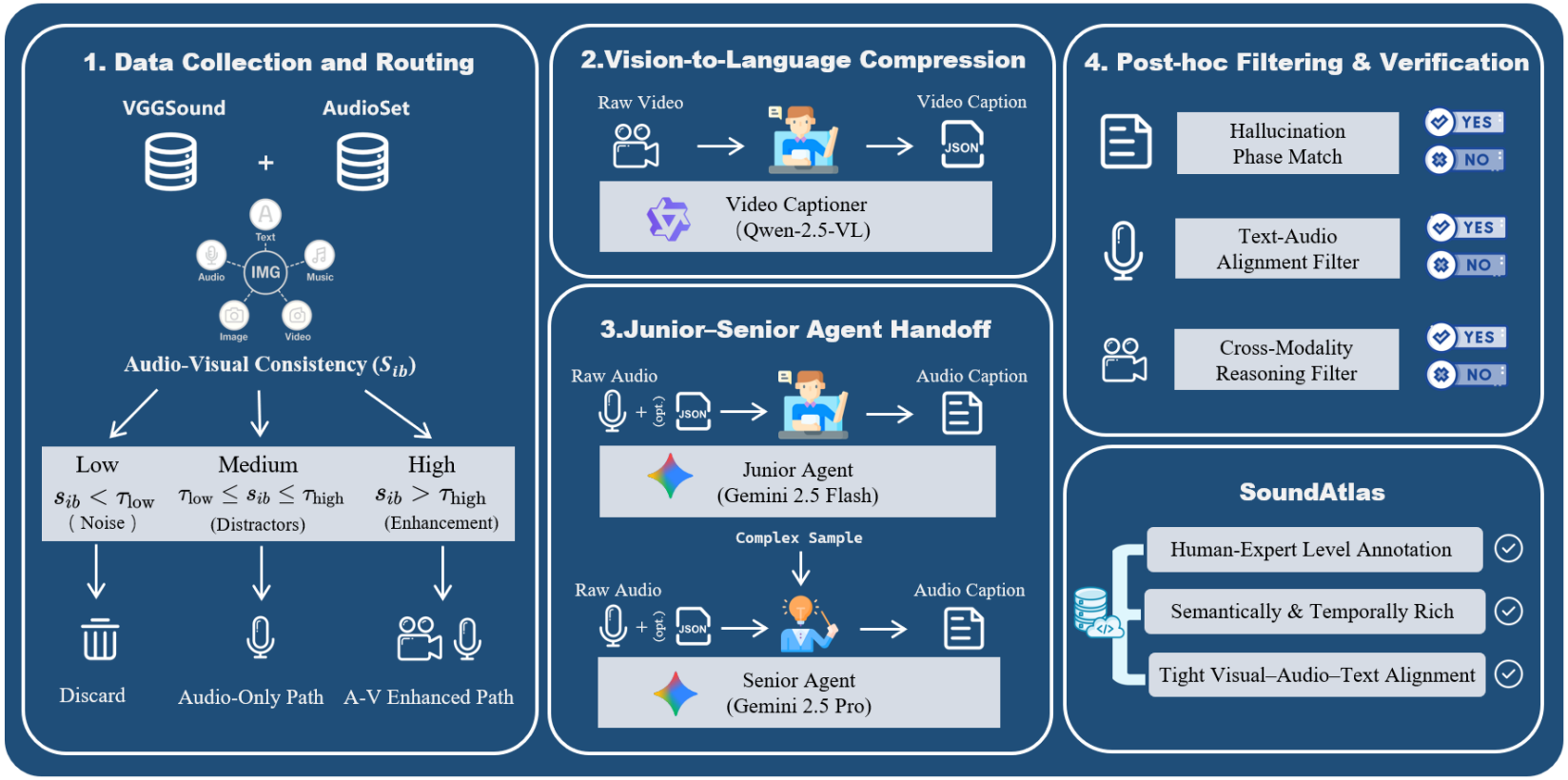
- text-bias → poor A-V synchronization
- video-bias → low text faithfulness in off-screen scenes



Without explicit scheduling, multi-objective optimization collapses one task to favor another.

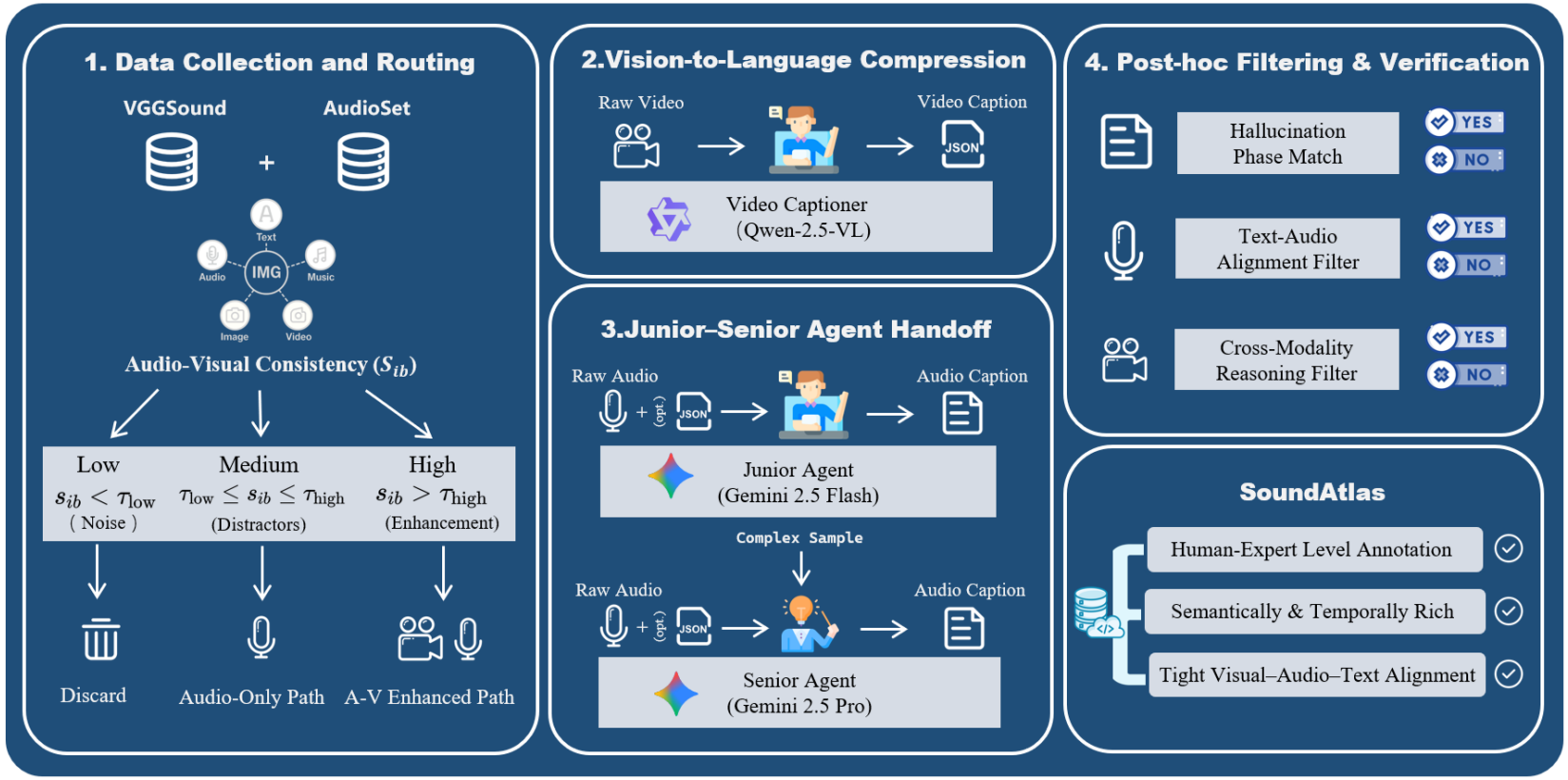
SoundAtlas: 470K high-quality V-A-T aligned captions

- *An agentic data-construction pipeline that* mitigates visual bias *and* halves annotation cost.
  - ➢ Vision-to-Language Compression
    - Qwen-2.5-VL distills video into compact text context
    - Captioning MLLM no longer sees raw frames → **mitigates visual bias**



SoundAtlas: 470K high-quality V-A-T aligned captions

- *An agentic data-construction pipeline that* mitigates visual bias *and* halves annotation cost.
  - ➢ Junior–Senior Agent Handoff → **5× Cost Reduction**
    - Junior (Gemini 2.5 Flash) handles common cases
    - Flagged samples (complexity + hallucination triggers + CLAP check) → Senior (Gemini 2.5 Pro)

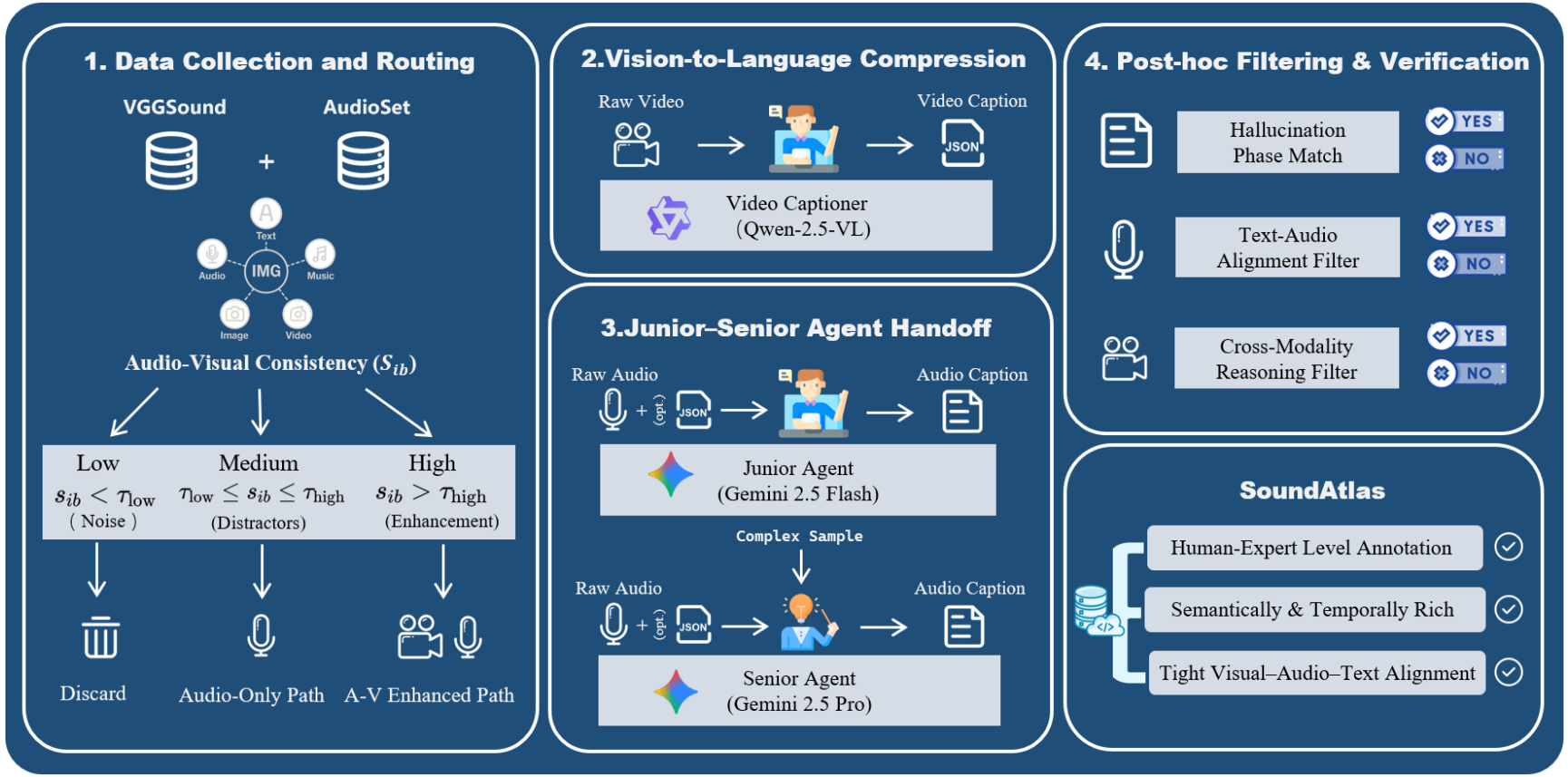


SoundAtlas: 470K high-quality V-A-T aligned captions

- *An agentic data-construction pipeline that* mitigates visual bias *and* halves annotation cost.

  - ➢ Post-hoc Hallucination Filtering

    - *Final fidelity guard that removes residual mismatches → tight V-A-T alignment.*

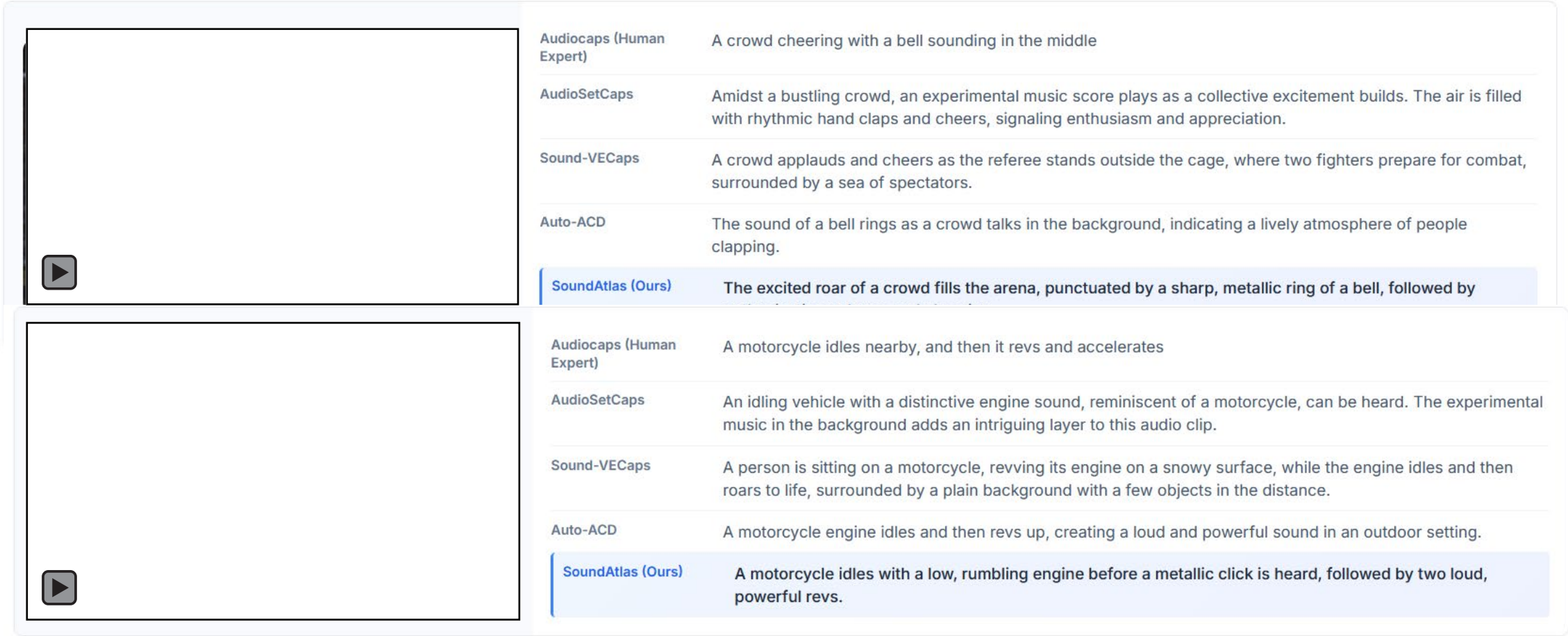


SoundAtlas: 470K high-quality V-A-T aligned captions

- Headline result:

  SoundAtlas quality > AudioCaps human-expert annotations

  on Mean Win Rate for Semantic alignment (MWR-S) and Temporal alignment (MWR-T)



Audiocaps (Human Expert)	A crowd cheering with a bell sounding in the middle
AudioSetCaps	Amidst a bustling crowd, an experimental music score plays as a collective excitement builds. The air is filled with rhythmic hand claps and cheers, signaling enthusiasm and appreciation.
Sound-VECaps	A crowd applauds and cheers as the referee stands outside the cage, where two fighters prepare for combat, surrounded by a sea of spectators.
Auto-ACD	The sound of a bell rings as a crowd talks in the background, indicating a lively atmosphere of people clapping.
SoundAtlas (Ours)	**The excited roar of a crowd fills the arena, punctuated by a sharp, metallic ring of a bell, followed by**

Audiocaps (Human Expert)	A motorcycle idles nearby, and then it revs and accelerates
AudioSetCaps	An idling vehicle with a distinctive engine sound, reminiscent of a motorcycle, can be heard. The experimental music in the background adds an intriguing layer to this audio clip.
Sound-VECaps	A person is sitting on a motorcycle, revving its engine on a snowy surface, while the engine idles and then roars to life, surrounded by a plain background with a few objects in the distance.
Auto-ACD	A motorcycle engine idles and then revs up, creating a loud and powerful sound in an outdoor setting.
SoundAtlas (Ours)	**A motorcycle idles with a low, rumbling engine before a metallic click is heard, followed by two loud, powerful revs.**

Training —— Three-stage progressive multi-task training

- Stage 1 — Large-scale T2A Pretraining

  Text-conditioned pretraining establishes a strong audio prior

  Later stages need only minimal high-quality T2A replay

- Stage 2 — Multi-task Interleaved Training

  Task-balanced sampling avoids in-batch gradient conflict

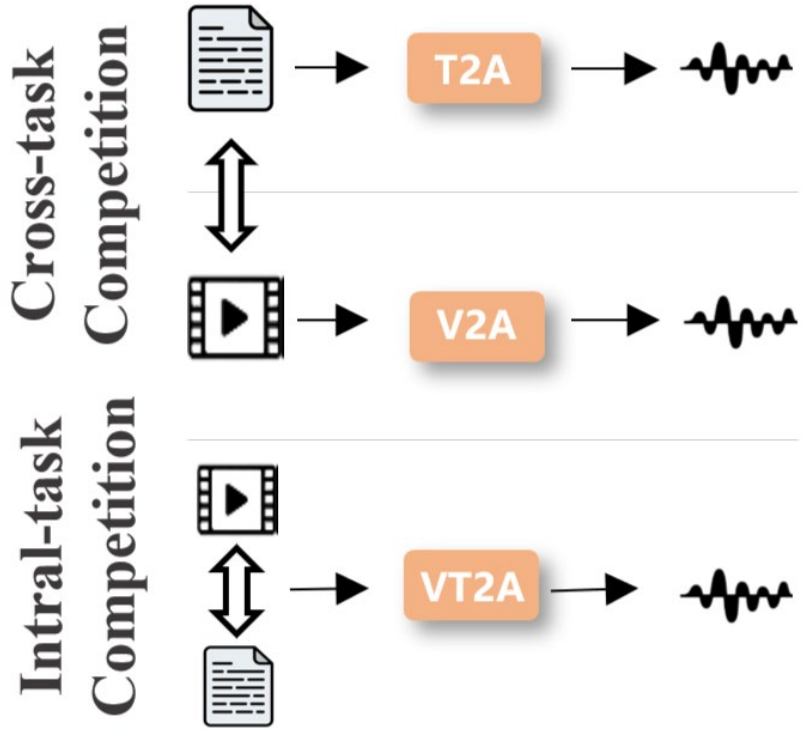
  High-quality VT2A acts as a semantic bridge — converts zero-sum cross-task competition into cooperative optimization

- Stage 3 — Decoupled Robustness Training

  Text Dropout: penalizes text bias → enhances A-V synchronization

  Off-screen Synthesis: counteracts video bias →

  ensures textual faithfulness in off-screen scenes



Model —— Omni2Sound Architecture

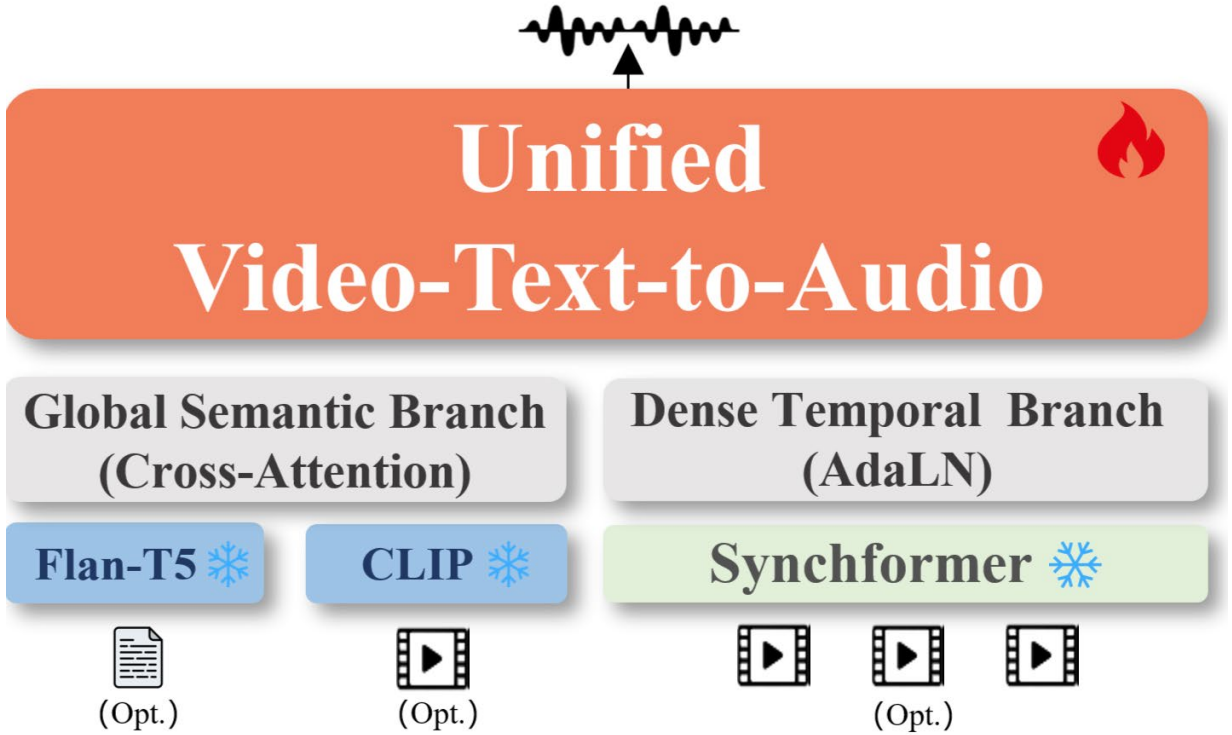
“Less is more ”—— a vanilla DiT with decoupled "What / When" conditioning

- Semantic Branch ("What")

Flan-T5 text + CLIP visual features concatenated → cross-attention (the high-level context)

- Temporal Branch ("When")

Synchformer features → global AdaLN (the fine-grained sync signal)



Benchmark —— VGGSound-Omni

The first comprehensive benchmark for unified VT2A, V2A, and T2A

- Standard tracks
  - **T2A · V2A · VT2A**
- Robustness tracks (new)
  - **Off-screen track** — VGGSound clips with low A-V correspondence

  (filtered by IB-Score & Desync-Score) + synthetic-music subset from MusicCaps

Experiments

Single checkpoint, evaluated on all three tasks. Omni2Sound surpasses both unified and specialist baselines

- Performance on VT2A & V2A & T2A
  - Leads on **distribution matching** & **Semantic alignment** & **Temporal alignment**

Table 3. Comparison on VGGSound-Omni benchmark: Omni2Sound against SOTA models on T2A, V2A, and VT2A tasks. The *w/ Video-LLaMA caps* row evaluates Omni2Sound's generalization to unseen captions generated by Video-LLaMA [37].

Task	Method	Distribution Matching				Audio Quality		Modality Alignment		
		KL↓	FD↓	FAD↓	FD_{PaSST}↓	PQ↑	IS↑	DS↓	IB↑	MS-CLAP↑
T2A	AudioX [14]	1.68	9.04	1.42	109.94	6.37	15.15	-	-	0.49
	MMAudio [13]	1.92	8.62	1.63	101.66	5.84	14.30	-	-	0.50
	Omni2Sound (ours)	**1.53**	**4.61**	**1.01**	**60.38**	**6.52**	**16.41**	-	-	**0.53**
	w/ Video-LLaMA caps	1.60	6.92	1.23	83.91	6.38	16.01	-	-	0.51
V2A	V-AURA [38]	2.28	16.43	2.34	245.25	5.74	10.82	0.69	0.28	0.32
	Frieren [6]	2.73	12.13	1.23	123.75	5.82	11.32	0.86	0.21	0.31
	AudioX [14]	2.96	12.73	1.42	121.82	**6.17**	13.34	1.22	0.24	0.34
	MMAudio [13]	2.11	5.65	0.81	69.33	5.72	11.85	0.48	0.28	0.43
	Omni2Sound (ours)	**2.04**	**3.41**	**0.51**	**50.19**	6.15	**16.18**	**0.47**	**0.35**	**0.44**
VT2A	ThinkSound (*w/o.* CoT) [12]	1.60	7.41	1.10	116.08	**6.21**	11.73	0.53	0.26	0.43
	HunyuanVideo-Foley [11]	1.74	10.02	2.36	100.53	6.18	11.58	0.57	0.32	0.45
	AudioX [14]	1.59	8.29	1.24	103.37	6.17	14.94	1.23	0.26	0.49
	MMAudio [13]	1.63	5.28	0.91	68.44	5.84	13.44	**0.49**	0.29	0.49
	Omni2Sound (ours)	**1.35**	**2.95**	**0.53**	**48.20**	6.21	**15.79**	0.49	**0.34**	**0.52**
	w/ Video-LLaMA caps	1.56	3.37	0.66	53.73	6.11	15.74	0.50	0.34	0.49

Experiments

Single checkpoint, evaluated on all three tasks. Omni2Sound surpasses both unified and specialist baselines

➢ Robustness & Generalization

- Holds on **low-quality text input** (Video-LLaMa)
- Holds on **Kling-Audio-Eval** (out-of-domain, professional video)

Table 4. Comparison on the Kling-Audio-Eval: Omni2Sound against SOTA models on T2A, V2A, and VT2A tasks.

Task	Method	Distribution Matching				Audio Quality		Modality Alignment		
		KL↓	FD↓	FAD↓	FD_{PaSST}↓	PQ↑	IS↑	DS↓	IB↑	LA-CLAP↑
T2A	AudioX [14]	2.73	19.43	3.32	171.60	5.98	**12.15**	-	-	0.28
	MMAudio [13]	2.54	**11.25**	5.07	**142.71**	5.54	9.28	-	-	0.28
	Omni2Sound (ours)	**2.36**	11.59	**2.62**	147.46	**6.26**	11.27	-	-	**0.28**
V2A	AudioX [14]	3.13	18.90	4.01	205.48	**5.87**	8.31	1.20	0.23	0.13
	MMAudio [13]	2.94	13.41	3.87	159.30	5.50	7.59	0.62	0.24	0.14
	Omni2Sound (ours)	**2.47**	**8.78**	**2.55**	**112.21**	5.78	**8.56**	**0.57**	**0.34**	**0.18**
VT2A	ThinkSound ($w/o.$ CoT) [12]	2.53	11.99	3.52	206.93	5.77	6.09	0.66	0.22	0.19
	HunyuanVideo-Foley [11]	2.13	8.06	3.58	**94.64**	**6.04**	8.17	**0.55**	**0.34**	0.23
	AudioX [14]	2.39	14.26	3.16	149.37	5.97	**10.23**	1.21	0.23	0.26
	MMAudio [13]	2.41	10.12	4.90	129.21	5.53	7.46	0.59	0.25	0.20
	Omni2Sound (ours)	**2.10**	**7.60**	**2.37**	106.55	5.98	8.22	0.58	0.32	**0.26**

Three observations worth highlighting

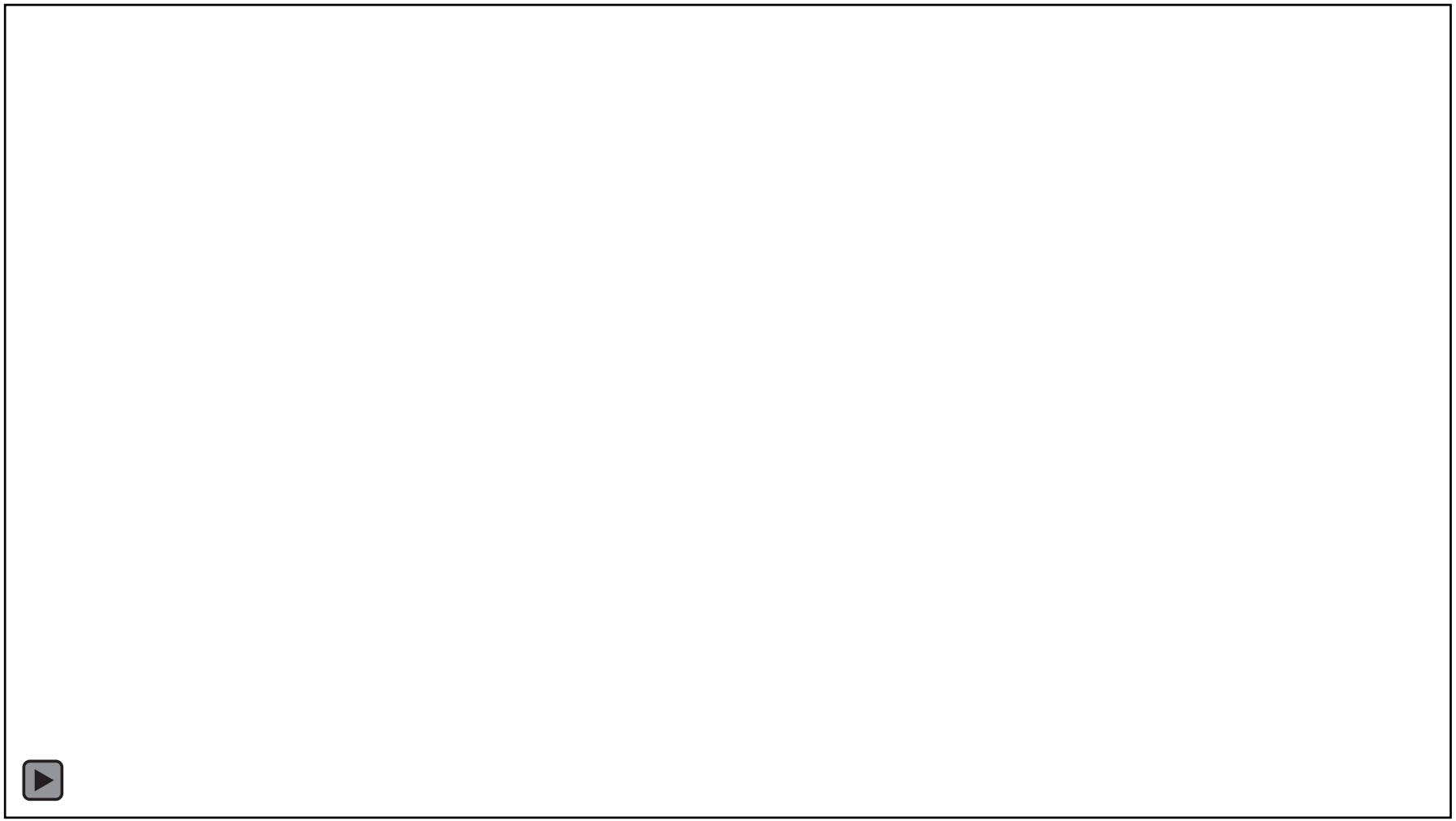
- VT2A as a semantic bridge
  - High-quality VT2A data converts cross-task competition into co-adaptation
- Caption quality > caption quantity
  - Small clean SoundAtlas data outperforms large noisy automated data
- Text dropout helps temporal sync
  - Masking text forces the model to attend to visual rhythm

What we want this work to argue

- The barrier to unified audio generation is alignment supervision and training dynamics, not architecture.
- Data + curriculum can make a vanilla DiT **outperform specialized expert designs**.

Less is More — Data & Strategy is all you need.

Demos



VT2A · T2A · V2A · Off-screen Task

No task-specific finetuning. No per-sample tuning.

✨ Summarize Highlights

—— A Unified VTA Toolkit with Model, Caption, Benchmark and Code

• 🏆 **Omni2Sound** — A Unified Audio Generation Foundation Model, supporting flexible input modalities with VT2A and V2A and T2A task in one model.

• 📀 **SoundAtlas** — The first large-scale, high-quality V-A-T aligned audio captions, **surpassing even human-expert annotation quality**

• 🧪 **VGGSound-Omni** — A unified VTA evaluation benchmark with a challenging track specifically for off-screen audio and background music generation.

• 🔓 **Open-source friendly** — A multimodal extension of stable-audio-tools, a good choice if you are searching for a multimodal audio generation codebase.

Everything is open source

- **Paper** — https://arxiv.org/pdf/2601.02731
- **Project** — https://omni2sound.github.io
- **Code** — https://github.com/omni2sound/Omni2Sound
- **Models & Benchmark** — https://huggingface.co/Dalision