Weaving Time, Space & Semantics:

Multimodal Alignment for Audio-Visual Generation

Danilo Comminiello

Sapienza
Università di Roma

1° Workshop on Generative AI for Audio-Visual Content Creation
International Conference on Computer Vision – ICCV 2025
19 October 2025, Honolulu, HI

Weaving signals into meaning: a multimodal loom

Figure 1. John William Waterhouse — Penelope and the Suitors (1912). Source: Wikipedia.

Outline

1 | Why Multimodal Alignment

2 | When: Temporal Alignment

3 | What: Semantic Alignment

4 | Where: Spatial Alignment

5 | Toward Joint AV Generation

6 | Conclusion

1 | Why Multimodal Alignment

Audio-visual generation tasks

Several **audio-visual generation tasks** can be in many fields of applications.

- **Realistic Sound Synthesis:** enabling high-quality, contextually relevant audio from silent video.
- **Cross-Domain Multimodal Synthesis**: connecting and leveraging visual and audio modalities.
- **Entertainment Applications:** advancing sound design for movies, video games, and virtual reality.
- **Assistive Technologies:** supporting accessibility tools to improve media inclusivity.
- **Physics-Informed Synthesis:** modeling acoustics and physical interactions from visual cues.
- **Scene Generation:** preventing dangerous events and improving safety and security.

Visual gen surged, audio lagged

- **Temporal is unforgiving.**

  Even small timing errors ruin realism; even advanced gen models struggle to keep precise sync.

- **Pairwise semantics are not enough.**

  CLIP/CLAP-style pairwise anchors do not guarantee that audio–video–text are jointly aligned.

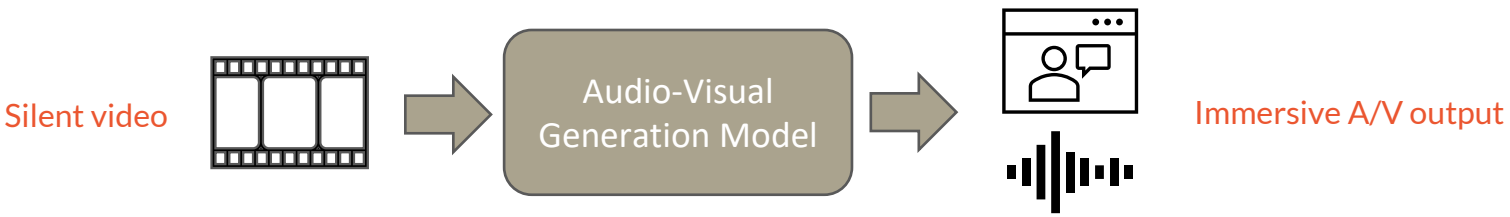
- **Space is often missing.**

  Most AV generative models stay mono or treat stereo as a post-effect.

- **Practical limitations.**

  Large paired AV datasets are scarce/expensive; heavy joint training is brittle.

Temporal, semantic and spatial alignment



In audio-visual generation, **three key aspects** may ensure a natural audio experience:

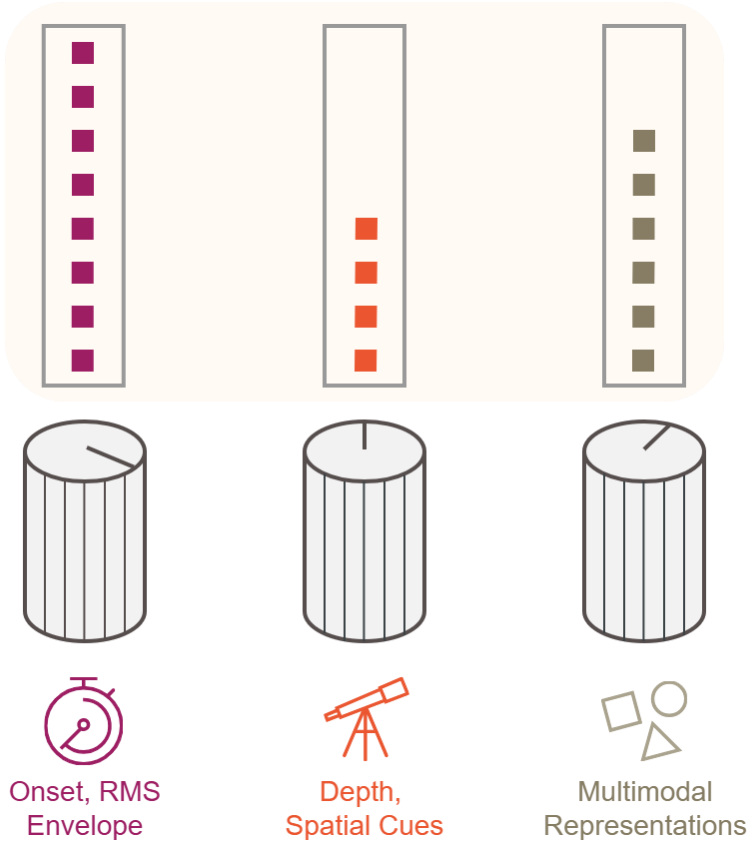
- **Temporal Alignment:** The timing of the audio must precisely match the visual events, ensuring synchronization between actions and sounds.
- **Semantic Alignment:** generated sounds must be contextually appropriate and represent events, objects, and actions.
- **Spatial Alignment:** The sound should reflect the spatial properties of the scene, including directionality, distance, and environmental acoustics.

AV creators need readable controls

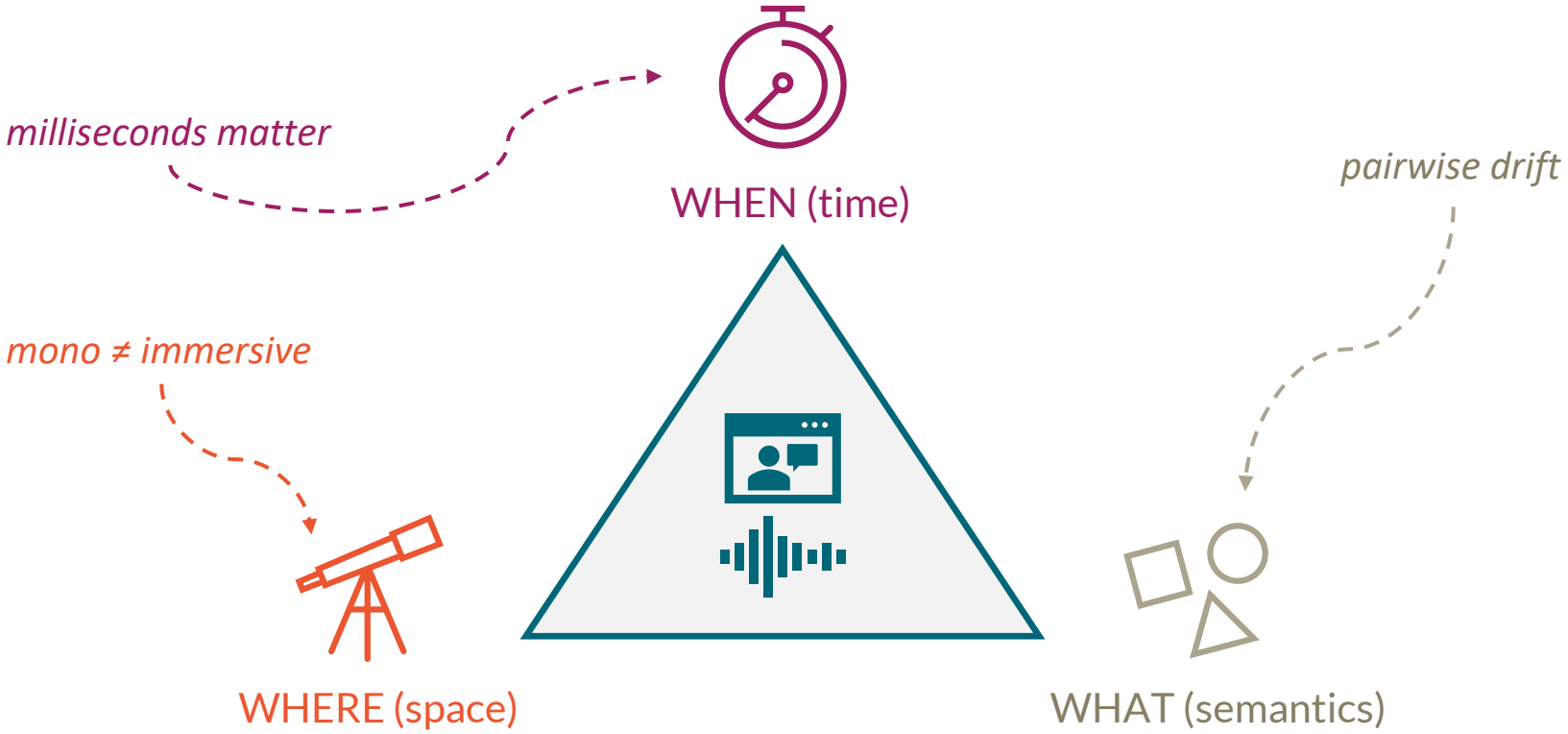
Generative AV is maturing but content creators need **readable controls**:

- **When (time):** onset, RMS envelope
- **Where (space):** depth, spatial cues
- **What (semantics):** multimodal embeddings, meaningful sounds

Alignment is the interface to controllable generation, raising creative quality.



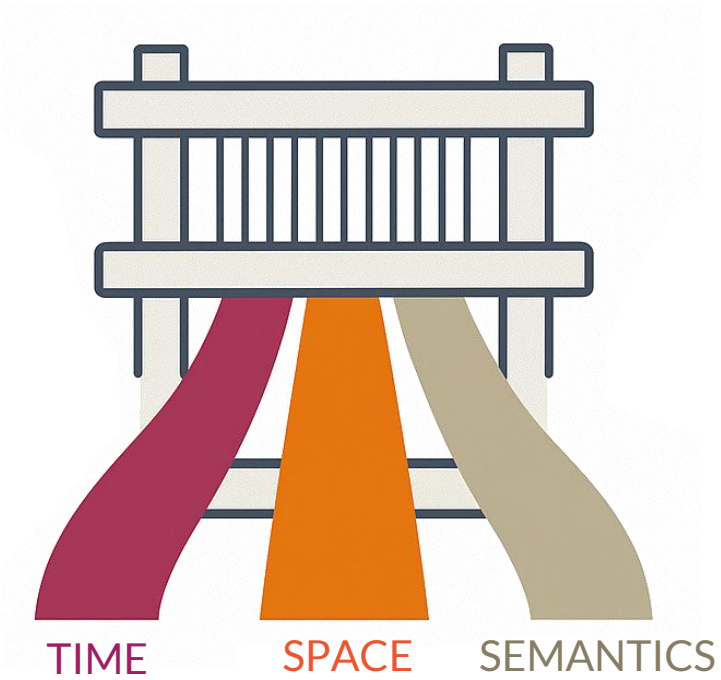
When, where, what



The three axes at a glance

Give models *when, where, what* as dials, then they act like collaborators.

- **WHEN (time):** simple, editable signals for temporal synchronization.
- **WHERE (space):** visual geometry for immersivity.
- **WHAT (semantics):** unified embeddings for meaningful sounds.



2 | When: Temporal Alignment

Why timing control matters

Even small timing errors (tens of ms) break immersion for Foley sound design.

Creators need timing they can see and edit.

We will move from discrete onsets to continuous envelopes as human-readable controls (HRC).

Video 1. Play at this link. Timing sync affects perception. Sample taken from the Diff-Foley project page.

[1] S. Luo, C. Yan, C. Hu, H. Zhao, Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models, *Advances in Neural Information Processing Systems* (NeurIPS), vol. 36, pp. 48855–48876, 2023.

[2] M. Comunità, R. F. Gramaccioni, E. Postolache, E. Rodolà, D. Comminiello, J. D. Reiss, SyncFusion: Multimodal Onset-Synchronized Video-to-Audio Foley Synthesis, *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 936–940, 2024.

SyncFusion: Onset track as a human-readable control

The **Onset Model** extracts the onsets for the actions in a video.

The **Diffusion Model** conditioned by the **onsets track** and a **CLAP embedding**, generates a synchronized audio that is used as *soundtrack* for the input video.

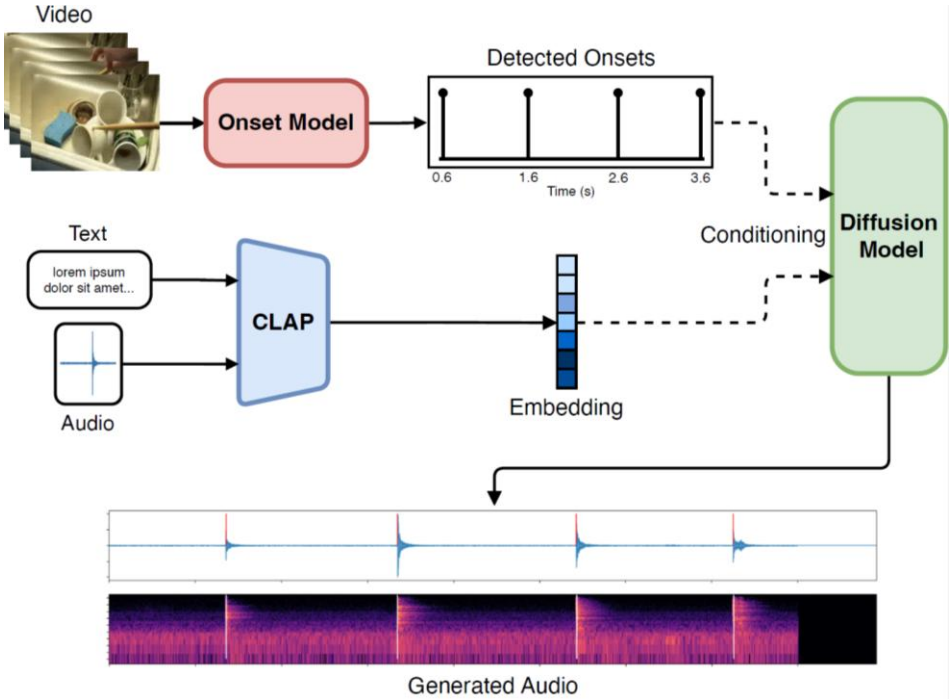


Figure 2. The SyncFusion architecture [2].

[2] M. Comunità, R. F. Gramaccioni, E. Postolache, E. Rodolà, D. Comminiello, J. D. Reiss, SyncFusion: Multimodal Onset-Synchronized Video-to-Audio Foley Synthesis, *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 936–940, 2024.

SyncFusion: metrics and results

The onset synchronization is evaluated via the Onset Accuracy (wrt the number of onsets detected correctly) and the Average Precision of the detected onsets on a confidence interval of 50ms.

The generated audio track quality is evaluated through the Frèchet Audio Distance (FAD).

	Params	Loss BCE	#Onset Acc. (%)	Onset Sync AP (%)
w/out augm.	31M	3.79	43.07	87.71
w/ augm.	31M	**2.64**	**49.39**	**88.83**

Table 1. Onset detection model evaluation [2].

Modality	Params	#Onset Acc. (%)	Onset Sync AP (%)	FAD
Text	215M	84.38	98.12	1.68
Audio+Text	215M	**89.38**	**98.75**	**1.48**

Table 2. Synthesis diffusion model evaluation [2].

Model	Params	#Onset Acc. (%)	Onset Sync AP (%)	FAD
CondFoleyGen [4]	408M	23.94	62.44	6.10
SyncFusion (w/ augm.) [2]	246M	49.38	79.11	5.38
SyncFusion (w/out augm.) [2]	**246M**	**56.87**	**84.37**	**5.50**

Table 3. Complete model evaluation [2].

SyncFusion: generation results

Discrete onsets give precise “when” and a simple editing surface.

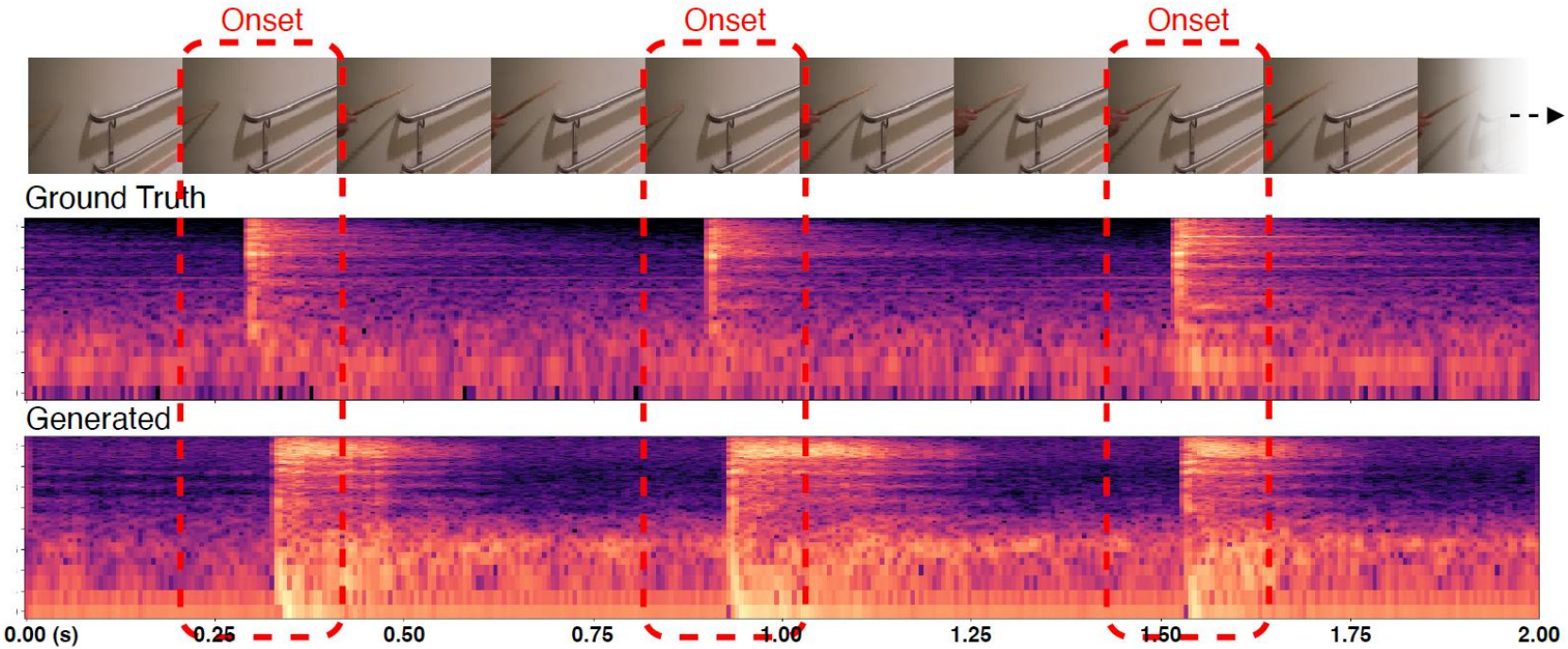


Figure 3. Example showing ground truth audio and video, detected onsets and generated audio [2].

From onsets to envelopes: why a continuous control?

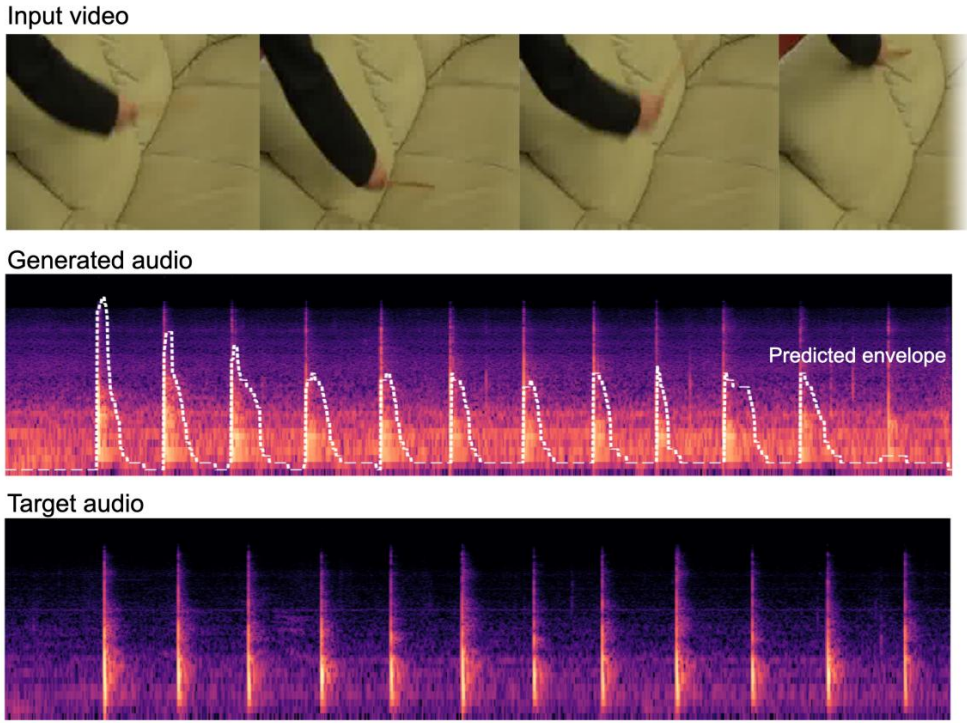


Figure 4. Ground truth, predicted RMS envelope and generated audio [3].

The envelope of an oscillating signal is a smooth curve outlining its extremes.

Onset: where events happen

Envelope: how they grow/decay

Envelopes are richer timing controls (e.g., intensity, duration).

[3] R. F. Gramaccioni, C. Marinoni, E. Postolache, M. Comunità, L. Cosmo, J. D. Reiss, D. Comminiello, FOL·AI: Synchronized Foley Sound Generation with Semantic and Temporal Alignment, *arXiv preprint arXiv:2412.15023v3*, May 2025.

FOL·AI: Two-stage temporal control with RMS envelope

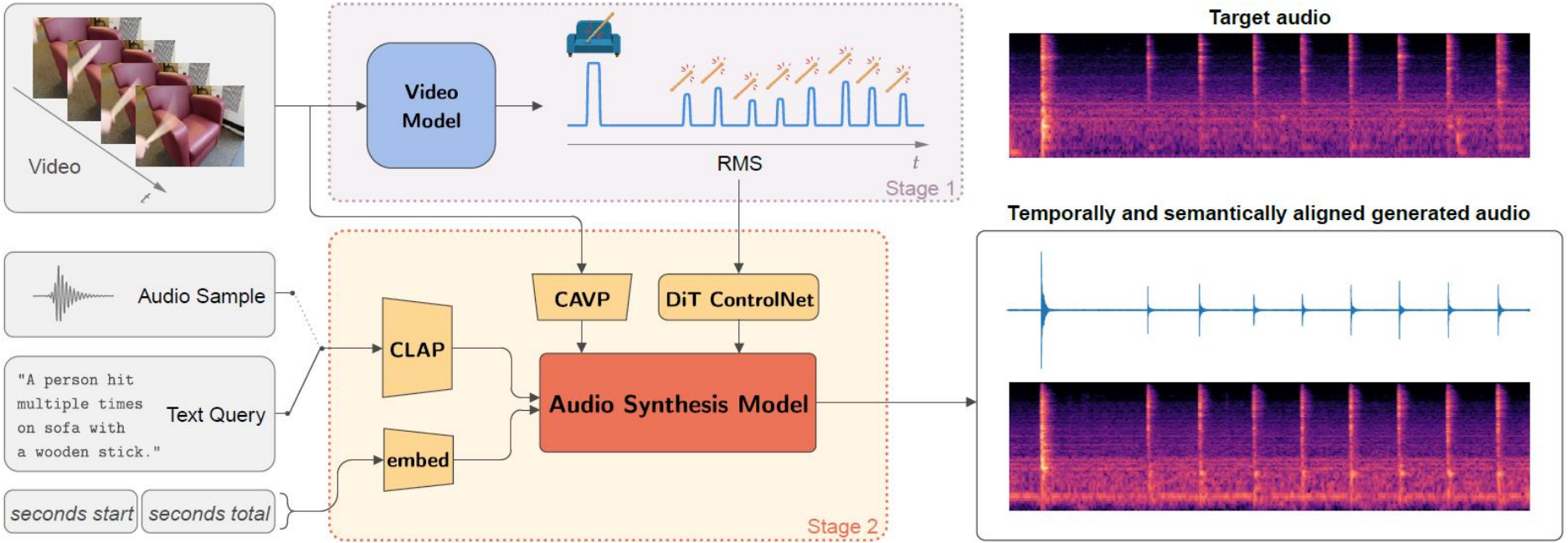
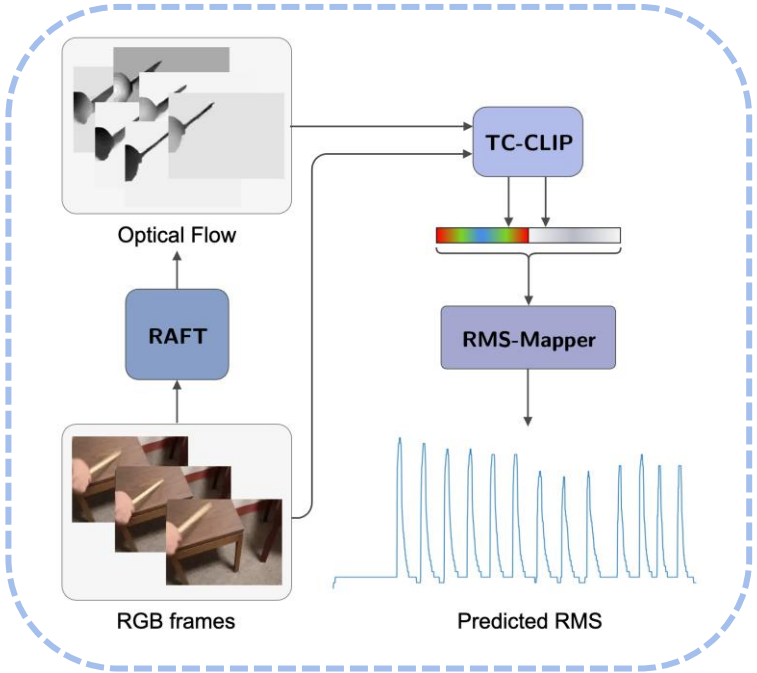
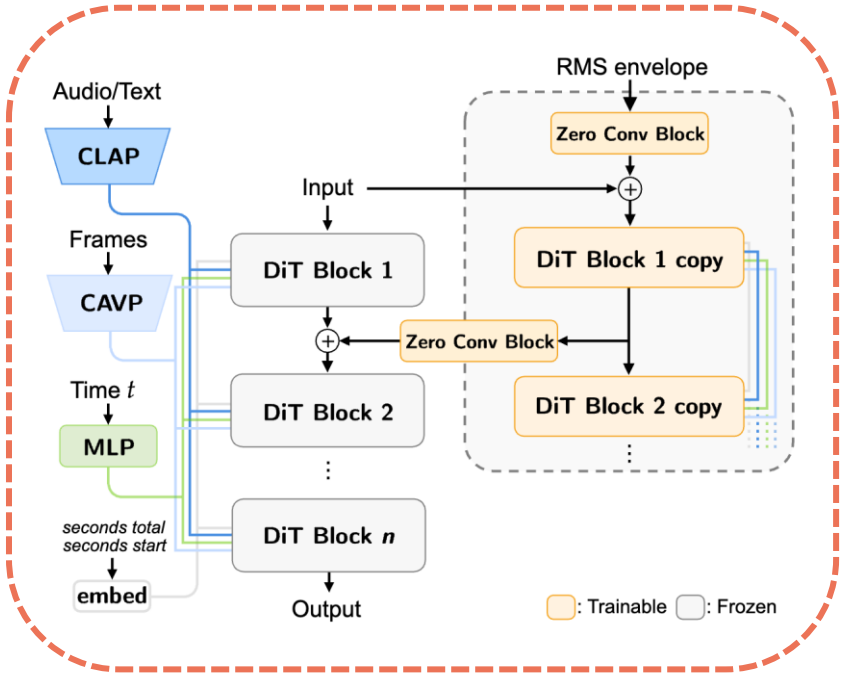


Figure 5. FOL·AI architecture consists of two distinct parts: the video model, that predicts an envelope representative for the audio directly from the input video, and the audio synthesis model for the controlled generation of the final audio effect. The generation is controlled temporally by the predicted RMS envelope through a DiT ControlNet, and semantically by CLAP and CAVP embeddings. The length of the output waveform can be controlled with the parameters *seconds start* and *seconds total*. [3].

FOL·AI: video and audio models





Video Model

Audio Model

Figure 6. In the Video Model, the RMS-Mapper predicts an envelope directly from the input video. The Audio Synthesis Model controls the generation of the final audio effect. The generation is controlled temporally by the predicted RMS envelope through a DiT ControlNet [3].

FOL·AI: temporal metrics

Temporal alignment: the **Envelope-L1 (E-L1)** metric evaluates the fitting of the generated sounds to the temporal condition of the event:

$$\mathrm{E}{-}L1 = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{r}_i - \hat{\mathbf{r}}_i\|$$

where $\mathbf{r}_i$ is the ground-truth envelope of the i-th frame, and $\hat{\mathbf{r}}_i$ is the predicted one.

Other metrics: accuracy metric, FAD (with 3 different audio encoders), CLAP-score.

[5] Y. Chung, J. Lee, J. Nam, T-Foley: A Controllable Waveform-Domain Diffusion Model for Temporal-Event-Guided Foley Sound Synthesis, *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 6820–6824 ,2024.

FOL·AI: datasets

Greatest Hits (10s chunks, curated hits/scratches) ↓

Walking The Maps for footsteps (893 clips; high-quality AV) →

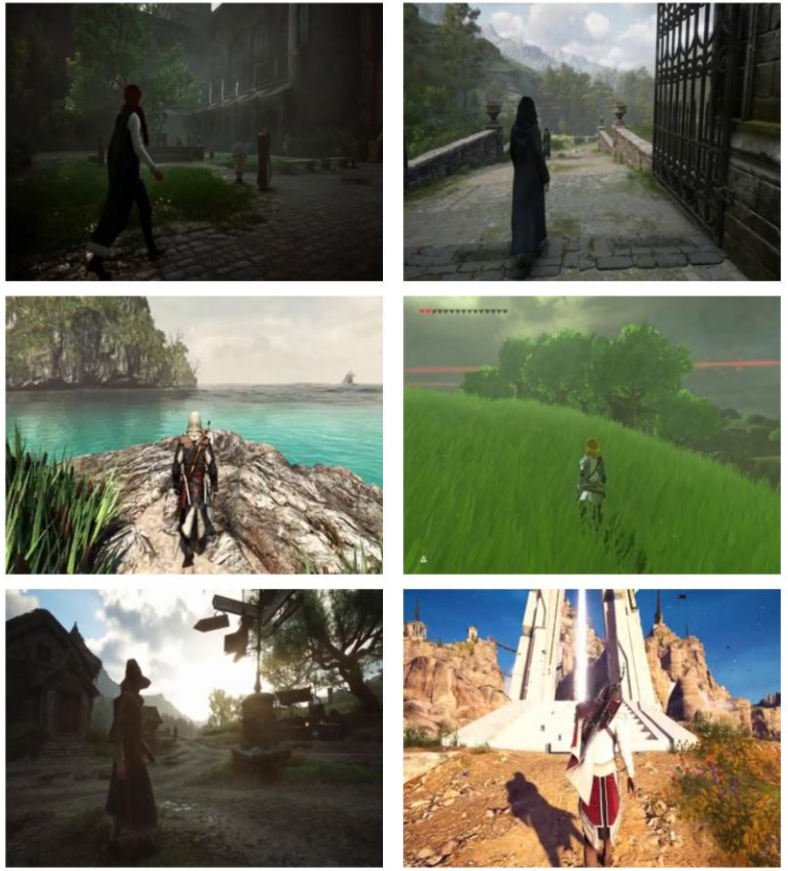


Figure 7. Samples from the adopted datasets: Greatest Hits (above) and Walking The Maps (on the right) [3].

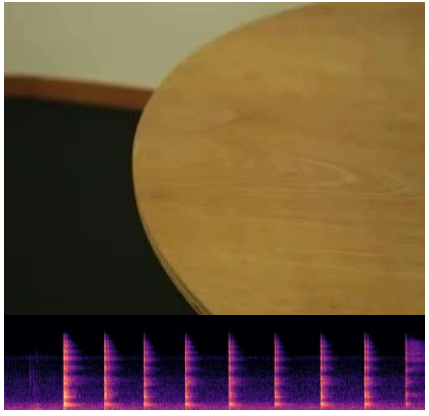
FOL·AI: Foley sound synthesis results on Greatest Hits

Model	Cond.	HRC	FAD-P ↓	FAD-C ↓	FAD-LC ↓	E-L1 ↓	CLAP ↑	FAVD ↓
SpecVQGAN [6]	None	NO	99.07	1001	0.710	0.043	0.142	6.513
Diff-Foley [1]	None	NO	85.70	654	0.469	0.045	0.373	4.618
CondFoleyGen [4]	Audio	NO	74.93	650	0.488	0.035	0.488	6.481
Video-Foley [7]	Text	YES	67.04	644	0.499	0.024	0.368	4.910
Video-Foley [7]	Audio	YES	28.45	435	0.167	0.018	0.678	2.207
SyncFusion [2]	Text	YES	35.64	591	0.436	0.023	0.515	4.302
SyncFusion [2]	Audio	YES	27.85	542	0.279	0.017	0.662	3.282
Fol·AI [3]	Text	YES	32.80	381	0.251	0.013	0.480	3.941
Fol·AI [3]	Audio	YES	**16.57**	**217**	**0.105**	**0.013**	**0.683**	**2.026**

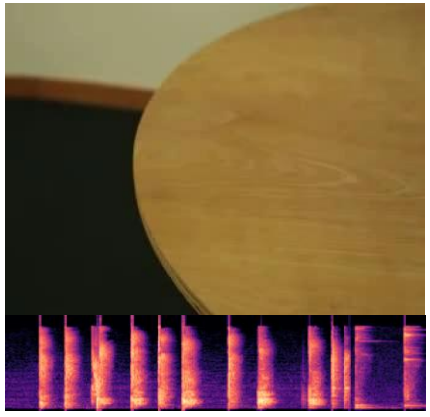
Table 4. Results for **FOL·AI** and comparison with other SOTA models on Greatest Hits. Table shows whether the model generates the output conditioned on audio or text prompt; HRC stands for Human Readable Control and refers to the use of time-varying interpretable signals that sound designers can use to control the generation process (i.e., envelope or onsets). FOL·AI provides the best results, even in text-conditioning generation [3].

FOL·AI: Foley sound synthesis results on Greatest Hits

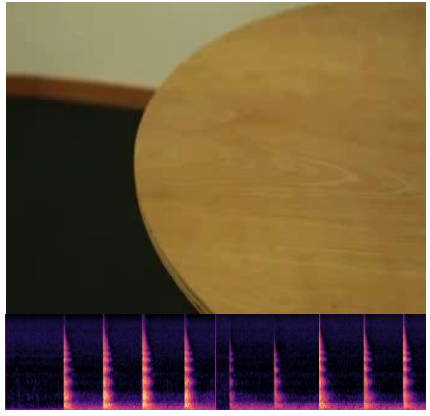
Video 2. FOL·AI synthesis results. These and other samples can be found on the FOL·AI project page.



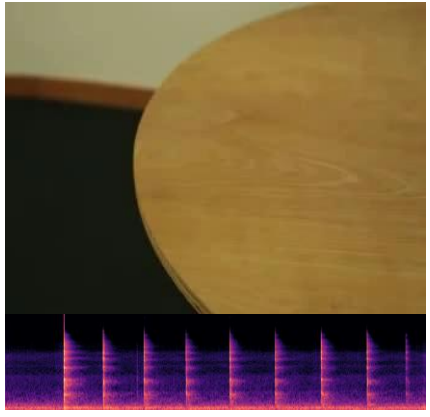
Ground Truth



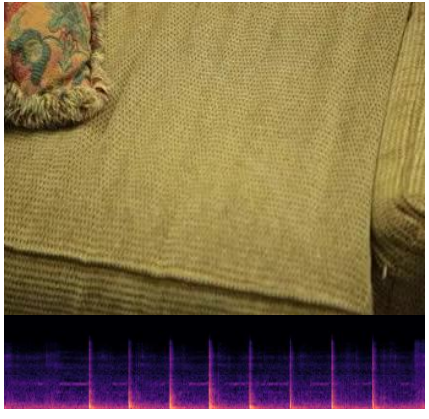
Diff-Foley



SyncFusion



Fol·AI



FOL·AI: Foley synthesis results on Walking the Maps

FAD-C ↓	FAD-LC ↓	E-L1 ↓	CLAP ↑	FAVD ↓
167	0.255	0.046	0.480	3.068

Table 5. Results for **FOL·AI** on Walking The Maps [3].

Results highlight that both the video model and the audio model succeed to obtain good performance for the synchronization between audio and video as well as for the quality of the produced waveforms, managing to generate realistic footstep sounds.

The sounds produced are diversified from each other according to both the character's walking style and the ground type.

FOL·AI: Foley synthesis results on Walking the Maps

Ground Truth

Baseline model

Fol·AI

Video 3. FOL·AI synthesis results on Walking The Maps. These and other samples can be found on the FOL·AI project page.

3 | WHAT: SEMANTIC ALIGNMENT

Multimodal representation learning

The goal of multimodal representation learning is to extract information from different sources to create a unified representation that captures underlying relationships and complementary information across the modalities.

Figure 8. Multimodal representation learning focused on learning meaningful representations from data that comes from multiple modalities.

Main issues in multimodal learning

Cause: contrastive learning in MRL is based on cosine similarity, which is not formally definite for more than two vectors.

***De-Facto* Standard:** it aligns two modalities at a time.

Problem: What if we choose an anchor and align more modalities in a pairwise fashion?

- No geometrical guarantees;
- Weak alignment between two generic modalities;
- Weak model explainability.

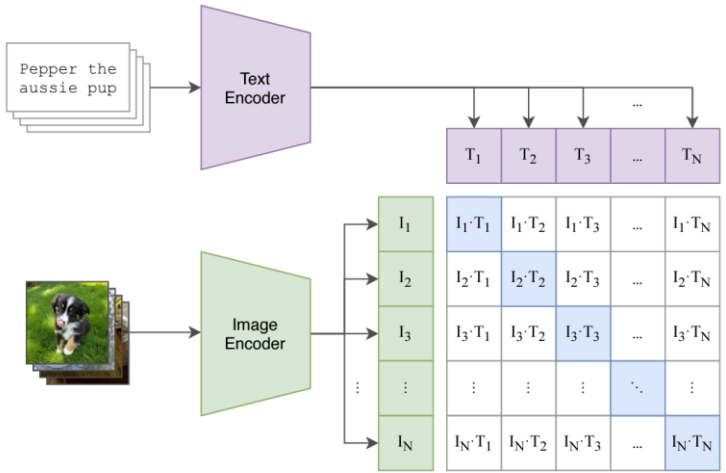


Figure 9. The conventional CLIP approach [8].

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, Learning Transferable Visual Models from Natural Language Supervision, *International Conference on Machine Learning* (ICML), 2021.

How to deal with more modalities consistently?

Main idea: an embedding of a single modality can be seen as a point in a hyperdimensional space $\mathbb{R}^N$.

If we have more than two modalities, they altogether semantically describe a shared geometrical entity in $\mathbb{R}^N$.

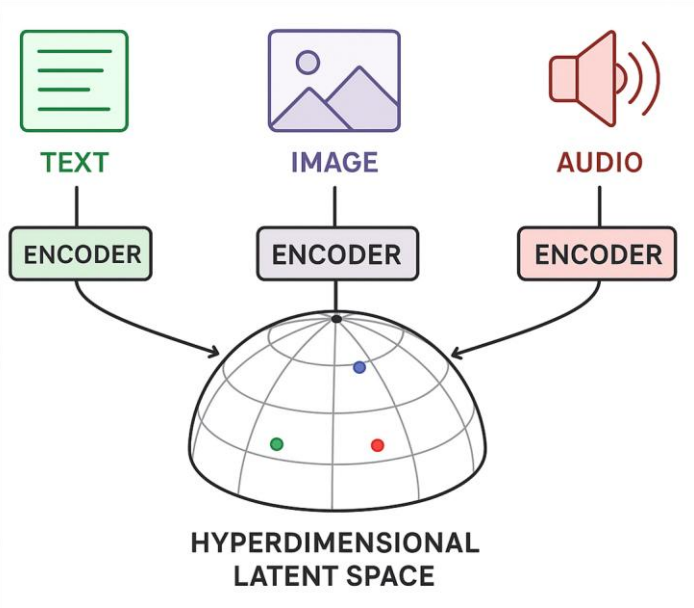


Figure 10. Learning representation in a joint hyperdimensional space. We want an embedding space where audio–video–text share a common semantic geometry.

GRAM intuition: semantics as a *volume*

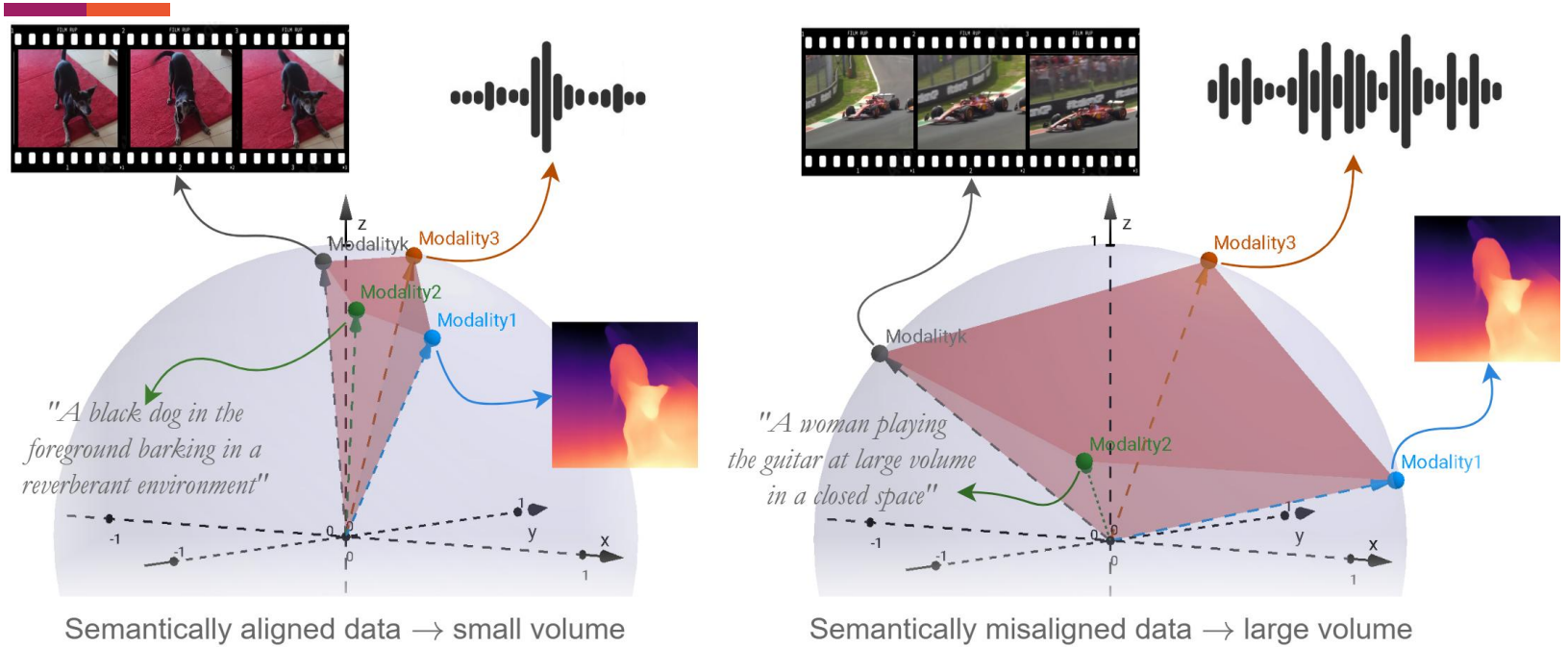


Figure 11. Visualization of the GRAM intuition: on the left, embedding vectors from semantically aligned multimodal data build a parallelotope with a small volume. On the right, where modalities are not aligned with each other, the formed parallelotope has a large volume [14].

[14] G. Cicchetti, E. Grassucci, L. Sigillo, and D. Comminiello, Gramian Multimodal Representation Learning and Alignment, *International Conference on Learning Representations* (ICLR), 2025.

Volume of the *k*-dimensional parallelotope

Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be vectors in $\mathbb{R}^k$, arranged as columns in a matrix $\mathbf{A} = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$. Then, the **Gram matrix** $\mathbf{G} = (\mathbf{v}_1, \ldots, \mathbf{v}_k) \in \mathbb{R}^{k \times \mathrm{k}}$ is defined as:

$$\mathbf{G}(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \mathbf{A}^\top \mathbf{A} = \begin{bmatrix} \langle \mathbf{v}_1, \mathbf{v}_1 \rangle & \langle \mathbf{v}_1, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_1, \mathbf{v}_k \rangle \\ \langle \mathbf{v}_2, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_2, \mathbf{v}_k \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{v}_k, \mathbf{v}_1 \rangle & \langle \mathbf{v}_k, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_k, \mathbf{v}_k \rangle \end{bmatrix}$$

The **determinant** of the **Gram Matrix** is directly linked to the volume of the *k*-dimensional parallelotope spanned by the *k* modality vectors as edges:

$$\mathrm{Vol}(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \sqrt{\det \mathbf{G}(\mathbf{v}_1, \ldots, \mathbf{v}_k)}$$

Beyond cosine similarity in contrastive learning

Exploiting the volume in a contrastive learning loss function.

- Cosine-based multimodal InfoNCE loss function:

$$\mathcal{L}_{(\mathbf{i}_i,\mathbf{t}_i)\sim(\mathcal{E}_\theta(I_i),\mathcal{E}_\phi(T_i))} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathbf{i}_i\cdot\mathbf{t}_i^\top)/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{i}_i\cdot\mathbf{t}_j^\top/\tau)}$$

Cosine-similarity

- Volume-based multimodal InfoNCE with Volume:

$$\mathcal{L}_{(\mathbf{m}_i^1,\cdots,\mathbf{m}_i^k)\sim(\mathcal{E}_\theta(M_i^1),\cdots,\mathcal{E}_\phi(M_i^k))} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(V(\mathbf{m}_i^1,\cdots,\mathbf{m}_i^k)/\tau)}{\sum_{j=1}^{B}\exp(V(\mathbf{m}_i^1,\cdots,\mathbf{m}_j^k)/\tau)}$$

Volume

The GRAM-based model architecture

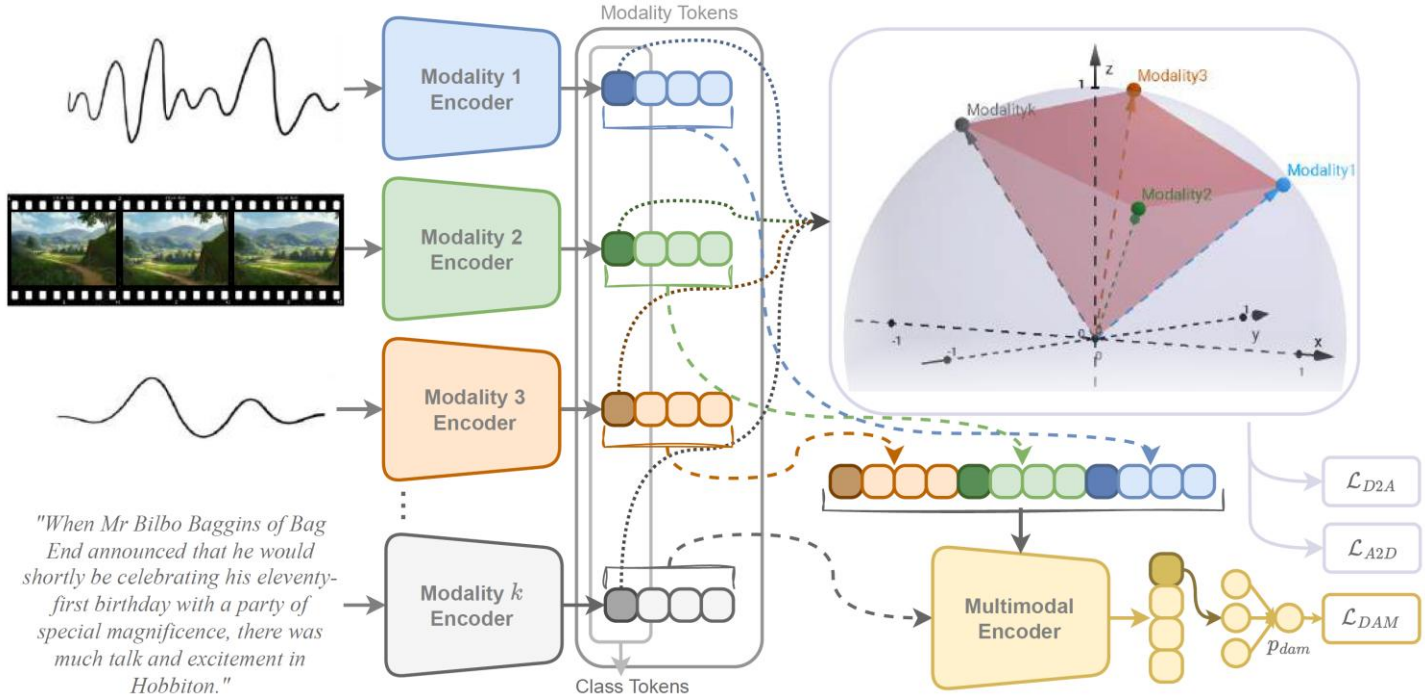
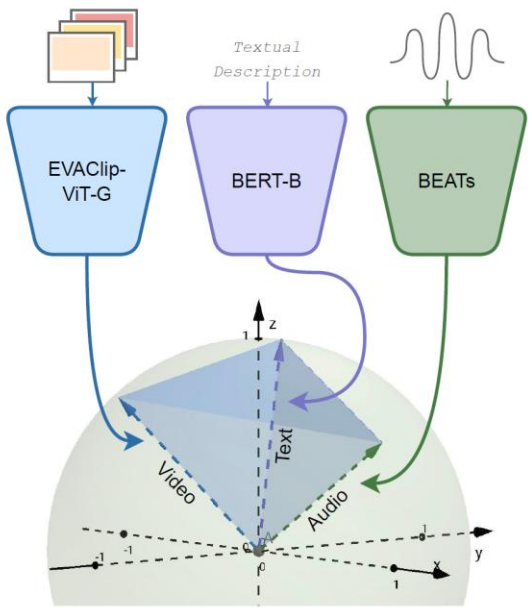


Figure 12. Class tokens from each modality are involved in shaping the k-dimensional parallelotope, whose volume indicates the semantic alignment of the modalities. All the tokens are then involved in the multimodal encoder to enhance the predictions [14].

Advantages of GRAM

- GRAM shows a novel approach to semantically aligning latent space of multiple modalities using a single loss function, going beyond classic pairwise cosine similarity approach.
- GRAM is able to scale up to N modalities, leveraging the true potential of multimodal AI.
- Geometrical theory background make the GRAM approach highly explainable.
- GRAM enables effective transfer learning showing state-of-the-art performance in a plethora of downstream tasks.

FoleyGRAM: GRAM-aligned encoders for Foley sounds



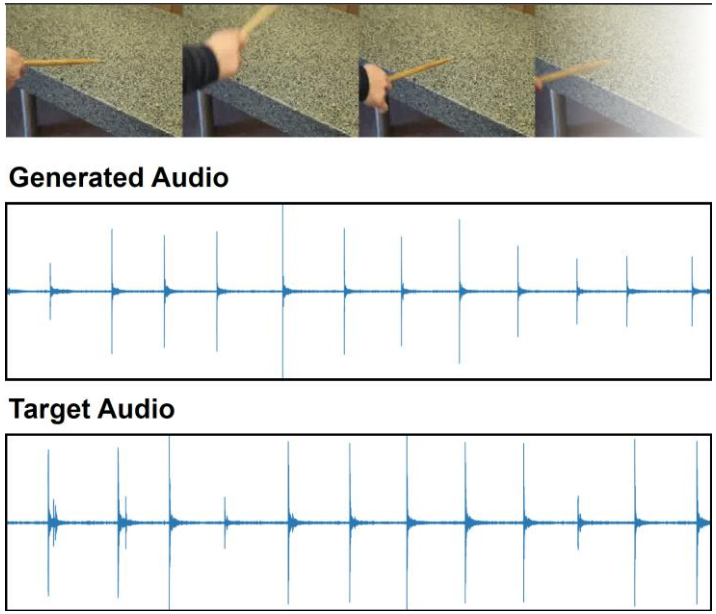


Figure 13. On the left, the GRAM framework applied to three encoders (EVAClip-ViT-G for video, BERT-B for text, and BEATs for audio), shaping the edges of the high-dimensional parallelotope to align modalities. On the right, ground truth and generated waveform [15].

[15] R. F. Gramaccioni, C. Marinoni, E. Grassucci, G. Cicchetti, A. Uncini, D. Comminiello, FoleyGRAM: Video-to-Audio Generation with GRAM-Aligned Multimodal Encoders, *International Joint Conference on Neural Networks* (IJCNN), Rome, Italy, July 2025.

FoleyGRAM: temporal and semantic alignment

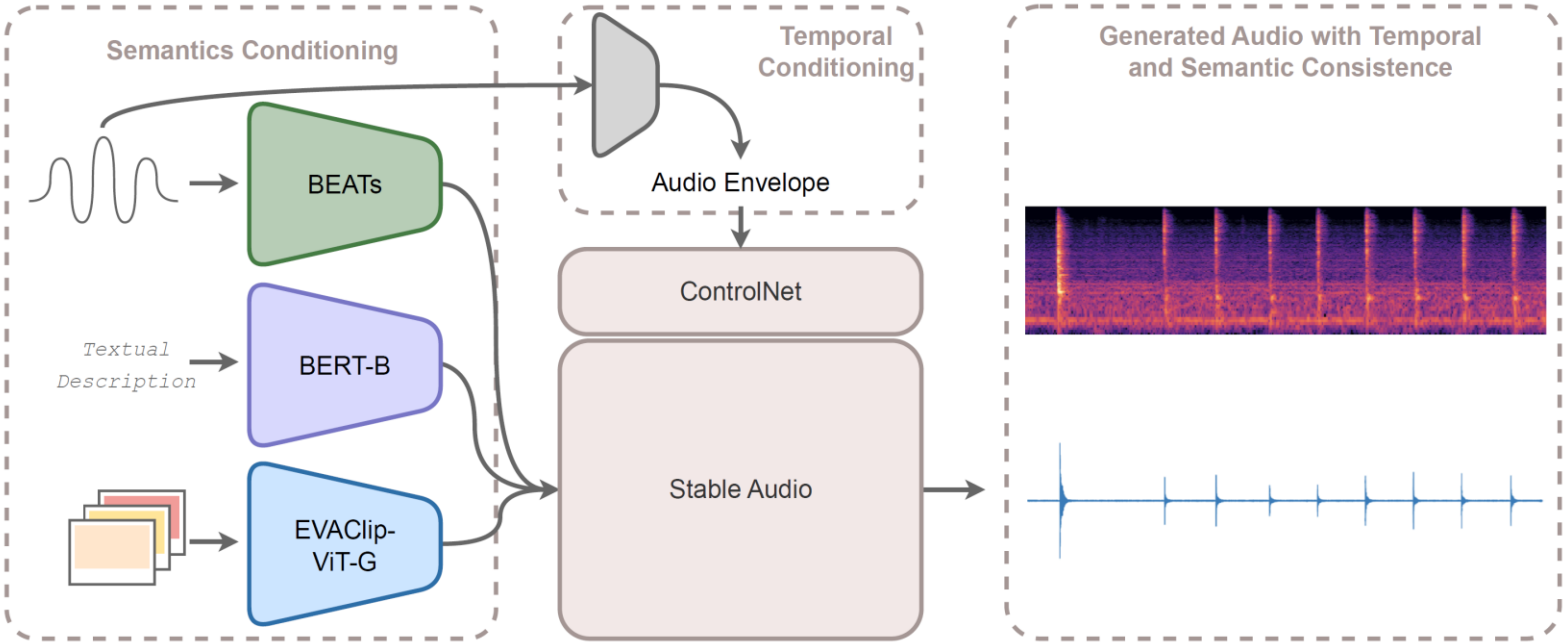


Figure 14. FoleyGRAM architecture: relevant semantic features are extracted from reference video, audio, and text through GRAM-aligned multimodal encoders. These features are used to condition an audio synthesis model that, together with the temporal information provided as an envelope signal used as input to a ControlNet, generates an audio that is temporally and semantically aligned with the reference video [15].

FoleyGRAM results: semantic fit without losing timing

Model	HRC	FAD-C ↓	FAD-LC ↓	CLAP ↑	FAVD ↓
SpecVQGAN [6]	NO	1001	0.710	0.142	6.513
Diff-Foley [1]	NO	654	0.469	0.373	4.618
CondFoleyGen [4]	NO	650	0.488	0.488	6.481
Video-Foley [7]	YES	435	0.167	0.678	2.207
SyncFusion [2]	YES	542	0.279	0.662	3.282
FoleyGRAM [15]	YES	**235**	**0.072**	**0.781**	**1.812**

Table 6. Results for **FoleyGRAM** and comparison with other SOTA models on Greatest Hits. HRC stands for Human Readable Control and refers to the use of time-varying interpretable signals that sound designers can use to control the generation process (i.e., envelope or onsets). FoleyGRAM provides the best results on all objective metrics [15].

Training-free multimodal diffusion guidance

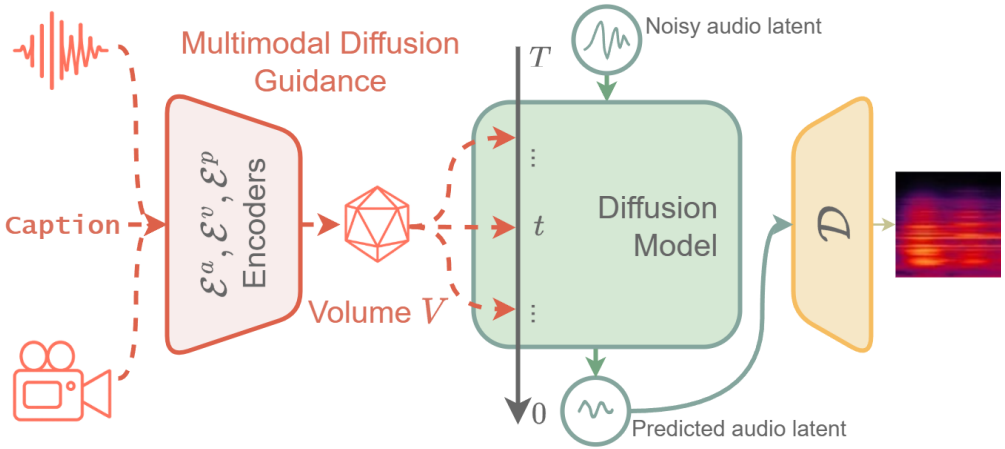


Figure 14. Overview of the multimodal diffusion guidance (MDG) for the video-to-audio generation process [16].

- Training-free guidance that enforces GRAM geometry during denoising.
- MDG computes the tri-modal volume (A/V/T) at each denoising step and nudges the latent to shrink it.
- It works on top of a pretrained AudioLDM: no retraining, no paired large datasets.

[16] E. Grassucci, G. Galadini, G. Cicchetti, A. Uncini, F. Antonacci, D. Comminiello, Training-Free Multimodal Guidance for Video to Audio Generation, *arXiv preprint arXiv:2509.24550*, Sept. 2025.

MDG metrics and results

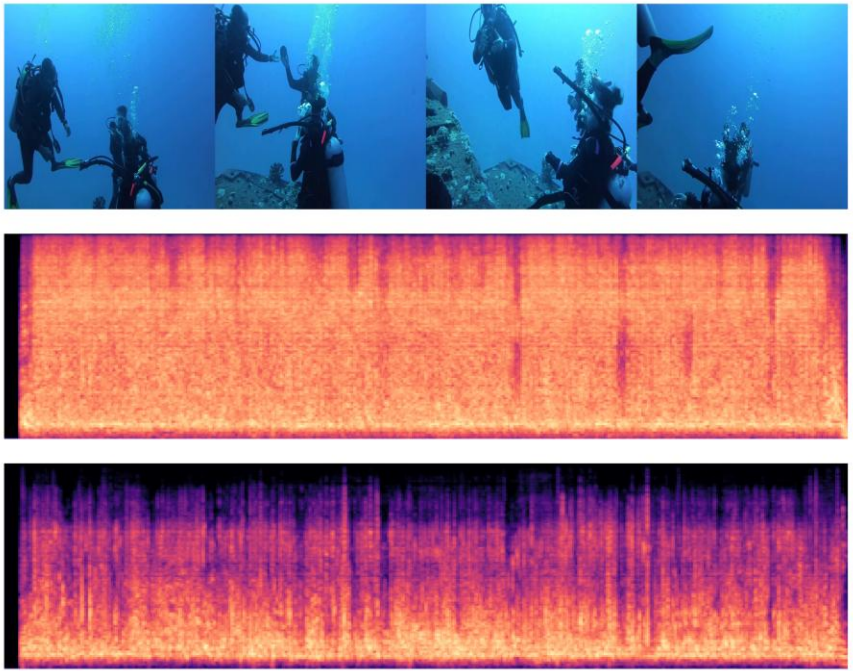


Figure 16. Generated samples from VGGSound by Seeing&Hearing [17] and by **MDG**, which better guides the generation towards a semantically meaningful synthesized audio.

Dataset	Model	FAD ↓	FAVD ↓	KL ↓	ISc ↑
VGGSound	SpecVQGAN [6]	7.74	-	3.29	5.11
	Diff-Foley [1]	8.91	3.57	3.31	4.28
	Seeing&H. [17]	7.80	3.44	3.35	4.88
	MDG [16]	**6.04**	**2.60**	**2.78**	**5.88**
AudioCaps	Seeing&H. [17]	11.04	4.44	3.43	4.58
	MDG [16]	**10.77**	**4.31**	**3.40**	**4.68**

Table 7. Quantitave results of generated audio samples [16].

Model	$V \downarrow$	$\delta_{(\cos)} \downarrow$	$\delta^{t,v}_{(\cos)} \downarrow$	$\delta^{t,a}_{(\cos)} \downarrow$	$\delta^{v,a}_{(\cos)} \downarrow$
Seeing&H. [17]	0.937	2.488	0.703	0.981	0.893
MDG [16]	**0.819**	**2.068**	**0.517**	**0.713**	**0.838**

Table 8. Semantic consistency results of generated audio with caption and video from the VGGSound test set. Comparison between the pairwise-guided Seeing&Hearing and the proposed MDG [16].

4 | Where: Spatial Alignment

Why “space” matters

- Most V2A systems remain monophonic, while focusing on semantic alignment and temporal synchronization.

- **Spatial coherence** remains largely unexplored. However, lack of spatial cues breaks immersion.

- Treating spatialization as native conditioning would complete multimodal audio-visual generation models.

StereoSync: advancing spatial awareness

StereoSync introduces spatially-aware stereo audio generation from video, enhancing immersion by accurately placing sound sources in the visual scene.

StereoSync leverages RollingDepth for depth estimation and MASA for object tracking without requiring extensive training.

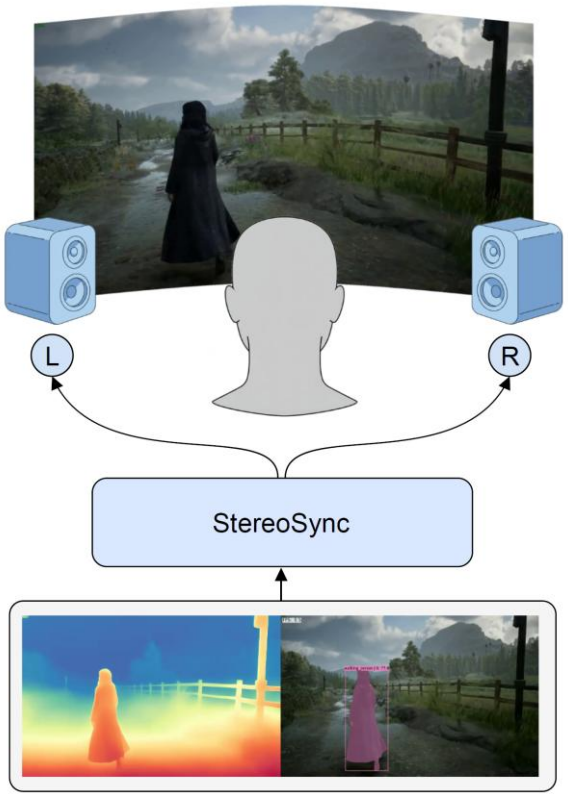


Figure 17. StereoSync generates a stereo audio that resembles the spatial context of an input video [18].

[18] C. Marinoni, R. F. Gramaccioni, K. Shimada, T. Shibuya, Y. Mitsufuji, D. Comminiello, StereoSync: Spatially-Aware Stereo Audio Generation from Video, *International Joint Conference on Neural Networks* (IJCNN), Rome, Italy, July 2025.

StereoSync: adding spatial controls to V2A gen models

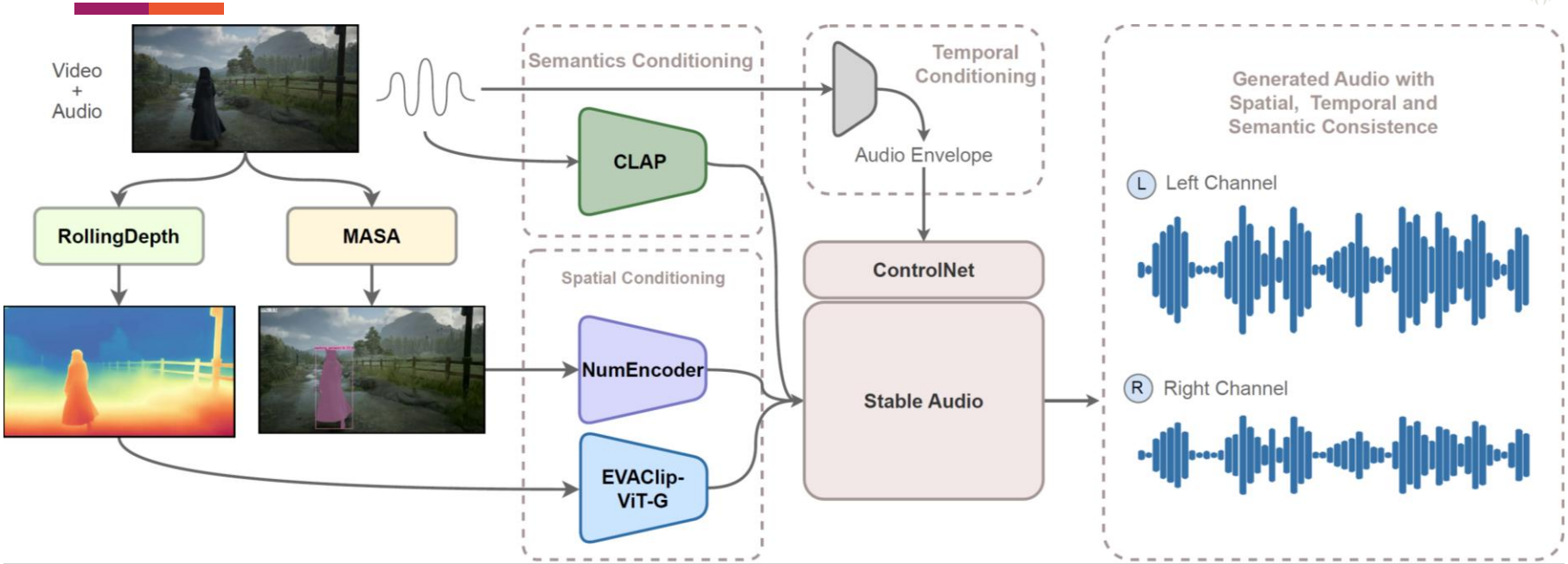


Figure 18. The StereoSync framework [18].

Depth maps (scene geometry) and **bounding boxes (object tracking)** are extracted from videos and integrated into a diffusion-based audio synthesis model using cross-attention conditioning to maintain spatial coherence.

How StereoSync encodes “where”

The goal is to make the spatial inputs concrete and editable.

Global cues: depth maps are fed into an EVA-CLIP encoder creating a temporal stack of embeddings.

Local cues: bounding box sequences are fed into a number conditioning encoder (four float streams: x_1, y_1, x_2, y_2).

Figure 18. Examples from the Walking The Maps showing the bounding box of the subjects and their corresponding movement trajectories [18].

Measuring spatial alignment

- **Spatial AV-Align** is a metric designed to evaluate the spatial synchronization between sound events in generated audio and corresponding objects in a video.

- The metric involves an object detector (YOLOX) to identify candidate sound source positions and a stereo SELD model detecting sound events. Then it aligns predicted sound direction with bbox position over time.

- Other metrics: FAD/FAVD (quality/AV fit) and E-L1 (timing).

StereoSync results for V2A gen on Walking the Maps

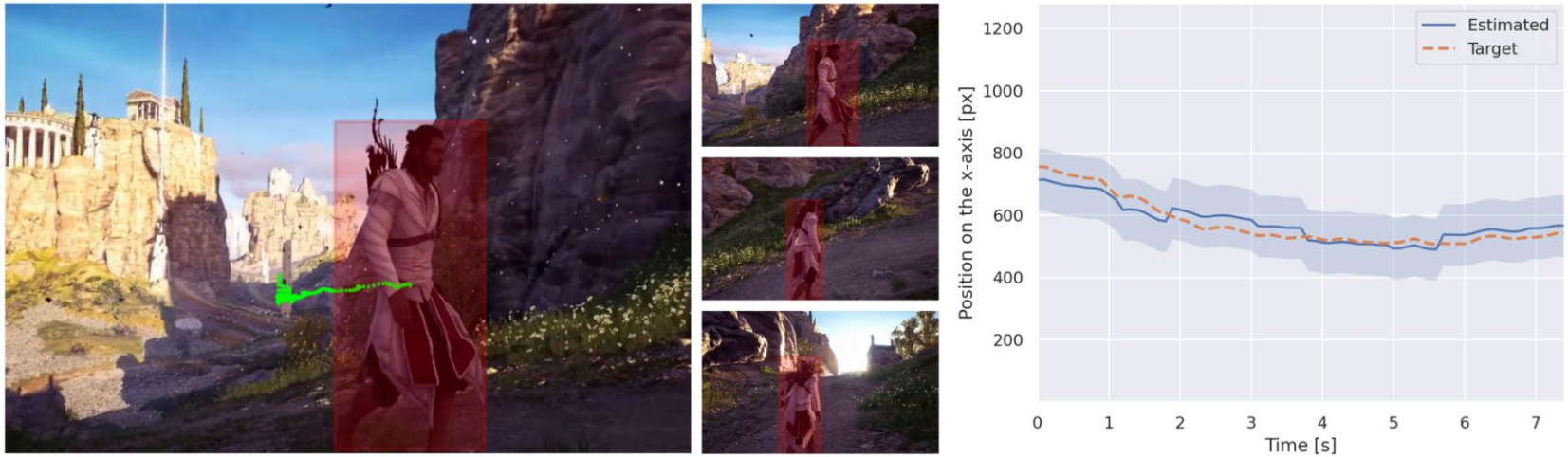


Figure 18. Visual example illustrating the behavior of the Spatial AV-Align metric [18].

Model	Spatial Alignment	Semantic Alignment		Temporal Alignment
	AV-Align ↑ (GT=0.89)	FAD-LC ↓	FAVD ↓	E-L1 ↓
StereoSync (w/out spatial cond.)	0.61	0.256	3.068	0.062
StereoSync	0.78	0.230	3.830	0.047

Table 9. Objective metric evaluation. Spatial-AV-Align indicates the spatial alignment between video and audio [18].

5 | Toward Joint AV Generation

From controls to co-generation

- So far we have seen how to control time, space and semantics knobs to generate consistent audio content from silent video.

- What if we generate audio-visual content jointly?

- Significant challenges need to be addressed, including:
  - High-dimensional AV data streams
  - More difficult synchronization
  - Different nature and structure of AV signals (frames vs waveforms)

360°-conditioned joint AV generation

Can we generate a video taking into account what is happening in the part of the world not shown in the video?

We are interested in generating viewpoints of a certain world, given

- saliency maps (salient points in the video);
- BASD maps (information on the direction of the viewpoint);
- a caption describing action taking place in the whole 360° space.

[19] C. Marinoni, R. F. Gramaccioni, E. Grassucci, D. Comminiello, Controllable Audio-Visual Viewpoint Generation from 360° Spatial Information, *arXiv preprint arXiv:2510.06060,* Sept. 2025.

Viewpoint generation

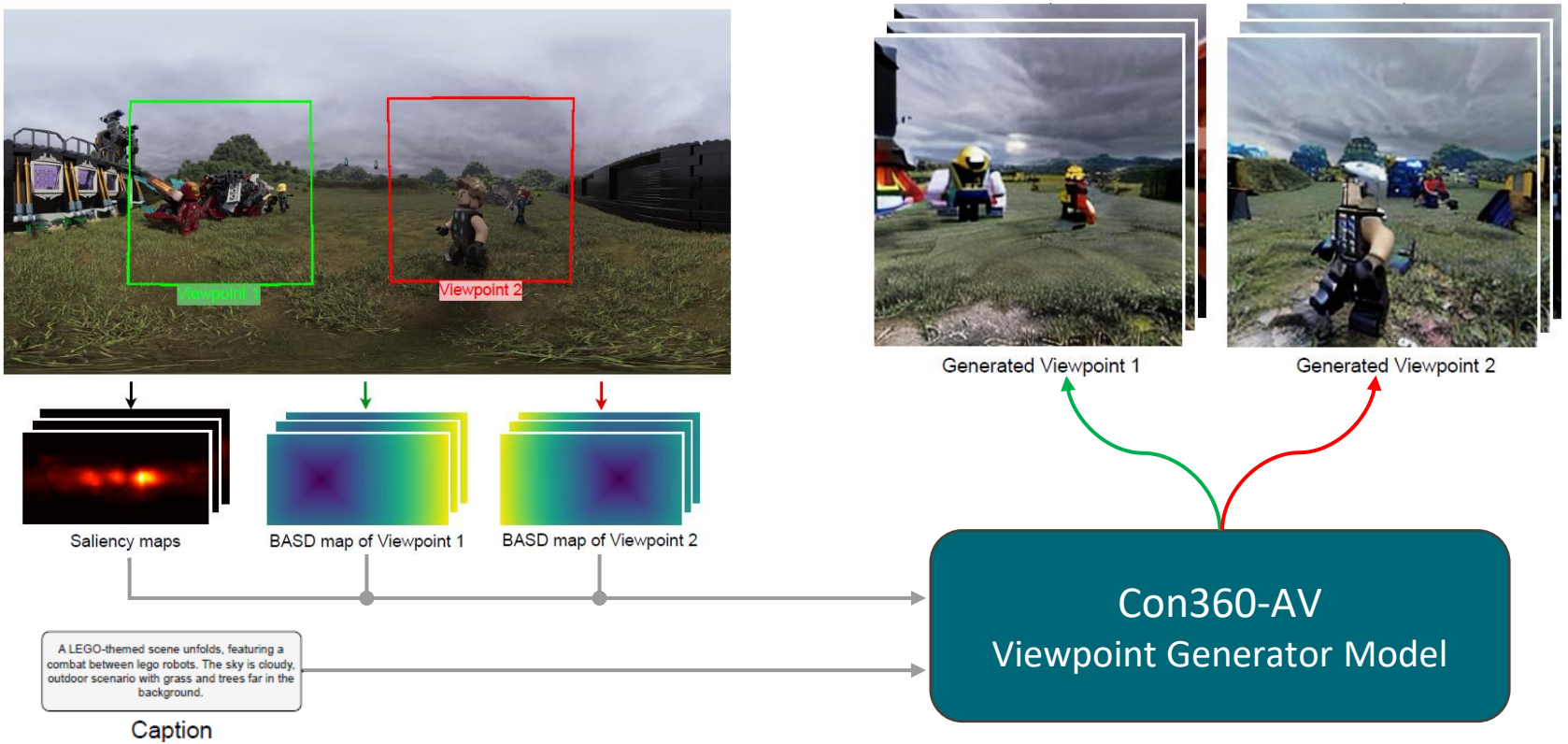


Figure 19. The Con360-AV model generates a viewpoint conditioned on global 360° saliency maps, a text caption, and BASD maps [19].

The Con360-AV feature extraction pipeline

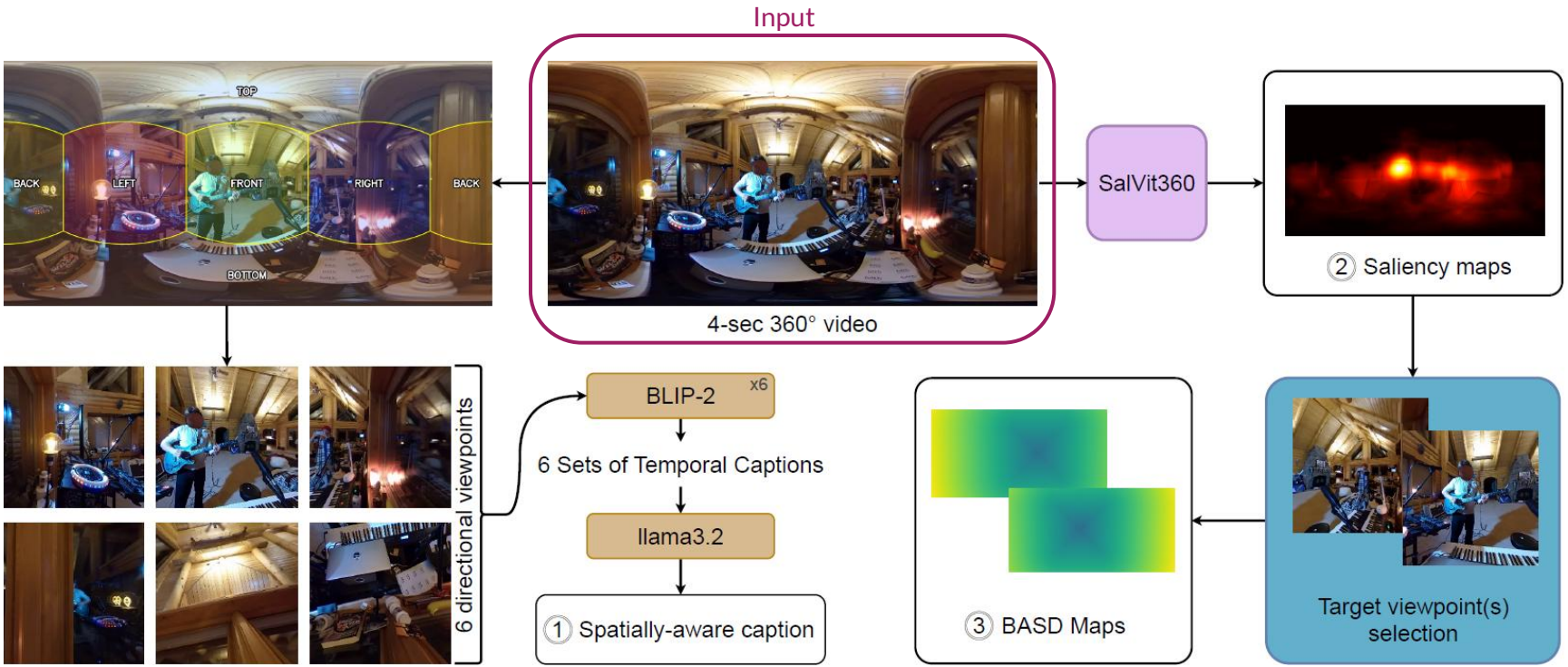
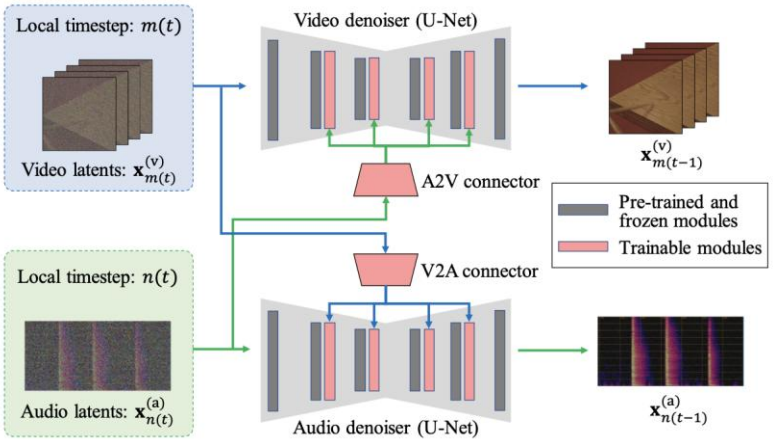


Figure 20. Extraction of the three contextual conditionings from a 360° video: the pipeline generates a textual prompt and two visual prompts [19].

Con360-AV: prepare the conditional information

Inspired by [20], we used two pretrained diffusion models, AudioLDM and AnimateDiff, and we adapted the Cross-Modal Conditioning as Positional Encoding (CMC-PE) strategy to take into account spatial cues.



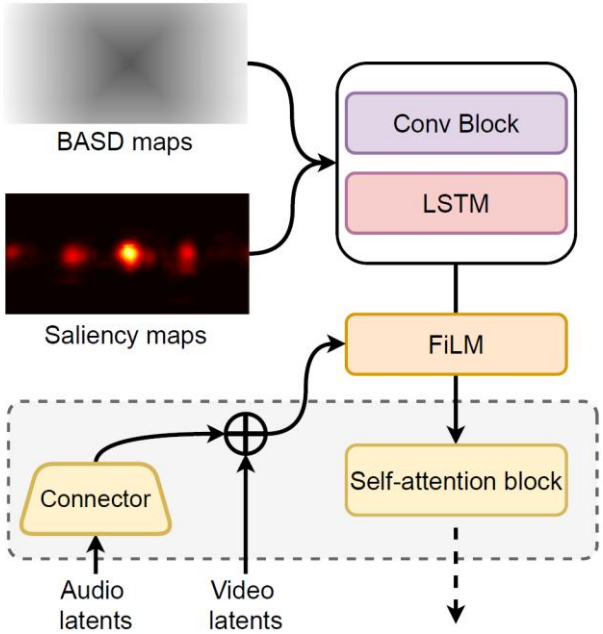


Figure 21. On the right, the join generation baseline [20], and, above, conditioning mechanism of Con360-AV [19].

[20] M. Ishii, A. Hayakawa, T. Shibuya, Y. Mitsufuji, A Simple but Strong Baseline for Sounding Video Generation: Effective Adaptation of Audio and Video Diffusion Models for Joint Generation. *International Joint Conference on Neural Networks* (IJCNN), Rome, Italy, July 2025.

Con360-AV results

The Con360-AV model is able to produce credible viewpoint videos for the same scene, showing excellent control over camera directionality and showcasing audio that takes off-screen events into account.

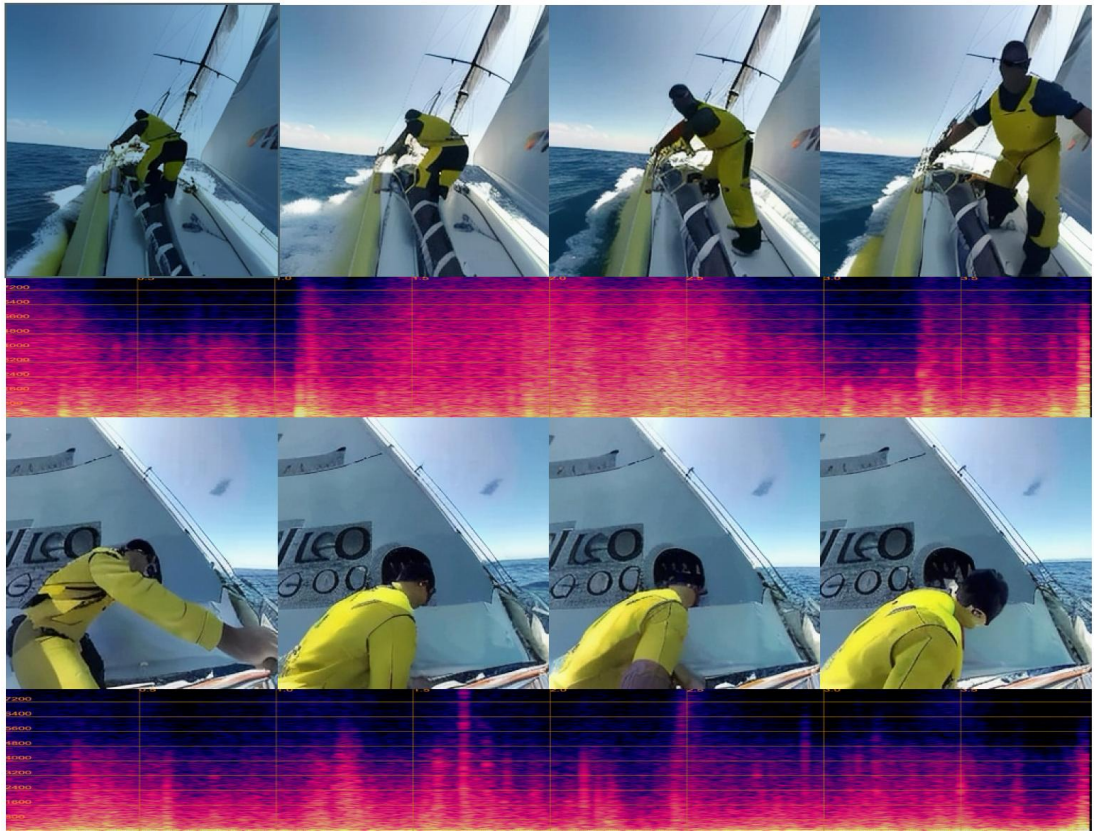


Figure 22. Example of 2 generated viewpoints of the same scene [19].

Con360-AV objective metrics

To evaluate the spatial alignment we adopt the S_{KL} measure that averages the frame-wise KL divergence between the saliency maps of the generated video and the target viewpoint:

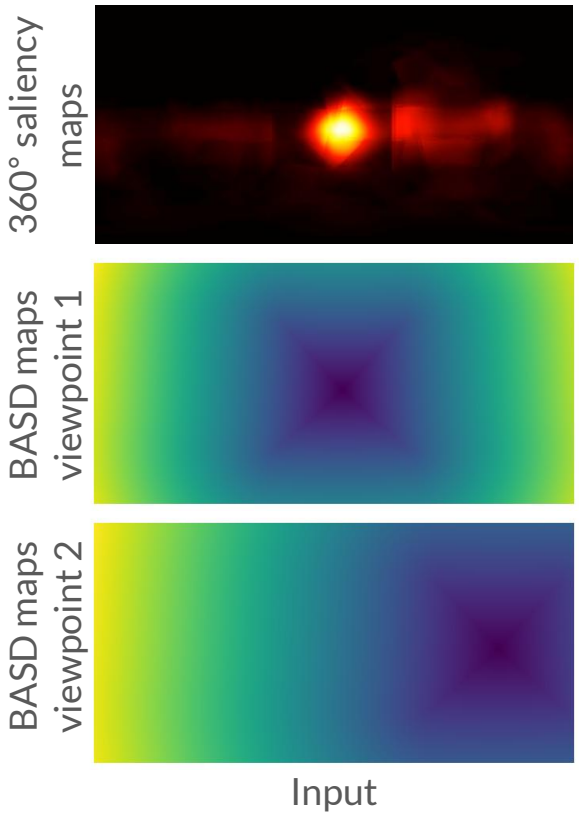
$$S_{KL} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}(\mathrm{Sal}(\hat{v}_t)\,||\mathrm{Sal}(v_t))$$

where T is the number of frames, Sal(·) the saliency map, $\hat{v}_t$ the generated frame, and v_t the target frame.

Model	$S_{\mathrm{KL}}\downarrow$	FAD ↓	FAVD ↓
Baseline [20]	1.953	8.23	1959
Con360-AV [19]	**0.645**	**3.21**	**1327**

Table 10. Objective metric evaluation [19].

Con360-AV viewpoint generation results



Generated viewpoint 1

Generated viewpoint 2

Video 4. Example of 2 viewpoints generation: play viewpoint1, play viewpoint2 [19].

6 | Conclusion

Conclusion: design patterns that generalize

- Generative AV becomes controllable when we expose three dials: **When** (temporal), **Where** (spatial), **What** (semantic).
- This results in a predictable, editable, high-fidelity AV generation.
- Decouple then compose: learn separate controls then recombine in the generator.
- Human readable controls help the content creator to have a complete control over the generated content.

Conclusion: open problems and future research

- Spatial scale-up: from stereo to binaural/FOA with scene geometry, including better perceptual spatial metrics.
- Long-form and edits: stable scene memory, local re-synthesis without global drift; streaming/low-latency inference.
- Datasets: high-SNR AV pairs, richer spatial annotations, synthetic-to-real adaptation, 360° AV, different downstream tasks.
- Joint training: tighter AV co-generation with controllable dials natively trained, not only guided.

ISPAMM Lab

CHRISTIAN MARINONI
PhD Student

RICCARDO GRAMACCIONI
PhD Student

ELEONORA GRASSUCCI
Assistant Professor

GIORDANO CICCHETTI
PhD Student

This talk would not have been possible without the hard and invaluable work of Christian, Riccardo, Eleonora and Giordano, members of the Intelligent Signal Processing And MultiMedia (ISPAMM) Lab of Sapienza University of Rome, Italy.

References

References

[1] S. Luo, C. Yan, C. Hu, H. Zhao, Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models, *Advances in Neural Information Processing Systems* (NeurIPS), vol. 36, pp. 48855–48876, 2023.

[2] M. Comunità, R. F. Gramaccioni, E. Postolache, E. Rodolà, D. Comminiello, J. D. Reiss, SyncFusion: Multimodal Onset-Synchronized Video-to-Audio Foley Synthesis, *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 936–940, 2024.

[3] R. F. Gramaccioni, C. Marinoni, E. Postolache, M. Comunità, L. Cosmo, J. D. Reiss, D. Comminiello, FOL·AI: Synchronized Foley Sound Generation with Semantic and Temporal Alignment, *arXiv preprint arXiv:2412.15023v3*, May 2025.

[4] Y. Du, Z. Chen, J. Salamon, B. Russell, A. Owens, Conditional Generation of Audio from Video via Foley Analogies, *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 2426–2436, 2023.

[5] Y. Chung, J. Lee, J. Nam, T-Foley: A Controllable Waveform-Domain Diffusion Model for Temporal-Event-Guided Foley Sound Synthesis,*IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 6820–6824 ,2024.

[6] V. Iashin, E. Rahtu, Taming Visually Guided Sound Generation, *British Machine Vision Conference* (BMVC), 2021.

[7] J. Lee, J.-Y. Im, D. Kim, J. Nam, Video-Foley: Two-Stage Video-to-Sound Generation via Temporal Event Condition for Foley Sound, *arXiv preprint arXiv:2408.11915,* 2024.

[8] S. Ghose, J. J. Prevost, FoleyGAN: Visually Guided Generative Adversarial Network-Based Synchronous Sound Generation in Silent Videos, *IEEE Transactions on Multimedia,* 25, 4508–4519, 2023.

[9] Y. Zhou, Z. Wang, C. Fang, T. Bui, T. L. Berg, Visual to Sound: Generating Natural Sound for Videos in the Wild, *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 3550–3558, 2018.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, Learning Transferable Visual Models from Natural Language Supervision, *International Conference on Machine Learning* (ICML), 2021.

References

[11] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, W. Zhang, Z. Li, W. Liu, and L. Yuan. LanguageBind: Extending Video-Language Pretraining to *n*-Modality by Language-Based Semantic Alignment, in *International Conference on Learning Representations* (ICLR), 2024.

[12] S. Chen, H. Li, Q. Wang, Z. Zhao, M.-T. Sun, X. Zhu, J. Liu, VAST: A Vision-Audio-Subtitle-Text Omni-Modality foundation Model and Dataset, *Advances in Neural Information Processing Systems* (NeurIPS), 2023.

[13] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. Vasudev Alwala, Armand Joulin, and Ishan Misra, ImageBind: One Embedding Space to Bind Them All, *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 15180–15190, 2023.

[14] G. Cicchetti, E. Grassucci, L. Sigillo, and D. Comminiello, Gramian Multimodal Representation Learning and Alignment, *International Conference on Learning Representations* (ICLR), 2025.

[15] R. F. Gramaccioni, C. Marinoni, E. Grassucci, G. Cicchetti, A. Uncini, D. Comminiello, FoleyGRAM: Video-to-Audio Generation with GRAM-Aligned Multimodal Encoders, *International Joint Conference on Neural Networks* (IJCNN), Rome, Italy, July 2025.

[16] E. Grassucci, G. Galadini, G. Cicchetti, A. Uncini, F. Antonacci, D. Comminiello, Training-Free Multimodal Guidance for Video to Audio Generation, *arXiv preprint arXiv:2509.24550,* Sept. 2025.

[17] Y. Xing, Y.-Y. He, Z-Tian, X.Wang, and Q. Chen, Seeing and Hearing: Open-Domain Visual-Audio Generation with Diffusion Latent Aligners, *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 7151–7161, 2024.

[18] C. Marinoni, R. F. Gramaccioni, K. Shimada, T. Shibuya, Y. Mitsufuji, D. Comminiello, StereoSync: Spatially-Aware Stereo Audio Generation from Video, *International Joint Conference on Neural Networks* (IJCNN), Rome, Italy, July 2025.

[19] C. Marinoni, R. F. Gramaccioni, E. Grassucci, D. Comminiello, Controllable Audio-Visual Viewpoint Generation from 360° Spatial Information, *arXiv preprint arXiv:2510.06060,* Sept. 2025.

References

[20] M. Ishii, A. Hayakawa, T. Shibuya, Y. Mitsufuji, A Simple but Strong Baseline for Sounding Video Generation: Effective Adaptation of Audio and Video Diffusion Models for Joint Generation. *International Joint Conference on Neural Networks* (IJCNN), Rome, Italy, July 2025.

[21] G. Chen, G. Wang, X. Huang, J. Sang, Semantically Consistent Video-to-Audio Generation Using Multimodal Language Large Model, *arXiv preprint arXiv:2404.16305*, Apr. 2024.

[22] M. Yi, M. Li, Efficient Video to Audio Mapper with Visual Scene Detection, Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Singapore, Oct. 2025.

[23] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, M. D. Plumbley, AudioLDM: Text-to-Audio Generation with Latent Diffusion Models, *International Conference on Machine Learning* (ICML), vol. 202, pp. 21450–21474, 2023.

[24] H. Liu, Q. Tian, Y. Yuan, X. Liu, X. Mei, Q. Kong, Y. Wang, W. Wang, M. D. Plumbley, AudioLDM 2: Learning Holistic Audio Generation with Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 2871–2883, 2023.

[25] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, J. Pons, Fast Timing-Conditioned Latent Audio Diffusion. *International Conference on Machine Learning* (ICML), 2024.

[26] S.-L. Wu, C. Donahue, S. Watanabe, N. J. Bryan, Music ControlNet: Multiple Time-Varying Controls for Music Generation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 32, 2692–2703, 2023.

[27] Z. Evans, J. Parker, C. Carr, Z. Zukowski, J. Taylor, J. Pons, Stable Audio Open, *arXiv preprint arXiv:2407.14358,* 2024.

[28] B. Elizalde, S. Deshmukh, M. A. Ismail, H. Wang, CLAP Learning Audio Concepts from Natural Language Supervision, *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 1–5, 2023.

[29] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo, MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 10219–10228, 2023.

THANK YOU FOR YOUR ATTENTION

danilo.comminiello@uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA