FATE: FRAME-LEVEL AUDIO-VISUAL TEMPORAL EMBEDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

When a dog opens its mouth and barks, humans naturally recognize what the sound is and when it occurs. Building audio-visual models with this same ability requires representations that capture both semantic and temporal alignment. Current approaches fall short on one side or the other: embedding models capture semantics but lose temporal information; synchronization models capture temporal offsets but lack semantic understanding. To bridge this gap, we propose **FATE**, **F**rame-level **A**udio-visual **T**emporal **E**mbedding. Unlike prior embedding models that pool each modality into a single embedding and discard temporal information, FATE retains frame-level sequences, aligns them on the physical timeline, and computes similarity over strictly aligned frame pairs. Unlike synchronization models that output only an offset prediction, FATE encodes synchronization in a reusable embedding space, trained with a joint objective combining cross-video semantic and within-video temporal contrastive learning to capture both *what* sounds and *when* it occurs. Across three tasks, FATE surpasses the strongest baseline on temporal and semantic retrieval by a large margin, matches fully supervised methods on event localization in a zero-shot setting, and achieves the best correlation with human judgments as a generation evaluation metric. The source code can be found at `https://anonymous.4open.science/r/FATE-ICLR`.

1 INTRODUCTION

When a dog opens its mouth to bark, the barking sound occurs at precisely the same moment. Any semantic contradiction or temporal misalignment immediately feels unnatural to human observers. This tight coupling between *what* and *when* is the foundation of human audio-visual perception.

Similarly, models designed to emulate human perception, like multimodal understanding or generation, must also capture both semantic and temporal information. Existing approaches, however, typically emphasize one dimension at the expense of the other. Audio-visual embedding models (Radford et al., 2021; Wu et al., 2023; Assran et al., 2025; Tuncay et al., 2025; Guzhov et al., 2021; Girdhar et al., 2023; Araujo et al., 2025; Vyas et al., 2025) train dual encoders with contrastive objectives to project audio and video into a shared embedding space. Although effective at semantic matching, they often discard fine-grained temporal information: as shown in Fig. 1 (bottom), such models score an unsynchronized segment even higher than the synchronized one. Conversely, synchronization models (Chen et al., 2021; Iashin et al., 2024) capture temporal offsets, but their task-specific objectives yield weak semantic representations: in Fig. 1 (top), the synchronization model fails to distinguish the barking dog from unrelated violin. Nor do recent Omni LLMs (qwen team, 2025; GeminiTeam, 2023) close this gap: their low sampling rates (e.g., 2 fps for Qwen3-Omni) preclude fine-grained synchronization, and they produce no reusable embeddings for downstream tasks(Fig. 1, bottom). Consequently, a unified embedding representation that jointly captures *what* sounds and *when* it occurs remains lacking.

In this paper, we propose **FATE**, **F**rame-level **A**udio-visual **T**emporal **E**mbedding. The key idea is simple: instead of compressing each modality into a single global vector, FATE preserves frame-level features and aligns the two modalities on the physical time axis, so that cross-modal similarity naturally reflects both semantic and temporal alignment. We train FATE with a joint objective com-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

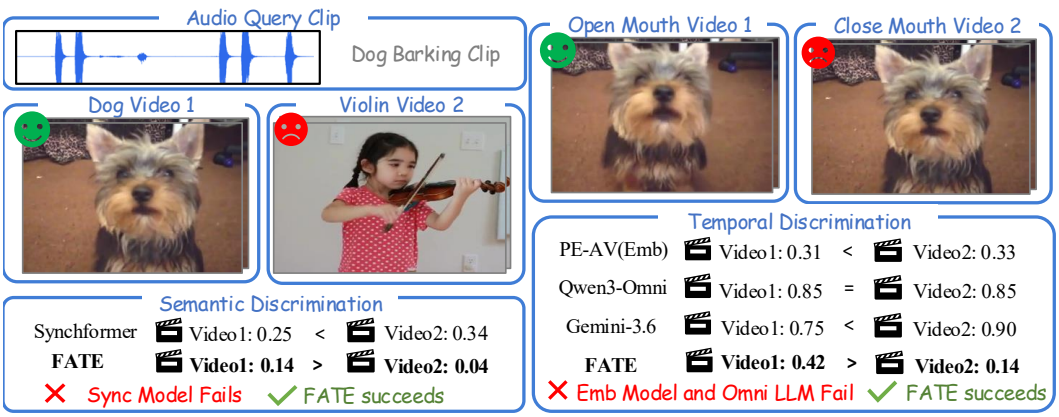

Figure 1: Semantic (what) vs. temporal (when) discrimination for a dog-barking audio query. Top: the synchronization model scores the unrelated violin video higher than the barking dog. Bottom: the embedding model (PE-AV) and Omni LLMs score the unsynchronized segment (mouth closed) higher than or equal to the synchronized one, whereas FATE succeeds in both dimensions. Embedding and synchronization scores are normalized to $[0, 1]$ (higher is better; Synchformer's offset output is inverted); Omni LLMs rate each clip independently via prompting (details in Appendix A.5).

bining semantic contrastive learning across videos and temporal soft-contrastive learning within each video, encouraging the embedding space to capture both *what* sounds and *when* they occur.

We evaluate FATE on three downstream tasks. For temporal cross-modal retrieval, FATE outperforms the strongest synchronization baseline by over 13 points in R@3, while simultaneously achieving over 20 points higher semantic retrieval accuracy. For audio-visual event localization, FATE in a zero-shot setting, surpasses supervised methods with 48.3% average accuracy. As an automatic metric for audio-visual generation, FATE agrees most closely with human synchronization judgments among all compared metrics. In summary, our main contributions are as follows:

(1) We introduce **FATE**, a reusable frame-level audio-visual embedding model. It preserves temporal feature sequences, aligns audio and video in physical time, and scores time-aligned frame pairs, representing synchronization without task-specific prediction heads.

(2) We train FATE with cross-video semantic contrastive learning and within-video temporal soft-contrastive learning. Controlled ablations show that the two objectives are complementary and that offset-aware soft targets improve temporal discrimination.

(3) We demonstrate FATE's transfer across temporal cross-modal retrieval, zero-shot event localization, and synchronization evaluation of generated videos. Without task-specific fine-tuning, FATE leads the retrieval baselines, matches supervised localization methods, and achieves the strongest agreement with human judgments among the compared metrics.

2 Related Work

2.1 Audio-Visual Representation Learning

Multi-modal representation learning has advanced significantly, driven by the success of contrastive learning (Oord et al., 2018). CLIP (Radford et al., 2021) pioneers this direction by aligning image and text modalities. This paradigm naturally extends to the audio-visual domain: AudioCLIP (Guzhov et al., 2021), ImageBind (Girdhar et al., 2023) LanguageBind (Zhu et al., 2023) and PE-AV (Vyas et al., 2025) scale this idea further to audio, image, text and video. However, all of these approaches capture *what* co-occurs across modalities but not *when*. Recent omni-modal LLMs (qwen team, 2025; GeminiTeam, 2023) also process audio and video jointly, but their sparse visual sampling (e.g., 2 fps in Qwen3-Omni) is far coarser than the granularity that synchronization requires, and they yield no reusable embeddings for downstream task; Fig. 1 illustrates this failure. This gap motivates our frame-level design.

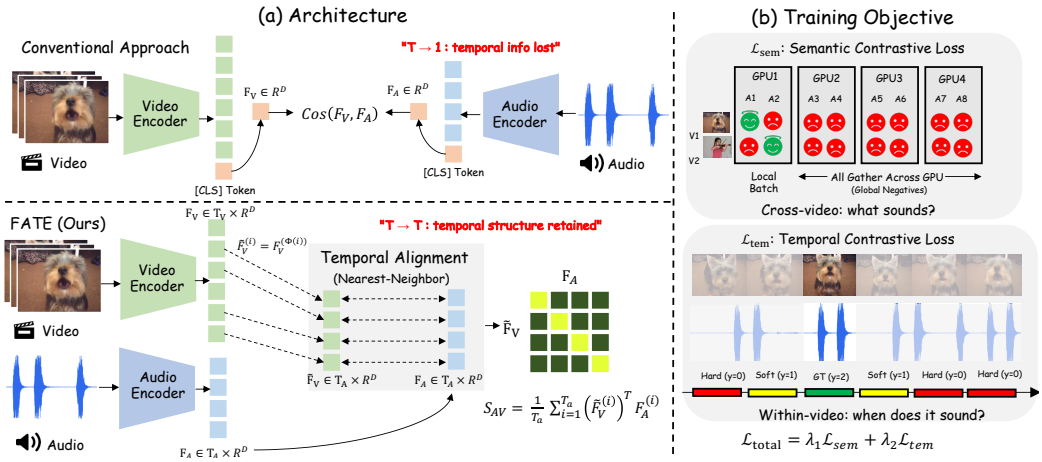

Figure 2: **Overview of FATE.** (a) **Architecture.** Conventional models pool encoder outputs into a single vector per modality ($T \to 1$), discarding temporal structure. FATE retains frame-level tokens ($T \to T$), aligns audio and video by nearest-neighbor resampling, and averages similarities over time-aligned pairs. (b) **Training objective.** The semantic loss $\mathcal{L}_{\text{sem}}$ separates videos using cross-GPU negatives, whereas the temporal loss $\mathcal{L}_{\text{temp}}$ distinguishes offsets within each video using offset-dependent soft targets (Eq. 6).

2.2 Audio-Visual Synchronization Models

Audio-visual synchronization aims to determine whether audio and video clips are temporally coherent. SyncNet (Chung & Zisserman, 2017) pioneered this direction with two-tower CNNs for lip-speech synchronization, and AVST (Chen et al., 2021) generalized it to open-domain videos. SparseSelector (Iashin et al., 2022) and Synchformer (Iashin et al., 2024) further moved from binary detection to quantitative offset estimation. On another front, CAVP (Luo et al., 2023) grounds synchronization in temporal contrastive learning, while PEAVS (Goncalves et al., 2024) directly regresses human opinion scores. These models output a single prediction (an offset or a score) but do not yield reusable representations, making them difficult to transfer to other tasks. In contrast, FATE learns a general-purpose embedding space where synchronization is reflected in frame-level embeddings, making it applicable across diverse downstream scenarios.

2.3 Evaluating Audio-Visual Generative Models

Audio-visual generative models (Wang et al., 2025b; Guan et al., 2025b; Liu et al., 2025; Low et al., 2025; HaCohen et al., 2026; Seedance et al., 2025; Wang et al., 2024; Cheng et al., 2025; Wang et al., 2025a; Cheng et al., 2026) are advancing rapidly, yet reliable evaluation of synchronization remains limited (Huang et al., 2024; Zheng et al., 2025; Huang et al., 2025; Guan et al., 2025a). Existing metrics fall into three paradigms. Rule-based metrics such as AV-Align (Yariv et al., 2023) match optical flow against audio onsets, but these low-level cues break down on complex generative artifacts. Offset predictors such as DeSync (Iashin et al., 2024) regress a global temporal shift, yet synchronization errors in generated videos are rarely uniform offsets. Embedding-based scores such as CAVP (Luo et al., 2023) measure similarity in the embedding space, but pool away the temporal dimension that synchronization depends on. FATE addresses all three by scoring synchronization over frame-level embeddings.

3 Method

3.1 Problem Formulation

We aim to learn audio-visual representations that jointly encode semantic and temporal alignment. Let $V \in \mathbb{R}^{T \times H \times W \times 3}$ denote a video clip of T frames, and let $A \in \mathbb{R}^{L \times C}$ denote its audio representation with L temporal frames and C channels. A dual-encoder architecture maps the two modalities

into a shared embedding space:

$$F_V = E_v(V), \quad F_A = E_a(A), \tag{1}$$

and the cross-modal similarity is computed as $S_{AV} = \mathrm{Sim}(F_A, F_V)$. We require S_{AV} to satisfy two properties: (i) *semantic discrimination*: $S_{AV}(V, A) > S_{AV}(V, A')$ for audio A' drawn from a semantically different clip; and (ii) *temporal discrimination*: $S_{AV}(V, A) > S_{AV}(V, A_{\Delta t})$ for any temporally shifted version $A_{\Delta t}$ of the same audio ($\Delta t \neq 0$). Both properties are required symmetrically in the video direction, i.e., $S_{AV}(V, A) > S_{AV}(V', A)$ and $S_{AV}(V, A) > S_{AV}(V_{\Delta t}, A)$. Conventional embedding models satisfy (i) but not (ii). Synchronization models, which train without cross-video contrast, satisfy (ii) but fail to separate semantically different clips, violating (i). FATE satisfies both: by retaining frame-level embedding and defining Sim over temporally aligned frame pairs, semantic and temporal discrimination are encoded in one single similarity.

3.2 Architecture of FATE

Fig. 2 illustrates the architecture of FATE. We build on the Perception Encoder Audio-Visual (PE-AV) (Vyas et al., 2025), retaining its dual-encoder backbone but discarding the global pooling head. The resulting pipeline operates in three stages: frame-level feature extraction, temporal alignment, and frame-level similarity computation.

Frame-level feature extraction. FATE bypasses the pooling step and retains full tokens. Given video V and audio A, the dual encoders produce $F_V \in \mathbb{R}^{T_v \times D}$ and $F_A \in \mathbb{R}^{T_a \times D}$, where D is the shared embedding dimension and all frame features are ℓ_2-normalized. The token counts T_v and T_a are determined by the video frame rate and the audio spectral hop length; in practice video encoders produce more tokens per second, yielding $T_v > T_a$ for the same clip duration. Element-wise similarity requires sequences of identical length, which motivates the alignment step below.

Temporal alignment. The two encoders sample the timeline independently and share only one common coordinate system: physical time. Comparing the i-th video token with the i-th audio token without alignment is therefore meaningless, as they cover different time intervals. We align the sequences with two design choices, both grounded in the asymmetry between the modalities. First, we use audio as the *temporal anchor* and downsample video to length T_a: video is temporally redundant — adjacent frames are near-identical, so subsampling loses little — whereas audio carries transient synchronization cues that resampling would destroy. Second, we use *nearest-neighbor* rather than linear interpolation, so that every aligned token remains an original frame feature; linear blending would fabricate intermediate states with no physical counterpart. Both choices are validated against their alternatives (linear interpolation, repeat padding, and upsampling) in Table 3 (ii). Formally, for each audio time step $i \in \{1, \ldots, T_a\}$, the aligned visual feature is:

$$\tilde{F}_V^{(i)} = F_V^{(\phi(i))}, \quad \phi(i) = \min\Big(\mathrm{round}\Big(i \cdot \tfrac{T_v}{T_a}\Big), T_v\Big). \tag{2}$$

yielding an aligned visual sequence $\tilde{F}_V \in \mathbb{R}^{T_a \times D}$.

Frame-level similarity. With both modalities on the same timeline, the cross-modal similarity is the mean inner product over aligned frame pairs:

$$S_{AV} = \frac{1}{T_a} \sum_{i=1}^{T_a} \left(\tilde{F}_V^{(i)}\right)^\top F_A^{(i)}. \tag{3}$$

This traces the main diagonal of the $T_a \times T_a$ pairwise similarity matrix, enforcing temporal order by construction. Under a temporal shift of Δt, the true correspondences move to the Δt-th off-diagonal, so S_{AV} decays as the fraction of correctly aligned content shrinks — an effect empirically visible as the sharp unimodal peak at $\Delta t = 0$ in Fig. 3. This built-in penalty is what FATE exploits for synchronization-aware retrieval and evaluation, without any task-specific prediction head.

3.3 Training Strategy

FATE is trained with two complementary objectives: semantic contrastive loss $\mathcal{L}_{sem}$ that separates embeddings *across* videos, and temporal soft-contrastive loss $\mathcal{L}_{temp}$ that supervises fine-grained

alignment *within* each video. The total objective is their weighted sum, which together captures both *what* sounds and *when* it occurs.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{sem} + \lambda_2 \mathcal{L}_{temp}, \tag{4}$$

We detail each term below:

Semantic contrastive loss. We apply symmetric InfoNCE over a batch of B audio-visual pairs. For each video anchor v_i, its paired audio a_i is the positive and all other in-batch audios are negatives:

$$\mathcal{L}_{v2a} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{AV}(v_i, a_i)/\tau)}{\sum_{j=1}^{B} \exp(S_{AV}(v_i, a_j)/\tau)}, \tag{5}$$

with $\mathcal{L}_{a2v}$ computed symmetrically and $\mathcal{L}_{sem} = \frac{1}{2}(\mathcal{L}_{v2a} + \mathcal{L}_{a2v})$. Since frame-level embeddings limit the per-GPU batch size, we all-gather embeddings across G GPUs, expanding the negative pool from B to $B \times G$ at no extra per-GPU memory cost (Fig. 2(b)).

Temporal soft-contrastive loss. The semantic loss structures the embedding space *across* videos but provides no supervision *within* a video: it cannot distinguish a synchronized clip from a slightly shifted one. The temporal loss fills this gap by casting alignment as probability matching over candidate offsets. Given a 2-second video anchor v, we slide a 2-second window over its 10-second source video with a 0.5-second stride, yielding $K = 17$ candidate audio clips whose offsets Δt_k range from fully aligned to strongly misaligned. The supervision target can be any non-increasing function of the absolute offset; we instantiate it as a raised-cosine kernel,

$$y(\Delta t_k) = \begin{cases} \cos^2\left(\dfrac{\pi \Delta t_k}{2T_0}\right) & |\Delta t_k| \le T_0, \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

which decays smoothly from its peak at $\Delta t_k = 0$ to zero at the support boundary T_0. We set $T_0 = 1$ s so that the half-maximum point, 0.5 s, matches the psychophysical threshold below which humans can barely perceive audio-visual asynchrony (Vatakis & Spence, 2006). The smooth roll-off also matches the structure of the candidates: adjacent clips share 75% temporal overlap, so assigning them zero weight would force the model to separate similar content. For example, on our 0.5s grid, Eq. (6) yields relative weights $2:1:0$ for the aligned, adjacent, and distant candidates; the kernel width is further validated by ablation (Table 3(i)). Normalizing $p_k = y(\Delta t_k)/\sum_m y(\Delta t_m)$, we minimize cross-entropy against the predicted similarity distribution:

$$\mathcal{L}_{temp_v2a} = -\sum_{k=1}^{K} p_k \log \frac{\exp(S_{AV}(v, a_k)/\tau)}{\sum_{m=1}^{K} \exp(S_{AV}(v, a_m)/\tau)}. \tag{7}$$

$\mathcal{L}_{temp_a2v}$ uses audio clips as anchors, and $\mathcal{L}_{temp} = \frac{1}{2}(\mathcal{L}_{temp_v2a} + \mathcal{L}_{temp_a2v})$. This graduated supervision shapes a smooth alignment landscape in which similarity peaks at zero offset and decreases with displacement, mirrored by the empirical sensitivity curve in Fig. 3.

4 Experimental Settings

4.1 Training Specification

We initialize FATE from PE-AV-small (Vyas et al., 2025), discarding its text encoder and global pooling head, and fine-tune with LoRA (Hu et al., 2022), which updates only 2.3% of the parameters while keeping the encoder weights frozen. Training uses the VGGSound (Chen et al., 2020) training split ($\sim$183k 10-second clips across 309 sound categories), where each video is segmented into 2-second clips with a 0.5-second stride, yielding $\sim$17 clips per video. The model is trained for 5 epochs on 8 NVIDIA A800 GPUs; preprocessing and optimization details are listed in Appendix.

4.2 Evaluation Protocol

We evaluate FATE on temporal cross-modal retrieval, zero-shot event localization, and generation evaluation, probing semantic-temporal discrimination, cross-task transfer, and agreement with human synchronization judgments, respectively.

Temporal Cross-Modal Retrieval. Given an audio query a_q and a video segmented into N clips $\mathcal{V} = \{v_1, \ldots, v_N\}$, the goal is to retrieve the temporally aligned clip:

$$\hat{v} = \arg\max_{v_i \in \mathcal{V}} S_{AV}(v_i, a_q). \tag{8}$$

We evaluate on the test splits of AVSync15 (Zhang et al., 2024) (150 videos) and VGG-Sync (Chen et al., 2021) (595 videos), each consisting of $\sim$10 s videos. Both benchmarks are curated from VGGSound, so we verify by video ID that no test video appears in training split. Each video is segmented into 2s clips with a 0.5s stride, yielding $N{\approx}17$ candidates per video; one segment is selected as the ground-truth query, with graded relevance of 2 for the aligned segment, 1 for segments within $\pm$0.5 s, and 0 otherwise. We report Recall@k (R@k) and NDCG@k (N@k) in both Video-to-Audio (V2A) and Audio-to-Video (A2V) directions.

We design two retrieval settings to disentangle model capabilities. In the **Intra-Video** setting, the candidate pool contains only the $\sim$17 segments of the same video; since all candidates share identical semantic content, this purely tests fine-grained temporal discrimination. In the **Inter-Video** setting, we append all segments from $M{=}50$ randomly sampled distractor videos ($\sim$850 candidates per query), requiring the model to jointly identify the correct video and locate the aligned segment within it. Together, the two settings reveal whether a model possesses one capability or both.

We compare against two categories of baselines. The first consists of audio-visual foundation models, including ImageBind (Girdhar et al., 2023), LanguageBind (Zhu et al., 2023), CAV-MAE Sync (Araujo et al., 2025), PE-AV (Vyas et al., 2025), CAVP (Luo et al., 2023), and PEAVS (Goncalves et al., 2024), scored by cosine similarity between query and candidate. The second is the specialized synchronization model Synchformer (Iashin et al., 2024), which predicts temporal offsets rather than embeddings; we use its predicted probability at zero offset as the retrieval score. All baselines use officially released checkpoints.

Audio-Visual Event Localization. Given an audio query A_e and an video V, we divide V into 1s segments and slide a window of the query's duration over them, predicting the window with the highest average similarity:

$$\hat{t} = \arg\max_{t} \frac{1}{|w|} \sum_{s \in w(t)} S_{AV}(V_s, A_e). \tag{9}$$

We evaluate on the Audio-Visual Event (AVE) dataset (Tian et al., 2018), which contains unconstrained videos with second-level temporal annotations across 28 event categories, and report segment-level accuracy for A2V, V2A, and their average. We compare against DCCA (Andrew et al., 2013), AVDLN (Tian et al., 2018), and DAM (Wu et al., 2019). Notably, all three baselines are *supervised*, trained on the AVE training set with ground-truth temporal annotations, whereas FATE is evaluated *zero-shot*, trained only on VGGSound without any AVE supervision.

Joint Audio-Video Generation Evaluation. We evaluate FATE as an automatic metric for audio-visual synchronization by measuring agreement between its score $S_{AV}(V_{\text{gen}}, A_{\text{gen}})$ and human judgments using Spearman's rank correlation ρ. Using text prompts from the AVSync-15 test set, we generate 150 videos per model (750 in total) with five joint audio-video generation models: BridgeDiT (Guan et al., 2025b), JavisDiT (Liu et al., 2025), Ovi (Low et al., 2025), LTX-2 (HaCohen et al., 2026), and JointDiT (Wang et al., 2025b). Ten independent annotators rate each video's temporal synchronization on a 1–5 scale; their average gives its mean opinion score (MOS). The annotation protocol is detailed in Appendix A.3.

We compare FATE against DeSync (Iashin et al., 2024), PEAVS (Goncalves et al., 2024), AV-Align (Yariv et al., 2023), and CAVP (Luo et al., 2023) at two levels. *Sample-level* ρ compares metric scores with human MOS across the 150 videos from each generation model, testing whether the metric ranks individual videos consistently with humans. *Model-ranking* ρ first averages metric scores and MOS separately over each model's videos, then correlates the five pairs of averages to assess agreement on the overall model ranking.

Table 1: Temporal audio-visual cross-modal retrieval results. Intra-Video: candidates are segments from the same video ($\sim$17 per query), testing pure temporal discrimination. Inter-Video: candidates additionally include segments from M=50 distractor videos ($\sim$850 per query), requiring joint semantic and temporal discrimination. We report Recall@k (R@k) and NDCG@k (N@k) for Video-to-Audio (V2A) and Audio-to-Video (A2V) tasks. Best in bold, second best underlined.

	Avsync15								**VGG-Sync**							
	V2A				A2V				V2A				A2V			
Method	R@1	R@3	N@1	N@3	R@1	R@3	N@1	N@3	R@1	R@3	N@1	N@3	R@1	R@3	N@1	N@3
Intra-Video ($\sim$17 candidates per query)																
ImageBind (Girdhar et al., 2023)	4.89	16.22	8.89	15.49	6.67	16.89	12.00	15.47	6.19	19.19	10.31	16.50	7.25	20.67	10.70	17.35
LanguageBind (Zhu et al., 2023)	5.11	18.89	7.78	14.81	0.00	4.22	0.00	2.81	5.43	17.09	8.91	14.48	0.06	3.59	0.06	2.15
CAV-MAE Sync (Araujo et al., 2025)	5.56	18.22	9.56	14.87	4.00	14.00	6.22	11.62	5.15	15.29	9.30	13.42	4.15	14.68	7.96	12.35
PE-AV (Vyas et al., 2025)	5.33	18.00	9.33	15.56	6.22	17.33	11.11	15.95	7.17	20.28	11.88	17.56	5.77	17.93	9.02	15.23
CAVP (Luo et al., 2023)	6.44	16.22	9.56	14.28	3.33	14.22	6.89	11.80	4.99	15.97	7.45	12.82	4.76	15.97	7.79	12.95
PEAVS (Goncalves et al., 2024)	12.45	28.57	15.14	19.47	11.76	27.57	13.63	18.23	11.05	25.85	16.34	20.35	10.75	24.85	13.71	16.19
Synchformer (Iashin et al., 2024)	15.56	38.89	28.44	38.46	13.78	32.89	29.78	34.98	12.04	34.12	22.35	31.57	10.92	32.21	21.46	30.97
FATE (Ours)	**22.37**	**55.74**	**53.56**	**59.24**	**21.56**	**56.22**	**51.78**	**59.66**	**26.39**	**48.37**	**60.83**	**69.04**	**16.50**	**44.83**	**33.61**	**44.49**
w/o frame-level	9.78	29.11	18.67	27.73	9.11	29.33	18.00	25.75	12.72	34.68	22.80	32.82	10.76	28.68	19.27	26.55
Inter-Video ($\sim$850 candidates per query, M=50)																
ImageBind (Girdhar et al., 2023)	2.89	9.78	5.11	9.96	2.89	7.78	5.11	7.24	2.80	7.73	4.37	6.71	3.03	7.90	5.71	7.56
LanguageBind (Zhu et al., 2023)	2.35	4.80	6.15	8.90	1.95	4.10	5.60	7.45	3.10	5.65	7.80	10.20	2.75	4.90	6.45	8.80
CAV-MAE Sync (Araujo et al., 2025)	0.22	0.44	0.67	0.56	0.22	0.44	0.22	0.40	0.06	0.11	0.06	0.07	0.06	0.39	0.17	0.31
PE-AV (Vyas et al., 2025)	3.33	10.54	5.56	9.05	4.67	10.12	8.22	10.21	4.15	10.70	6.61	9.20	2.41	7.11	3.98	6.14
CAVP (Luo et al., 2023)	3.15	9.42	5.05	6.85	2.68	8.35	4.77	5.92	2.95	8.15	4.23	6.10	2.45	7.20	4.15	5.25
PEAVS (Goncalves et al., 2024)	3.68	12.64	6.44	10.29	2.99	10.34	5.75	9.23	5.05	15.85	7.98	13.35	4.75	14.85	8.74	13.19
SyncFormer (Iashin et al., 2024)	4.44	13.56	8.00	11.98	2.44	3.56	6.00	5.17	4.82	12.49	7.96	10.78	0.90	2.07	1.90	2.58
FATE (Ours)	**14.22**	**34.13**	**33.53**	**37.66**	**11.11**	**30.44**	**23.11**	**30.11**	**15.35**	**37.76**	**35.85**	**42.24**	**11.93**	**31.54**	**24.93**	**31.99**
w/o frame-level	6.67	19.33	12.44	18.83	6.44	20.44	12.22	17.54	7.90	22.02	14.06	20.87	7.23	19.44	13.28	18.34

5 Experimental Results

5.1 Temporal Cross-Modal Retrieval

Table 1 provides a controlled diagnosis of audio–visual retrieval: intra-video retrieval isolates *when* an event occurs, whereas inter-video retrieval requires identifying both *what* occurs and *when*.

Global pooling provides little usable temporal information. In the intra-video setting, all candidates share the same video-level semantics, and chance performance is 5.9% R@1. Global-embedding baselines remain near this level on AVSync-15 (e.g., 4.89 for ImageBind and 5.33 for PE-AV), whereas FATE reaches 22.37 R@1 and 55.74 R@3. The same trend holds on VGG-Sync and for A2V retrieval, showing that semantic similarity alone cannot reliably localize events in time.

Temporal supervision and frame-level structure provide complementary gains. With the same PE-AV backbone, applying our objectives without frame-level tokens raises intra-video V2A R@1 from 5.33 to 9.78 on AVSync-15. Restoring frame-level tokens further increases it to 22.37. The corresponding progression on VGG-Sync is similarly consistent: 7.17 $\rightarrow$ 12.72 $\rightarrow$ 26.39. This controlled comparison separates the two effects: temporal supervision introduces synchronization awareness, while frame-level alignment preserves the temporal resolution needed to exploit it.

Realistic retrieval requires both semantic grounding and temporal precision. With approximately 850 candidates, FATE achieves 34.13 V2A R@3 on AVSync-15 and 37.76 on VGG-Sync. The strongest baselines reach only 13.56 and 15.85, respectively, and the same ordering holds for A2V. For example, Synchformer decreases from 38.89 R@3 intra-video to 13.56 inter-video on AVSync-15 once semantic distractors are introduced. Thus, recognizing temporal offsets within a known video is insufficient: a transferable audio–visual embedding must represent both *what* event occurs and *when* it occurs.

5.2 Cross-Modal Event Localization

Table 2 evaluates zero-shot event localization on AVE and yields two findings. First, among zero-shot methods, temporal structure in the representation is the decisive factor. All four share VGGSound training data and see no AVE annotations, so their ordering isolates representation design: ImageBind's global embeddings barely ground events in time (21.6%), Synchformer's offset pre-

Table 2: Audio-visual event localization on AVE. The upper and lower blocks contain supervised and zero-shot methods, respectively. Best and second-best results are bold and underlined.

Method	A2V	V2A	Avg.
Supervised on AVE			
DCCA (2013)	34.8	34.1	34.5
AVDLN (2018)	44.8	35.6	40.2
DAM (2019)	<u>48.5</u>	**47.1**	<u>47.8</u>
Zero-shot on VGGSound			
ImageBind (2023)	23.4	19.8	21.6
Synchformer (2024)	28.5	25.1	26.8
PE-AV (2025)	34.2	29.4	31.8
FATE (Ours)	**51.8**	<u>44.7</u>	**48.3**
w/o frame-level	37.9	33.2	35.6

Table 3: Ablation of FATE's core design choices on AVSync-15 inter-video retrieval. Parentheses show relative drops from the full model.

Variant	V2A R@3	A2V R@3
FATE (full)	**34.13**	**30.44**
(i) Temporal label design		
Hard binary (0/1)	28.50 (−16.5%)	24.35 (−20.0%)
Soft label, 0.25 s	29.78 (−12.7%)	28.12 (−7.6%)
Soft label, 0.75 s	31.56 (−7.5%)	28.67 (−5.8%)
(ii) Alignment strategy		
Linear interpolation	30.12 (−11.7%)	26.55 (−12.8%)
Repeat padding	29.40 (−13.9%)	25.80 (−15.2%)
Audio→video	29.74 (−12.9%)	26.13 (−14.2%)
(iii) Loss components		
$\mathcal{L}_{\text{sem}}$ only	19.78 (−42.0%)	11.46 (−62.4%)
$\mathcal{L}_{\text{temp}}$ only	15.56 (−54.4%)	12.34 (−59.5%)

diction helps (26.8%), and FATE, which differs from PE-AV (31.8%) only in preserving frame-level structure, reaches 48.3%, a 16.5-point gain with an identical encoder. Second, the learned temporal correspondence transfers beyond its training domain. Without training on AVE, FATE achieves 48.3% average accuracy under our audio-query localization protocol, outperforming the evaluated zero-shot baselines. Compared with PE-AV using the same backbone, preserving frame-level temporal structure improves the average accuracy from 31.8% to 48.3%, demonstrating effective transfer of the learned temporal correspondence.

5.3 Joint Audio-Video Generation Evaluation

Table 4 evaluates agreement with human synchronization rankings. Since annotators rank five outputs within each prompt, we report two averaged Spearman correlations: *cross-prompt relative-rank* correlation, computed across prompts for each generator and then averaged over generators, and *within-prompt model-rank* correlation, computed among the fiveouputs for each prompt and then averaged over prompts. FATE performs best on both measures, achieving 17.24 in cross-prompt correlation and 44.41 in within-prompt correlation, compared with 14.15 for CAVP and 24.99 for AV-Align, respectively. This suggests that FATE more closely reproduces human relative rankings, particularly when comparing outputs generated from the same prompt. Nevertheless, the absolute correlations remain moderate, and each within-prompt correlation contains only five videos. We therefore interpret these results as evidence of relative ranking agreement rather than calibrated prediction of absolute synchronization quality.

Table 4: Agreement with prompt-conditional human synchronization rankings on AVSync-15. Results are reported as average Spearman correlation $\bar{\rho} \times 100$. Best and second-best results are bold and underlined.

Metric	Cross-prompt relative-rank $\bar{\rho} \times 100 \uparrow$	Within-prompt model-rank $\bar{\rho} \times 100 \uparrow$
DeSync (2024)	8.22	24.13
PEAVS (2024)	8.04	15.71
AV-Align (2023)	8.04	<u>24.99</u>
CAVP (2023)	<u>14.15</u>	14.21
FATE (Ours)	**17.24**	**44.41**

5.4 Ablation Study

Impact of Frame-Level Embeddings. To disentangle training from representation design, we compare FATE with the *w/o frame-level* variant in Table 1, trained using identical data, backbone, LoRA configuration, and budget. Replacing frame-level tokens with global pooling consistently reduced V2A R@3: from 55.74 to 29.11 intra-video and from 34.13 to 19.33 inter-video on AVSync-15, with the same trend on VGG-Sync (48.37 to 34.68 and 37.76 to 22.02). The pooled variant nevertheless slightly surpassed Synchformer on VGG-Sync intra-video R@1 (12.72 vs. 12.04), showing that the joint objective alone learns useful synchronization cues. Restoring frame-level structure raised R@1

to 26.39. Thus, the objective learns synchronization cues, while frame-level alignment provides the temporal precision needed for accurate retrieval.

Temporal Sensitivity Analysis. We plot S_{AV} against temporal offset Δt, averaged over all 150 AVSync-15 test videos. PE-AV with global pooling is nearly offset-invariant: even a 4-second shift yields almost the same score as perfect alignment. FATE instead exhibits a clear maximum at $\Delta t = 0$, followed by a steady decay in both directions. This behavior follows directly from frame-level matching (Eq. 3): shifting either modality breaks correspondence between physical moments and lowers the mean similarity. Its gradual decay further reflects the offset-aware soft labels in Eq. 6, suggesting that FATE captures decreasing temporal overlap rather than imposing a binary aligned–misaligned boundary.

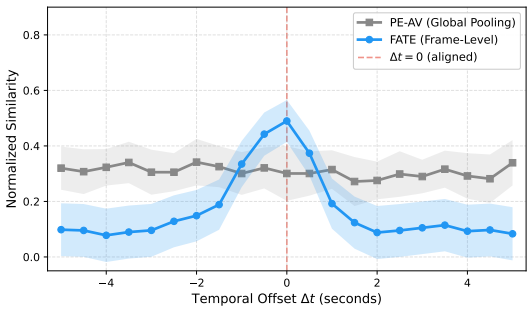

Figure 3: Temporal sensitivity of normalized audio–visual similarity. Shading denotes ± 1 standard deviation.

Effectiveness of Loss Components. Table 3 (iii) evaluates each loss term. Training with $\mathcal{L}_{sem}$ alone yields 19.78 V2A R@3: the model learns cross-video semantics but lacks within-video temporal precision. Training with $\mathcal{L}_{temp}$ alone performs worse still (15.56 V2A R@3), as it provides no semantic signal to separate different videos. Combining both reaches 34.13 V2A R@3, exceeding the sum of their individual gains over PE-AV. This super-additive effect indicates the two objectives are complementary: $\mathcal{L}_{sem}$ structures the embedding space across videos, while $\mathcal{L}_{temp}$ refines temporal ordering within each video.

Soft vs. Hard Temporal Labels. Table 3 (i) compares our raised-cosine soft targets, which yield relative weights of 2:1:0 at $|\Delta t| = 0$, 0.5 s, and $\geq$ 1 s, against hard labels (aligned=1, else=0). Hard labels degrade V2A R@3 from 34.13 to 28.50 and A2V R@3 from 30.44 to 24.35. The reason is straightforward: with a 0.5-second stride, adjacent segments share 75% temporal overlap with the anchor, and treating them as hard negatives forces the model to push apart near-identical content, introducing conflicting gradients. Soft labels instead assign intermediate supervision to these ambiguous neighbors, letting the model learn a smooth similarity decay rather than an abrupt boundary. Varying the soft-label window confirms the choice: a narrower 0.25 s window (29.78 V2A R@3) mislabels positives as negatives, a wider 0.75 s window (31.56) blurs the boundary.

Influence of Alignment Strategy. Table 3 (ii) evaluates alternatives to nearest-neighbor interpolation for aligning the higher-rate visual sequence (T_v tokens) to the lower-rate audio sequence (T_a tokens). Linear interpolation (30.12 V2A R@3) averages adjacent frames, producing blended features that correspond to no real frame and weaken temporal boundaries. Repeat padding (29.40) duplicates visual tokens to match T_a, skewing the similarity computation toward repeated content. Nearest-neighbor (34.13) selects the closest original frame for each audio time step, so every aligned token remains an unmodified representation. Reversing the resampling direction (audio→video, 29.74) also underperforms, as upsampling audio degrades transient synchronization cues, supporting audio as the proper temporal anchor.

6 Conclusion

We propose FATE, an audio-visual embedding model that unifies semantic correspondence and temporal synchronization within a single representation space. By preserving dense frame-level features, aligning them on the physical time axis, and training with a joint semantic-temporal objective, FATE avoids the information bottleneck imposed by global pooling. Experiments across three complementary tasks validate this design: FATE sets the state of the art on retrieval that demands both semantic and temporal discrimination, matches fully supervised methods on event localization in a purely zero-shot setting, and agrees most closely with human synchronization judgments as a generation evaluation metric. These results suggest that fine-grained temporal awareness can emerge from a general-purpose embedding space without task-specific prediction heads.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the clarity of the manuscript. They were not used to generate experimental results, conduct data analysis, or make scientific conclusions. All content was reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

We provide the complete model architecture, training objective, data preprocessing procedure, hyperparameters, and evaluation protocols in the main paper and supplementary material. Training details are provided in Appendix A.2, and the human annotation protocol is described in Appendix A.3. Anonymous source code is also provided with the submission.

ETHICS STATEMENT

Our human evaluation involves annotators assessing the audio-visual synchronization quality of generated videos. Annotators were informed of the evaluation task and compensated for their participation. No sensitive personal information was collected. The evaluated content was screened to avoid harmful or inappropriate material.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/andrew13.html.

Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R Glass, and Hilde Kuehne. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18794–18803, 2025.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Audio-visual synchronisation in the wild. *arXiv preprint arXiv:2112.04432*, 2021.

Xin Cheng, Xihua Wang, Yihan Wu, Yuyue Wang, and Ruihua Song. Lova: Long-form video-to-audio generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888085.

Xin Cheng, Yuyue Wang, Xihua Wang, Yihan Wu, Kaisi Guan, Yijing Chen, Peng Zhang, Xiaojiang Liu, Meng Cao, and Ruihua Song. Vssflow: Unifying video-conditioned sound and speech generation via joint learning, 2026. URL https://arxiv.org/abs/2509.24773.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Computer Vision – ACCV 2016 Workshops*, pp. 251–263. Springer, 2017.

GeminiTeam. Gemini: A family of highly capable multimodal models, 2023. URL https://arxiv.org/abs/2312.11805.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

Lucas Goncalves, Prashant Mathur, Chandrashekhar Lavania, Metehan Cekic, Marcello Federico, and Kyu J Han. Perceptual evaluation of audio-visual synchrony grounded in viewers' opinion scores. In *European Conference on Computer Vision*, pp. 288–305. Springer, 2024.

Kaisi Guan, Zhengfeng Lai, Yuchong Sun, Peng Zhang, Wei Liu, Kieran Liu, Meng Cao, and Ruihua Song. Etva: Evaluation of text-to-video alignment via fine-grained question generation and answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21299–21309, 2025a.

Kaisi Guan, Xihua Wang, Zhengfeng Lai, Xin Cheng, Peng Zhang, XiaoJiang Liu, Ruihua Song, and Meng Cao. Taming text-to-sounding video generation via advanced modality condition and interaction, 2025b. URL https://arxiv.org/abs/2510.03117.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3633890.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Sparse in space and time: Audio-visual synchronisation with trainable selectors. *arXiv preprint arXiv:2210.07055*, 2022.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329. IEEE, 2024.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 48855–48876, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

qwen team. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 247–263, 2018.

Ludovic Tuncay, Etienne Labbé, Emmanouil Benetos, and Thomas Pellegrini. Audio-jepa: Joint-embedding predictive architecture for audio representation learning. *arXiv preprint arXiv:2507.02915*, 2025.

Argiro Vatakis and Charles Spence. Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, 1111(1):134–142, 2006.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matt Le, Piotr Dollár, Christoph Feichtenhofer, et al. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning. *arXiv preprint arXiv:2512.19687*, 2025.

Xihua Wang, Yuyue Wang, Yihan Wu, Ruihua Song, Xu Tan, Zehua Chen, Hongteng Xu, and Guodong Sui. Tiva: Time-aligned video-to-audio generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 573–582, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681027. URL https://doi.org/10.1145/3664647.3681027.

Xihua Wang, Xin Cheng, Yuyue Wang, Ruihua Song, and Yunfeng Wang. Vaflow: Video-to-audio generation with cross-modality flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11777–11786, October 2025a.

Xihua Wang, Ruihua Song, Chongxuan Li, Xin Cheng, Boyuan Li, Yihan Wu, Yuyue Wang, Hongteng Xu, and Yunfeng Wang. Animate and sound an image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23369–23378, 2025b.

Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6292–6300, 2019.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095969. URL https://doi.org/10.1109/ICASSP49357.2023.10095969.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation, 2023. URL https://arxiv.org/abs/2309.16429.

Lin Zhang, Shentong Mo, Yijing Zhang, and Pedro Morgado. Audio-synchronized visual animation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2023.

A Supplementary Material for FATE

A.1 Efficiency Analysis

Table 5 compares the per-video inference time and embedding storage on AVSync-15. FATE (10.8 s) is nearly identical to its backbone PE-AV (9.6 s), as both share the dual-encoder architecture. In contrast, Synchformer requires 168.0 s per video—roughly 15× slower—because its single-encoder architecture cannot pre-extract modality-independent embeddings, requiring a full forward pass for every query. This highlights a key advantage of FATE: it achieves superior temporal discrimination over Synchformer while maintaining better inference efficiency. Retaining frame-level embeddings increases per-video storage from 136 KB to 3.3 MB, since temporal information must be kept rather than pooled away. This remains an order of magnitude smaller than the raw video itself and is a justified trade-off for the fine-grained temporal capability that pooling-based models cannot provide, while query-time inference speed is preserved.

Method	**Storage (KB)↓**	**Time (s)↓**
Global Pooling (PE-AV)	**136.0**	**9.6**
SynchFormer (Iashin et al., 2024)	–	168.0
FATE (Ours)	3400.0	10.8

Table 5: Efficiency comparison on AVSync-15: average per-video inference time (in seconds) for feature extraction and cross-modal similarity computation, and per-video embedding storage (∼10 s videos, single NVIDIA A800 GPU). SyncFormer does not produce reusable embeddings, hence no storage entry.

A.2 Training Details

Table 6 lists the complete preprocessing and optimization configuration. Video frames are resized to 336 × 336 and audio is resampled to 48 kHz before feature extraction.

A.3 Human Annotation Protocol

Task setup. We collect human judgments on 750 generated videos, organized into 150 groups corresponding to the 150 text prompts from the AVSync-15 test set. Each group contains five videos, one from each generation model, allowing annotators to inspect outputs generated from the same

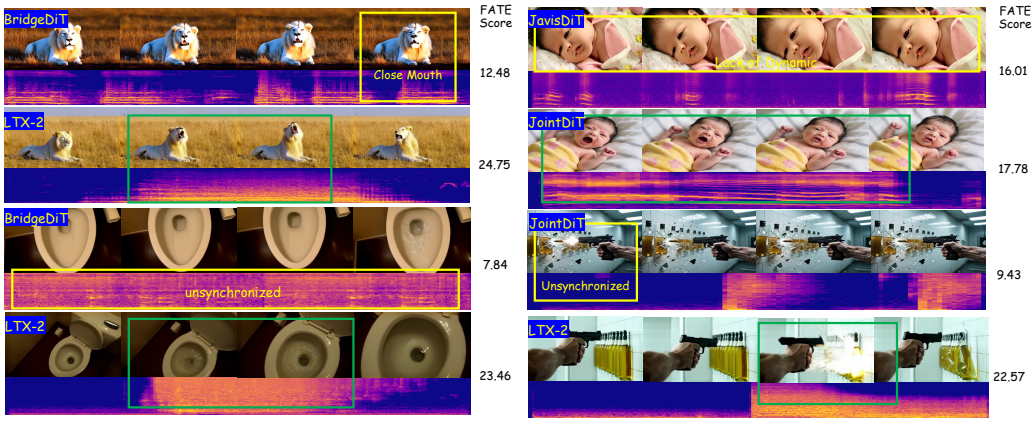

Figure 4: Qualitative visualization of FATE as a generation evaluation metric. For each case, we compare outputs from two generation models, showing key frames, mel-spectrograms, and FATE scores. FATE assigns higher scores to videos with well-aligned sound events (green boxes) and penalizes unsynchronized contents, closely matching human perceptual judgments.

Table 6: Training hyperparameters of FATE.

Component	Setting
Backbone	PE-AV-small (visual + audio encoders)
Video resolution	336×336
Audio sampling rate	48 kHz
Clip segmentation	2 s window, 0.5 s stride
LoRA rank r / α	16 / 32
LoRA dropout	0.1
LoRA target modules	Q, K, V, O projections (all attention layers)
Trainable parameters	2.3%
Optimizer	AdamW (β_1=0.9, β_2=0.999)
Weight decay	0.0
Peak learning rate	1×10^{-4}
LR schedule	linear decay
Epochs	5
Precision	`bfloat16`
Per-GPU batch size	4
Gradient accumulation	2
GPUs	8 $\times$ NVIDIA A800
Loss weights λ_1 / λ_2	0.5 (semantic) / 1.0 (temporal)
Temperature τ	0.07

prompt. Within each group, the videos are randomly shuffled and assigned filenames `1.mp4`–`5.mp4`. Model identities are hidden, and file indices do not consistently correspond to particular models across groups, preventing annotators from identifying generators through presentation order.

Evaluation criterion. Annotators assess *audio-visual synchronization*, considering both temporal alignment and semantic consistency between sounds and visible events. Temporal alignment concerns whether a sound occurs at the same time as its corresponding visual event, without perceptible lead or lag. Semantic consistency concerns whether the sound is physically plausible for the visible event. For example, a barking sound should coincide with a dog's visible barking action; the presence of a dog alone does not establish temporal synchronization. Annotators are instructed to focus on these audio-visual relationships and disregard general content quality, visual aesthetics, and audio fidelity.

Annotation procedure. For each prompt, annotators view the five generated videos and assign a separate synchronization rating to each video on a 1–5 scale, with higher scores indicating better synchronization. Each video is independently evaluated by ten annotators. Let $r_{p,m}^{(k)} \in \{1, 2, 3, 4, 5\}$ denote the rating assigned by annotator k to the video generated by model m for prompt p, where $p \in \{1, \ldots, 150\}$ and $m \in \{1, \ldots, 5\}$. The mean opinion score (MOS) for this video is

$$\mathrm{MOS}_{p,m} = \frac{1}{10} \sum_{k=1}^{10} r_{p,m}^{(k)}. \tag{10}$$

Thus, each of the 750 videos receives one MOS, obtained by averaging its ten individual ratings. This video-level MOS serves as the human reference score in our correlation analysis. Annotators are encouraged to replay videos and scrub the timeline around corresponding sound and visual events when their alignment is difficult to judge.

A.4 Case Study

Fig. 4 provides qualitative evidence that FATE scores align with human perception of synchronization quality. We present four pairs of generated videos, each produced by two different generation models from the same text prompt. In the first example, BridgeDiT generates a lion video where the lion's mouth closes prematurely before the sound ends, receiving a low score of 12.48, while LTX-2

produces a version with well-aligned mouth movements and audio onsets, scoring 24.75. For the toilet flushing prompt (bottom-left), BridgeDiT generates audio that is clearly unsynchronized with the visual action (7.84), whereas LTX-2 produces a temporally coherent flush sequence (23.46). Beyond temporal misalignment, FATE also penalizes static visuals lacking dynamics (JavisDiT, 16.01) and missing sound-source correspondence (JointDiT, 9.43). Across all cases, the score rankings closely match human perceptual judgments, corroborating the quantitative results in Table 4.

A.5 Prompts for MLLM-based Evaluation

We benchmark FATE against two general-purpose omni-modal LLMs: `Qwen3-Omni-30B-A3B` (open-weight, served locally with vLLM) and `Gemini-3.6-Flash` (accessed through its official API). Both models receive the identical zero-shot prompt shown below; no model-specific tuning, few-shot exemplars, or chain-of-thought elicitation is used. The prompt isolates temporal synchronization from semantic relevance, forbids free-form explanation so that outputs can be parsed automatically, and fixes the output to a two-decimal scalar in $[0, 1]$ that is directly comparable to normalized FATE scores.

Each video is submitted as a single request with its native audio track and no additional text context. Decoding is greedy (temperature 0, top-p 1.0, maximum 8 output tokens) to eliminate sampling variance, and each video is queried once. Responses that fail to match the expected numeric format are re-queried once; the small number of persistent failures are excluded from the correlation analysis. Both models ingest the same preprocessed clips used for FATE, so no model receives an input advantage. We note that Qwen3-Omni samples video at 2 fps, which upper-bounds the temporal granularity it can in principle resolve; this sampling rate is a property of the released model and cannot be increased through prompting.

Evaluation prompt for Omni LLM

```
You are an expert evaluator of audio-visual temporal synchronization.

You will be given a short video clip with its audio track.  Your task
is to rate ONLY the temporal synchronization between the audio and the
visual content:  whether sound events occur at exactly the same
moments as their visible sources (e.g., a bark occurs while the dog's
mouth opens, a drum hit occurs at the moment of impact).
```

Scoring rules:

- `Output a single score between 0.00 and 1.00, with exactly two decimal places.`
- `1.00 = perfectly synchronized:  every audible event coincides with its visible cause.`
- `0.50 = partially synchronized:  some events align, others are noticeably shifted.`
- `0.00 = completely unsynchronized:  sound events and visible actions do not co-occur at all.`
- `Judge synchronization continuously; use the full range (e.g., 0.15, 0.38, 0.85), not only the anchor values above.`

Strict constraints:

- `Ignore semantic relevance.  Even if the sound type matches the scene (a dog video with barking audio), the score must be LOW if the timing is off (e.g., barking while the mouth is closed).`
- `Ignore video quality, aesthetics, audio fidelity, and content plausibility.`
- `If the sounding source is never visible on screen, output 0.00.`
- `Do not explain.  Output ONLY the score in the format:  X.XX`