SYNCREWARD: HUMAN-ALIGNED SYNCHRONIZATION REWARD FOR AUDIO-VISUAL GENERATION

Anonymous authors
Paper under double-blind review

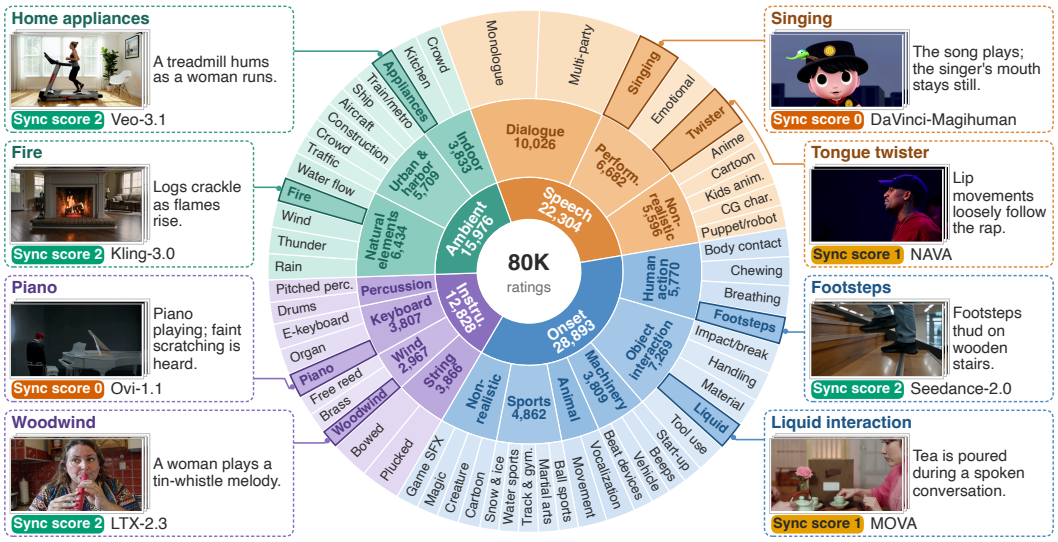

Figure 1: **Illustration of SyncReward-Data-80K**: human annotations of synchronization in generated audio–visual clips. The dataset contains 80K professional synchronization ratings on outputs from eight generators, covering four sound-event families: Speech, Onset, Instrument, and Ambient. A three-level taxonomy organizes these events into 16 classes and 58 fine-grained categories. Surrounding examples show sampled frames and brief descriptions of visible sound sources and their associated sounds, linked to the corresponding categories.

ABSTRACT

Recent years have witnessed rapid progress in joint audio–visual generation. However, audio–visual synchronization remains a challenge, as generated speech and sound often temporally misalign with visual events. A key bottleneck is the lack of human-annotated data, reliable evaluation benchmarks, and synchronization-aware evaluation models for generated audio–visual outputs. To address these limitations, we conduct a comprehensive study spanning data, benchmarking, and reward modeling. We first introduce **SyncReward-Data-80K**, a dataset containing 80K professional synchronization ratings for clips from eight audio–visual generators. A taxonomy-guided construction pipeline covers four sound-event families: Speech, Onset, Instrument, and Ambient. We further construct **SyncReward-Bench**, a held-out benchmark of 1K clips, each independently rated by multiple professionals. Using this benchmark, we systematically evaluate specialized synchronization metrics and omni-modal models, revealing their substantial misalignment with human judgments. Finally, we develop **SyncReward**, a reward model that combines semantic–temporal contrastive pretraining on real clips with human supervision on generated clips. As an evaluation metric, SyncReward achieves the strongest agreement with human ratings among the evaluated methods. As a reward, it improves synchronization through inference-time best-of-N selection and reinforcement learning based post-training. We will release the data, benchmark, model, and code. Additional demos are available at `https://anonymous.4open.science/r/SyncReward`.

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.

1 INTRODUCTION

Recent advances in joint audio–visual generation (Low et al., 2025; HaCohen et al., 2026; SII-OpenMOSS Team et al., 2026; MiniMaxAI, 2026) have enabled models to produce realistic videos together with speech, music, and sound effects. However, even state-of-the-art models (Bytedance, 2025; Kuaishou, 2024; Google, 2025) still exhibit noticeable audio–visual synchronization failures. For example, a speaker's voice may continue while the mouth remains still, or an impact sound may occur before the corresponding object contact. These errors make generated videos less realistic. Detecting them is important for evaluating and improving audio–visual generation.

Existing approaches, however, remain limited for this purpose. Specialized synchronization metrics (Yariv et al., 2024; Iashin et al., 2024; Luo et al., 2023; Guan et al., 2026b) estimate temporal offsets, or build audio and visual representations. While useful, these proxy objectives may not fully reflect human judgments on generated audio–visual content, where synchronization failures can be semantic ambiguity, missing or hallucinated events rather than simple global temporal shifts. General-purpose omni-modal large language models (Xu et al., 2025a;b) offer strong semantic understanding, but are not trained to capture fine-grained temporal misalignments.

To address these limitations, we provide a comprehensive study covering data construction, benchmarking, and reward model design. First, we introduce **SyncReward-Data-80K**, a large-scale human-annotated dataset for studying synchronization in generated audio–visual content. A three-level sound-event taxonomy guides caption sampling and prompt refinement, covering Speech, Onset, Instrument, and Ambient events. We then collect outputs from eight proprietary and open-source audio–visual generation models, capturing diverse generation behaviors and failure patterns. Finally, professional annotators evaluate audio–visual synchronization with a three-level rating protocol, yielding human synchronization annotations for 80K generated clips. Overall, the dataset provides systematic semantic coverage, diverse model outputs, and reliable human supervision.

We further introduce **SyncReward-Bench**, a disjoint evaluation set of 1,000 generated clips across sound-event categories and generators, each independently rated by multiple professional annotators. Using this benchmark, we comprehensively evaluate synchronization metrics, off-the-shelf omni-modal judges, and fine-tuned omni-modal models. Extensive experiments show that existing evaluators remain limited, while fine-tuning on human-rated generated clips substantially improves agreement with human judgments, highlighting the value of task-specific annotations.

Building on `SyncReward-Data-80K`, we develop **SyncReward**, a reward model for both synchronization evaluation and reward-guided generation. SyncReward first learns general temporal correspondence from real audio–visual clips via semantic–temporal contrastive learning, and is then trained to predict human synchronization ratings on generated clips. **As an evaluation metric**, it achieves the strongest agreement with human ratings across all four sound-event families on `SyncReward-Bench`, outperforming larger omni-modal models fine-tuned on the same data. **As a reward**, it guides inference-time Best-of-N selection and reinforcement-learning-based post-training, yielding better synchronization than the evaluated alternative rewards in both settings. These results demonstrate that a human-aligned synchronization reward can serve as both a reliable evaluator and an effective optimization signal for audio–visual generation.

Our contributions are threefold:

- **Balanced, diverse, and high-quality data.** We introduce SyncReward-Data-80K, with 80K professional annotations spanning eight audio–visual generators and four sound-event families.
- **Comprehensive benchmarking.** We build SyncReward-Bench and systematically benchmark specialized synchronization metrics, off-the-shelf omni-modal judges, and fine-tuned models.
- **Reliable reward modeling.** We develop SyncReward, which better aligns with human judgments and scales effectively to inference-time selection and generative post-training.

2 RELATED WORK

Audio-Visual Joint Generation. Existing joint audio–visual generators can be classified into two categories. Dual-tower models (Low et al., 2025; HaCohen et al., 2026; Guan et al., 2026a; SII-OpenMOSS Team et al., 2026; Ji et al., 2026) connect modality-specific branches through cross-

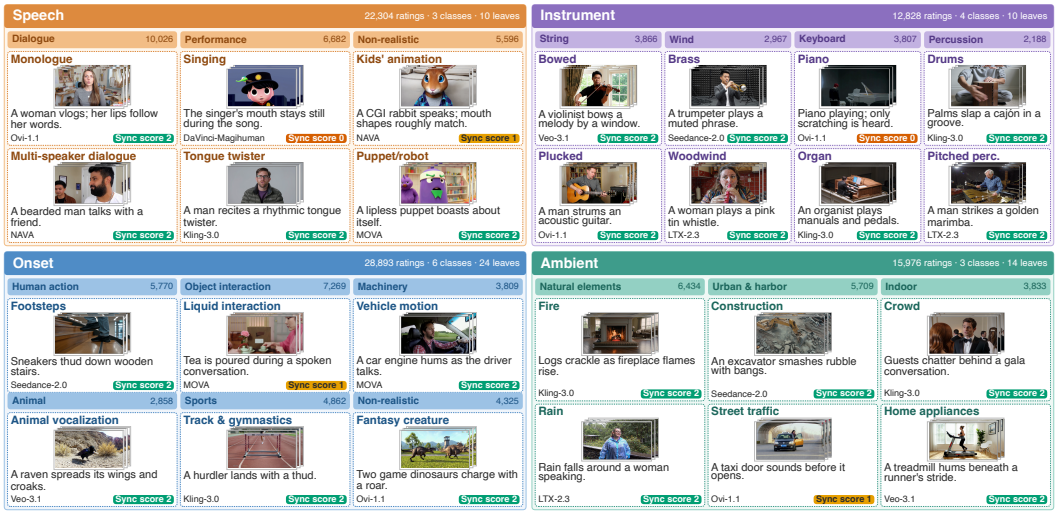

Figure 2: **Sound-event taxonomy and human-rated examples.** Examples span four sound-event families, 16 classes, and eight audio–visual generators. Headers report dataset-level rating counts for each family and class. Each card presents three sampled video frames, an event description, the generator name, and a human synchronization score. Scores indicate unsynchronized (0), partially synchronized (1), and fully synchronized (2) clips. Examples are selected to illustrate category coverage and different synchronization levels, rather than the dataset's score distribution.

modal layers, allowing pretrained components to be reused. Single-tower models (SII-GAIR et al., 2026; Sandai, 2026; MiniMaxAI, 2026) process audio and video tokens within a shared Transformer backbone. These architectures learn cross-modal dependencies, but their generative objectives do not directly optimize human judgments of synchronization. In this paper, we explore how to evaluate and improve audio–video synchronization without changing the generator architecture.

Reward Models for Generative Models. Reward models learn human preferences over generated outputs and provide objectives for optimization. ImageReward (Xu et al., 2023) and PickScore (Kirstain et al., 2023) establish this approach for text-to-image generation. Later work broadens the assessment beyond a single preference score: VisionReward (Xu et al., 2024) models multiple dimensions of image and video quality, while EditReward (Wu et al., 2026) targets instruction-guided image editing. VideoScore (He et al., 2024) and VideoAlign (Liu et al., 2025c) learn human assessments of generated videos, including visual quality, temporal consistency, and text alignment. These visual rewards do not directly assess whether sounds coincide with visible actions. In this work, we address this gap through a comprehensive study of audio–visual synchronization, including data construction, benchmarking, and reward model design.

Audio-Visual Synchronization Evaluation. Existing methods assess audio–visual synchronization from different perspectives. AV-Align (Yariv et al., 2024) compares audio onsets with visual motion peaks, whereas SparseSelector (Iashin et al., 2022) and Synchformer (Iashin et al., 2024) predict temporal offsets. Embedding-based methods assess cross-modal similarity, including SyncNet (Chung & Zisserman, 2017) for lip synchronization and CAVP (Luo et al., 2023) and FATE (Guan et al., 2026b) for broader audio–visual events. PEAVS (Goncalves et al., 2024) learns synchronization scores from human ratings. Beyond Time Shifts (Qian et al., 2026b) further adapts an omni-modal LLM using SynthSync, which pairs real videos with generated audio and human preferences. Despite these developments, existing works do not fully capture the native patterns of joint audio–video generators. We address this limitation with diverse generated outputs and high-quality human annotations, and further evaluate whether the learned score can improve generation.

3 SyncReward

We first present the task definition, then detail the data collection, construction, and annotation procedures, and finally introduce the design of the reward model.

Task Formulation. Given a condition c, such as a text prompt, a joint audio–visual generator produces a video clip and its accompanying audio, $x = (x^v, x^a)$. We evaluate whether sounds are aligned with their visible sources, such as speech with lip motion and impact sounds with object contact. Our evaluation focuses on clips whose audio is related to the visible scene and distinguishes synchronization from perceptual quality and prompt adherence. Our goal is to obtain a clip-level score $R(x^v, x^a) \in [0, 2]$ that agrees with human judgments of synchronization quality. The score serves both as an evaluation metric and as a reward for candidate selection and post-training.

3.1 SyncReward-Data-80K and SyncReward-Bench

Taxonomy. We use Qwen2.5-Omni-7B (Xu et al., 2025a) to caption real audio–visual clips from AudioSet (Gemmeke et al., 2017) and VGGSound (Chen et al., 2020), describing visible events and their accompanying sounds. We organize these descriptions into a three-level taxonomy comprising four event families, 16 intermediate categories, and 58 fine-grained subcategories. The four families are Speech, covering visible speaking and singing; Onset, covering sounds associated with visible actions and discrete events; Instrument, covering visible musical performances; and Ambient, covering environmental sounds with identifiable visual sources. Off-screen narration and background music are excluded from the target events. This taxonomy guides prompt sampling and refinement and defines the categories used for evaluation. The complete hierarchy and six additional language tags for Speech are provided in the Appendix.

Prompt Construction. The caption pool collected from AudioSet (Gemmeke et al., 2017) and VGGSound (Chen et al., 2020) is highly imbalanced across sound-event categories. We therefore map each caption to the taxonomy and resample the pool according to predefined category quotas, balancing category coverage with relevance to synchronization evaluation. This produces three disjoint sets of 10K captions and reduces the dominance of frequent categories in the original pool. We then use an LLM to rewrite each sampled caption into a generation prompt that explicitly describes the temporal correspondence between a visible sound source, its action, and the associated sound. The rewriting preserves the original scene where possible and removes sound descriptions that cannot be linked to a visible source. These prompts specify the intended synchronization, while the generated clips may still contain synchronization errors. The resulting category distribution is shown in Figure 1; the sampling quotas and rewriting instructions are provided in the Appendix.

Audio–Visual Generation. We synthesize audio–visual clips using eight close-source and open-source generators: Kling 3.0 (Kuaishou, 2024), Seedance 2.0 (Bytedance, 2025), Veo 3 (Google, 2025), LTX-2.3 (HaCohen et al., 2026), daVinci (SII-GAIR et al., 2026), MOVA (SII-OpenMOSS Team et al., 2026), NAVA (Ji et al., 2026), and Ovi (Low et al., 2025). The generation pool contains approximately 10K clips per model. Using multiple generators enables the dataset to cover a diverse range of synchronization error patterns rather than artifacts specific to a single generation system.

Human Annotation. Each generated clip is independently assessed by five professional annotators under a two-stage protocol, with generator identities hidden. Annotators first determine whether the audio is related to the visible scene, then assign related clips an ordinal synchronization rating $y \in \{0, 1, 2\}$, indicating clearly desynchronized, partially or uncertainly synchronized, and well synchronized, respectively. Finally, the constructed `SyncReward-Data-80K` contains 80K professional synchronization ratings. We aggregate annotations into bias-corrected consensus targets and construct training and validation splits with no shared $(\mathrm{uid}, \mathrm{model})$ pairs, where uid identifies the source prompt. Representative examples can be found in Figure 2. More details on the annotation guidelines, annotation platform, label aggregation are reported in the Appendix.

SyncReward-Bench. We further construct `SyncReward-Bench`, an independently annotated benchmark containing 1,000 generated clips. Samples are drawn across generators according to predefined taxonomy proportions and are strictly disjoint from `SyncReward-Data-80K`. Each clip is independently rated by multiple professional annotators, followed by cross-checking and score aggregation to obtain robust human reference scores. Spearman's rank correlation and pairwise ranking accuracy measure agreement with human ratings and their induced orderings, respectively. The Appendix details benchmark composition, annotation checks, and pair construction. Section 4.2 compares synchronization metrics and omni-modal judges.

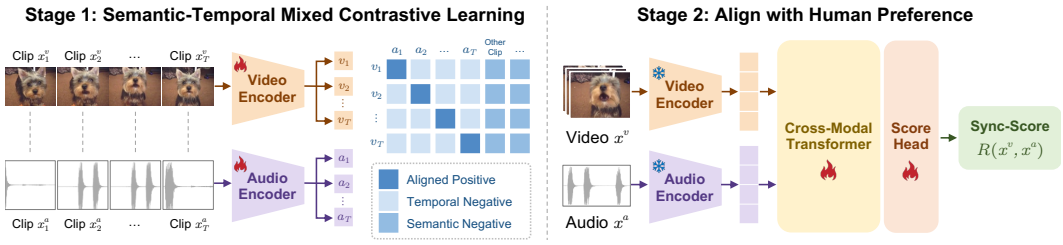

Figure 3: **Two-stage training of SyncReward.** Stage 1 trains the audio and video encoders through contrastive learning with aligned segment pairs, within-clip temporal negatives, and cross-clip semantic negatives from the same batch. Stage 2 freezes both encoders and trains a cross-modal Transformer and score head using human synchronization ratings, combining regression and pairwise ranking losses. The model outputs a clip-level synchronization score in $[0, 2]$. Flame and snowflake icons indicate trainable and frozen modules, respectively.

3.2 SyncReward Architecture and Two-Stage Training

SyncReward comprises audio and video encoders, a cross-modal Transformer, and a scalar score head. In Stage 1, we first train the encoders to capture semantic and temporal correspondence using real audio–visual clips. In Stage 2, we then freeze the encoders and train the remaining components on human-rated generated clips. The overall pipeline is illustrated in Figure 3.

Stage 1: Semantic–Temporal Mixed Contrastive Learning. We divide each real audio–visual clip into T temporally corresponding segments and encode them as $v_t = f_v(x_t^v)$ and $a_t = f_a(x_t^a)$. Video and audio segments from the same clip and temporal position form *aligned positives*. Segments from different temporal positions within the same clip serve as *temporal negatives*, while segments from different clips serve as *semantic negatives*. Together, these negatives encourage the encoders to distinguish both audio–visual semantic correspondence and temporal alignment.

We further introduce a batch index $b \in \{1, 2, \ldots, B\}$ for samples in a training batch, yielding $v_{b,t}$ and $a_{b,t}$. We then optimize a symmetric contrastive objective:

$$\mathcal{L}_{\mathrm{S1}} = \frac{1}{2}\left(\mathcal{L}_{v\to a} + \mathcal{L}_{a\to v}\right),$$
$$\mathcal{L}_{v\to a} = -\frac{1}{BT}\sum_{b=1}^{B}\sum_{t=1}^{T}\log\frac{\exp(\mathrm{sim}(v_{b,t}, a_{b,t})/\tau)}{\sum_{b'=1}^{B}\sum_{t'=1}^{T}\exp(\mathrm{sim}(v_{b,t}, a_{b',t'})/\tau)}. \tag{1}$$

Here, $\mathrm{sim}(\cdot)$ measures embedding similarity, τ is the temperature, and $\mathcal{L}_{a\to v}$ is defined by exchanging the two modalities. Each denominator includes the aligned positive and both temporal and semantic negatives. Within-clip negatives are particularly useful because recognizing shared semantic content alone is insufficient to distinguish temporal positions.

Stage 2: Align with Human Preference. We freeze the Stage-1 encoders and learn a cross-modal Transformer g_ϕ and score head r_ψ using `SyncReward-Data-80K`. For a generated clip x, the encoders produce video and audio sequences $V = \{v_t\}_{t=1}^T$ and $A = \{a_t\}_{t=1}^T$. We concatenate these sequences with a learnable reward token REW and a modality separator MOD:

$$H = g_\phi([\mathrm{REW}; V; \mathrm{MOD}; A]),$$
$$R_{\phi,\psi}(x) = 2\sigma(r_\psi(h_{\mathrm{REW}})). \tag{2}$$

The output h_{REW} aggregates information across temporal positions and modalities. The sigmoid $\sigma(\cdot)$ maps the scalar prediction to the human rating range $[0, 2]$.

We train the model with regression and pairwise ranking supervision. For a batch of B clips, let $R_b = R_{\phi,\psi}(x_b)$ and let y_b denote the aggregated human target. We derive pairwise ranking super-

Table 1: **SyncReward as a metric.** We report Spearman's rank correlation ($\rho \times 100$) and pairwise ranking accuracy (Pair Acc., %) for four on-screen sound categories: Speech, Onset, Instrument, and Ambient. Clips are grouped by the sound category used to construct their prompts. All comparisons are rank-based and require no score calibration. The best and second-best results are shown in **bold** and underlined, respectively. The Instrument subset is smaller because music-dominant clips were removed from the evaluation set.

Method	Params	Speech		Onset		Instrument		Ambient	
		Spearman ρ	Pair Acc	Spearman ρ	Pair Acc	Spearman ρ	Pair Acc	Spearman ρ	Pair Acc
Synchronization metrics									
AV-Align (Yariv et al., 2024)	–	4.4	51.8	7.0	47.3	14.9	43.2	15.3	55.9
CAVP (Luo et al., 2023)	114M	31.1	62.2	36.2	64.2	42.5	66.9	43.7	67.4
SparseSelector (Iashin et al., 2022)	55M	47.2	65.5	48.3	68.1	48.4	68.9	53.1	69.3
Synchformer (Iashin et al., 2024)	237M	32.9	60.4	47.3	67.9	32.0	61.9	56.5	70.7
PEAVS (Goncalves et al., 2024)	108M	16.5	56.5	1.6	50.6	5.0	47.9	12.7	45.0
Omni-LLM based									
Qwen3-Omni-30B-A3B	30BA3B	36.3	64.6	26.8	60.6	23.6	59.1	52.4	71.2
Qwen2.5-Omni-3B	3B	23.9	59.5	26.0	60.2	29.8	61.5	43.7	67.3
SynthSync (Qian et al., 2026b)	3B	4.9	51.9	3.1	51.2	21.9	58.5	15.1	55.7
Qwen2.5-Omni-3B-FT	3B	61.9	75.7	64.7	76.6	51.8	71.0	73.9	81.3
Qwen3-Omni-30B-A3B-FT	30BA3B	69.5	79.4	68.6	78.8	64.7	77.1	77.2	82.9
Ours									
SyncReward	1.6B	**79.7**	**85.1**	**70.5**	**79.9**	**68.1**	**79.3**	**82.1**	**85.9**

vision from these targets as $\mathcal{P} = \{(i, j) : 1 \le i, j \le B,\ y_i > y_j\}$. The objective is:

$$\begin{aligned} \mathcal{L}_{\mathrm{S2}} &= \mathcal{L}_{\mathrm{reg}} + \lambda_{\mathrm{rank}} \mathcal{L}_{\mathrm{rank}}, \\ \mathcal{L}_{\mathrm{reg}} &= \frac{1}{B} \sum_{b=1}^{B} \mathrm{SmoothL1}(R_b, y_b), \\ \mathcal{L}_{\mathrm{rank}} &= -\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \log \sigma(R_i - R_j). \end{aligned} \tag{3}$$

We set $\mathcal{L}_{\mathrm{rank}} = 0$ when no valid pairs are available. Regression fits the human rating scale, while the ranking term penalizes incorrect relative orderings. Only the cross-modal Transformer, its learnable tokens, and the score head are updated. Training settings are reported in Section 4.1.

4 Experiments

4.1 Experimental Setup

Base Model and Training Data. We initialize SyncReward from the Perception Encoder Audio–Visual checkpoint (`pe-av-large`) (Vyas et al., 2025). For Stage 1, we perform semantic–temporal contrastive learning on AudioSet (Gemmeke et al., 2017) and VGGSound (Chen et al., 2020), which contain totally 2.2M natural audio–visual clips. For Stage 2, we train the human-aligned reward model on `SyncReward-Data-80K`, which comprises 80K generated audio–visual clips with professional ordinal synchronization ratings. The clips are produced by the eight proprietary and open-source audio–visual generators described in Section 3.1.

Baselines. We compare SyncReward against two groups of baselines. First, we consider synchronization-specific metrics, including AV-Align (Yariv et al., 2024), CAVP (Luo et al., 2023), SparseSelector (Iashin et al., 2022), Synchformer (Iashin et al., 2024), and PEAVS (Goncalves et al., 2024). Second, we evaluate general-purpose omni-modal LLM judges, including Qwen2.5-Omni-3B (Xu et al., 2025a), Qwen3-Omni-30B-A3B (Xu et al., 2025b), and SynthSync (Qian et al., 2026b). To construct stronger data-matched baselines, we additionally fine-tune Qwen2.5-Omni-3B and Qwen3-Omni-30B-A3B on the same `SyncReward-Data-80K`.

Implementation Details. In Stage 1, each video is represented by 18 overlapping segments, each spanning 0.5 s with a stride of 0.25 s. We optimize the audio and visual encoders using AdamW with a learning rate of 1×10^{-5} on 4×8 NVIDIA H20 GPUs and a per-GPU batch size of 2. In Stage 2, we freeze both modality encoders and train only the cross-modal interaction Transformer

Table 2: **SyncReward as a reward.** We evaluate synchronization rewards for inference-time Best-of-16 selection and DiffusionNFT post-training. Human and Gemini preference scores are measured on matched held-out prompts, while complementary automatic metrics are computed on the same generated videos. The current preference scores are provisional estimates from base-referenced comparisons and will be replaced by complete all-pairs evaluation.

Method	Pairwise Preference Score (%)		Automatic Metrics			
	Human ↑	Gemini ↑	SyncR. ↑	DeSync ↓	AV-Align ↑	JavisScore ↑
LTX 2.3 (Base Model)	32.31	47.97	0.935	0.327	0.306	0.191
Inference-time Best-of-16 selection						
AV-Align	36.15	40.83	0.953	0.300	**0.448**	0.184
Synchformer	38.90	41.16	0.985	**0.125**	0.293	0.192
SynthSync	52.57	54.20	0.991	0.267	0.292	0.197
SyncReward	**63.67**	**71.94**	**1.269**	0.205	0.313	**0.207**
DiffusionNFT post-training						
AV-Align	40.73	37.89	0.897	0.380	**0.420**	0.185
Synchformer	45.03	35.17	1.053	**0.185**	0.299	0.220
SynthSync	53.71	52.63	0.885	0.417	0.311	0.189
SyncReward	**65.78**	**67.86**	**1.129**	0.187	0.323	**0.223**

and scalar reward head. The reward head produces $R(x) = 2\sigma(\cdot) \in [0, 2]$. We optimize on 2×8 NVIDIA H20 GPUs, using a per-GPU batch size of 8.

Evaluation Protocol. We evaluate SyncReward on the proposed `SyncReward-Bench`, both as a synchronization metric and as a reward for improving audio–visual generation. For metric evaluation, we report Spearman's rank correlation ρ and pairwise ranking accuracy. Spearman's ρ measures monotonic agreement with human ratings. For pairwise accuracy, we construct clip pairs with different mean human ratings and treat the higher-rated clip as the preferred one; a prediction is correct if the metric recovers this ordering. For reward evaluation, we apply SyncReward to Best-of-16 selection and DiffusionNFT (Zheng et al., 2026) post-training. The resulting videos are evaluated using human and Gemini pairwise preferences together with four automatic synchronization metrics. Detailed protocols for the human evaluation and Gemini judge are provided in the appendix.

4.2 SyncReward as a Synchronization Metric

Table 1 compares all methods with human synchronization ratings across four sound categories in `SyncReward-Bench`. SyncReward achieves the highest Spearman correlation and pairwise accuracy in every category. Compared with the strongest baseline, Qwen3-Omni-30B-A3B-FT, it improves Spearman correlation by 1.9–10.2 points and pairwise accuracy by 1.1–5.7 points, with the largest margin on Speech. Methods trained mainly on real videos transfer less reliably to generated content: AV-Align, for example, falls below 50% pairwise accuracy on Onset and Instrument, while the other specialized metrics show only moderate agreement with human ratings. Fine-tuning the Qwen-Omni models on the same human-rated data produces large gains, showing that generated-video failures differ from general-purpose pretraining. Even with the same training data, both Qwen baselines remain below the smaller SyncReward model across all four categories.

4.3 SyncReward as a Reward

Setting. We evaluate each reward in two settings. For inference-time selection, we sample 16 LTX-2.3 candidates for each prompt and retain the candidate with the highest reward. For post-training, we use each reward as the optimization signal to fine-tune LTX-2.3 with DiffusionNFT (Zheng et al., 2026). Further implementation and evaluation details are provided in the Appendix.

Inference-Time Selection Table 2 reports the results of Best-of-16 selection. SyncReward achieves human and Gemini preference scores of 63.67 and 71.94, exceeding the next-best reward by 11.1 and 17.7 points, respectively. It also ranks first on SyncR. and JavisScore and second on DeSync and AV-Align. By comparison, AV-Align and Synchformer perform best on the metrics they directly target but receive much lower preference scores. Improvements on a specialized synchronization metric therefore do not necessarily identify the candidates preferred by human evaluators.

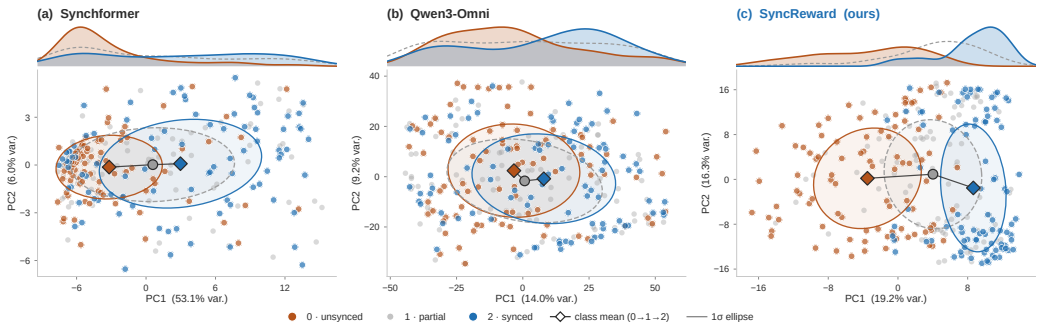

Figure 4: **Clip representations and human synchronization ratings.** Each panel shows a separate PCA projection fitted on validation features, with a class-balanced subset of held-out test clips colored by human synchronization ratings. Curves show class-wise PC1 densities; ellipses indicate one-standard-deviation covariance contours, and connected markers denote class means in rating order. Axis labels report explained variance, and coordinates are not shared across models.

Post-Training As shown in Table 2, SyncReward remains the strongest reward after Diffusion-NFT (Zheng et al., 2026) post-training, leading the next-best method by 12.1 points in human preference and 15.2 points under the Gemini judge. It also achieves the best SyncReward Score and JavisScore and ranks second on DeSync and AV-Align. AV-Align post-training raises its own metric from 0.306 to 0.420, but degrades the other three automatic metrics relative to the base model. Synchformer similarly obtains the lowest DeSync while producing substantially lower preference scores. The consistent results across selection and post-training show that SyncReward provides a more reliable optimization signal than other models.

4.4 Ablation and Analysis

Ablation Study. Table 3 examines the contributions of the training stages and objectives. The full model outperforms both single-stage variants, with human alignment improving Spearman correlation by 17.7 points and pairwise accuracy by 9.0 points over Stage 1 alone. Training Stage 2 directly from PE-AV yields substantially lower performance, supporting the importance of temporal pretraining before human alignment. Within Stage 1, removing temporal negatives causes a larger correlation drop than removing semantic negatives (7.6 versus 4.1 points). This comparison supports learning temporal correspondence beyond semantic similarity, while retaining both types of negatives gives the best results. In Stage 2, removing either regression or ranking supervision reduces performance. Omitting the ranking loss lowers pairwise accuracy by 3.8 points, compared with 1.6 points without regression, supporting explicit ranking supervision even when scalar ratings are available.

Table 3: **Key ablations of SyncReward.** We evaluate the contributions of the training stages and the semantic-temporal contrastive objective on SyncReward-Bench. Additional design ablations are reported in the Appendix.

Variant	Spearman $\rho \uparrow$	Pair Acc. (%) $\uparrow$
Training stages		
PE-AV initialization	26.6	60.4
Stage 1 only	60.5	74.6
Stage 2 only	40.9	66.3
Stage 1 ablation		
w/o semantic	74.1	81.4
w/o temporal	70.6	78.9
Stage 2 ablation		
w/o regression loss	75.4	82.0
w/o rank loss	73.8	79.8
SyncReward	**78.2**	**83.6**

Representation Analysis. We examine whether the dominant variation in clip representations reflects human synchronization judgments. We fit PCA on validation features and apply the projections to held-out test clips (Figure 4). Correlations are computed on all 1,000 test clips using unmodified representations. SyncReward's first principal component has a higher absolute Spearman correlation with human ratings (0.47) than those of Synchformer (0.29) and Qwen3-Omni (0.04). A linear probe trained on validation features also achieves higher test correlation with SyncReward

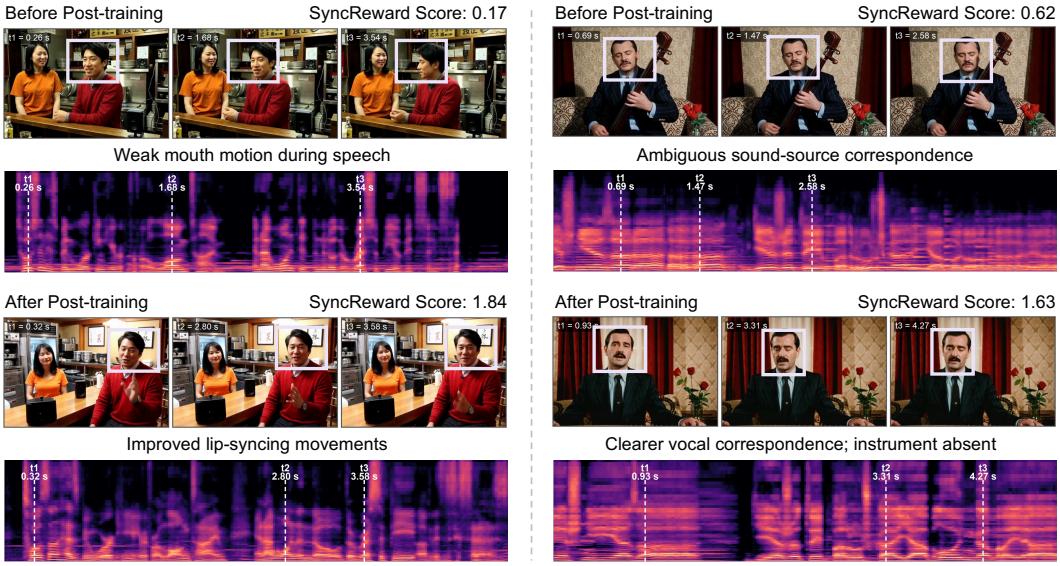

Figure 5: **Qualitative effects of SyncReward-guided DiffusionNFT post-training.** Each example compares outputs before (top) and after (bottom) post-training using three video frames and the corresponding log-mel spectrogram. Dashed lines mark frame timestamps t_1–t_3, and displayed scores are SyncReward predictions. Left: Post-training improves visible lip articulation during speech. Right: Vocal correspondence improves, but the instrument specified by the prompt is absent, illustrating a trade-off between synchronization and prompt fidelity.

features (0.41 versus 0.28 and 0.25). These results suggest that synchronization information is more linearly accessible in SyncReward's representation. However, Qwen3-Omni's low PC1 correlation does not rule out such information in other dimensions.

4.5 Case Study

Figure 5 presents two qualitative examples of SyncReward-guided DiffusionNFT post-training. In the dialogue example, post-training produces clearer mouth articulation during speech, while the SyncReward score increases from 0.17 to 1.84. The singing example shows clearer correspondence between vocal audio and the visible speaker, with the score increasing from 0.62 to 1.63. However, the post-trained output omits the instrument specified by the prompt, highlighting a potential trade-off between synchronization and prompt fidelity. This example motivates combining synchronization rewards with complementary objectives that preserve the requested scene content. Additional qualitative comparisons are available on our anonymous demo page.

5 Conclusion

In this paper, we introduce **SyncReward-Data-80K**, **SyncReward-Bench**, and **SyncReward** to support the evaluation and improvement of audio–visual synchronization. The dataset and disjoint benchmark provide professional synchronization annotations across diverse sound events and generators. Building on these resources, SyncReward combines temporal pretraining on real clips with human supervision on generated clips. As an evaluation metric, it achieves the strongest agreement with human ratings across all four sound-event families among the evaluated methods. As a reward, it yields higher human preference scores than the evaluated alternatives in both best-of-N selection and reinforcement learning based post-training. These results highlight the value of human-annotated generated content for learning effective synchronization rewards. Our case studies also show that optimizing synchronization alone can reduce prompt fidelity, motivating complementary objectives that preserve the intended content. We will release the data, benchmark, model, and code to support further research on audio–visual generation.

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 details of data construction, human annotation, model architecture, training, and evaluation in the main paper and appendix. Anonymous source code is provided with the submission, and the data, benchmark, model, and code will be released upon publication.

ETHICS STATEMENT

Professional annotators evaluated the synchronization quality of generated audio–visual clips and were compensated for their participation. No sensitive personal information was collected. The evaluated content was screened to avoid harmful or inappropriate material.

REFERENCES

Inc Bytedance. Seedance video generation, 2025. URL https://seed.bytedance.com/en/seedance2_0.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset, 2020. URL https://arxiv.org/abs/2004.14368.

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.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

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.

Inc Google. Veo3 video generation, 2025. URL https://deepmind.google/models/veo/.

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, 2026a. URL https://arxiv.org/abs/2510.03117.

Kaisi Guan, Bingzi Zhang, Xihua Wang, Ying Ba, Xin Cheng, Yijing Chen, and Ruihua Song. Fate: Frame-level audio-visual temporal embedding, 2026b. URL https://arxiv.org/abs/2608.01310.

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.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.127/.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. Vabench: A comprehensive benchmark for audio-video generation. *CoRR*, abs/2512.09299, 2025. doi: 10.48550/ARXIV.2512.09299. URL https://doi.org/10.48550/arXiv.2512.09299.

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.

Longbin Ji, Guan Wang, Xuan Wei, Chenye Yang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Jingzhou He. Native audio-visual alignment for generation, 2026. URL https://arxiv.org/abs/2605.30073.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.01569.

Inc Kuaishou. Kling video generation, 2024. URL https://klingai.com/.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8a57aa8e8b57e64a42e95f7dceb0adb9-Abstract-Conference.html.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online RL. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025a. URL http://papers.nips.cc/paper_files/paper/2025/hash/3a10c46572628d58cb44fb705f25cbbf-Abstract-Conference.html.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/76227feb18ea0ee40bd15cf02c33e18e-Abstract-Conference.html.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025c. URL https://arxiv.org/abs/2501.13918.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, and Tat-Seng Chua. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *CoRR*, abs/2503.23377, 2025d. doi: 10.48550/ARXIV.2503.23377. URL https://doi.org/10.48550/arXiv.2503.23377.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 8009–8019. Computer Vision Foundation / IEEE, 2025e. doi: 10.1109/CVPR52734.2025.00750. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_VideoDPO_Omni-Preference_Alignment_for_Video_Diffusion_Generation_CVPR_2025_paper.html.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation, 2025. URL https://arxiv.org/abs/2510.01284.

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.

MiniMaxAI. Minimax h3. 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-H3.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *CoRR*, abs/2407.08737, 2024. doi: 10.48550/ARXIV.2407.08737. URL https://doi.org/10.48550/arXiv.2407.08737.

Yijie Qian, Juncheng Wang, Chao Xu, Huihan Wang, Yuxiang Feng, Yang Liu, Baigui Sun, Yong Liu, and Shujun Wang. Beyond time shifts: Adapting omni-llm as a reference-free evaluator for generative audio-visual models. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part LV*, volume 17055 of *Lecture Notes in Computer Science*, pp. 39–58. Springer, 2026a. doi: 10.1007/978-3-032-37261-1_3. URL https://doi.org/10.1007/978-3-032-37261-1_3.

Yijie Qian, Juncheng Wang, Chao Xu, Huihan Wang, Yuxiang Feng, Yang Liu, Baigui Sun, Yong Liu, and Shujun Wang. Beyond time shifts: Adapting omni-llm as a reference-free evaluator for generative audio-visual models, 2026b. URL https://arxiv.org/abs/2607.09091.

Sandai. Magi-2 preview: Scaling video generation models efficiently. 2026. URL https://sand.ai/blog/magi-2-preview.

SII-GAIR, Sand. ai, Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, Lingzhi Li, Lyumanshan Ye, Min Hu, Qiangang Wang, Quanwei Qi, Steffi Chern, Tao Bu, Taoran Wang, Teren Xu, Tianning Zhang, Tiantian Mi, Weixian Xu, Wenqiang Zhang, Wentai Zhang, Xianping Yi, Xiaojie Cai, Xiaoyang Kang, Yan Ma, Yixiu Liu, Yunbo Zhang, Yunpeng Huang, Yutong Lin, Zewei Tao, Zhaoliang Liu, Zheng Zhang, Zhiyao Cen, Zhixuan Yu, Zhongshu Wang, Zhulin Hu, Zijin Zhou, Zinan Guo, Yue Cao, and Pengfei Liu. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

SII-OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyang Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu, Ke Chen, Shimin Li, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. Mova: Towards scalable and synchronized video-audio generation, February 2026. URL https://arxiv.org/abs/2602.08794. Technical report. Corresponding authors: Xie Chen and Xipeng Qiu. Project leaders: Qinyuan Cheng and Tianyi Liang.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matt Le, Piotr Dollár, Christoph Feichtenhofer, Ann Lee, and Wei-Ning Hsu. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning, 2025. URL https://arxiv.org/abs/2512.19687.

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *CoRR*, abs/2412.04814, 2024. doi: 10.48550/ARXIV.2412.04814. URL https://doi.org/10.48550/arXiv.2412.04814.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. EditReward: A human-aligned reward model for instruction-guided image editing. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=eZu358JOOR.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.05977.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. URL https://arxiv.org/abs/2412.21059.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b. URL https://arxiv.org/abs/2509.17765.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing GRPO on visual generation. *CoRR*, abs/2505.07818, 2025. doi: 10.48550/ARXIV.2505.07818. URL https://doi.org/10.48550/arXiv.2505.07818.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6639–6647, 2024.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 6463–6474. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00618. URL https://doi.org/10.1109/CVPR52733.2024.00618.

Jiacheng Zhang, Jie Wu, Weifeng Chen, Yatai Ji, Xuefeng Xiao, Weilin Huang, and Kai Han. Onlinevpo: Align video diffusion model with online video-centric preference optimization. *CoRR*, abs/2412.15159, 2024. doi: 10.48550/ARXIV.2412.15159. URL https://doi.org/10.48550/arXiv.2412.15159.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

Ziwei Zhou, Zeyuan Lai, Rui Wang, Yifan Yang, Zhen Xing, Yuqing Yang, Qi Dai, Lili Qiu, and Chong Luo. Avgen-bench: A task-driven benchmark for multi-granular evaluation of text-to-audio-video generation. *CoRR*, abs/2604.08540, 2026. doi: 10.48550/ARXIV.2604.08540. URL https://doi.org/10.48550/arXiv.2604.08540.

A Related Work

Audio–visual Datasets and Benchmarks. Recent benchmarks increasingly evaluate joint audio–visual generation from multiple dimensions. JavisBench (Liu et al., 2025d) evaluates generation quality, semantic consistency, and audio–visual spatio-temporal synchronization, while VABench (Hua et al., 2025) covers audio and visual quality, cross-modal consistency, synchronization, and lip–speech alignment. AVGen-Bench (Zhou et al., 2026) further provides multi-granular evaluation spanning perceptual quality, cross-modal alignment, and fine-grained semantic controllability. These benchmarks primarily evaluate generation systems rather than provide large-scale human supervision for learning synchronization rewards. SynthSync (Qian et al., 2026a) provides pairwise human synchronization preferences over generated audio conditioned on real videos. In contrast, SyncReward-Data-80K provides large-scale professional synchronization annotations on diverse jointly generated audio–visual outputs, while SyncReward-Bench provides an independently annotated benchmark for evaluating agreement with human synchronization judgments.

Post-Training for Video Generation. Post-training has become an important paradigm for improving video generation quality, temporal consistency, and alignment with human preferences. Early approaches directly optimize video generators with reward feedback, including InstructVideo (Yuan et al., 2024), VADER (Prabhudesai et al., 2024) and T2V-Turbo (Li et al., 2024). Another line of work learns from preference data: LiFT (Wang et al., 2024) adopts reward-weighted optimization, VideoDPO (Liu et al., 2025e) extends direct preference optimization to video generation, and OnlineVPO (Zhang et al., 2024) further performs preference optimization with samples generated by the evolving policy. With the increasing adoption of flow matching and rectified flow, recent methods have developed post-training algorithms tailored to flow-based video generators. Flow-DPO (Liu et al., 2025b) and Flow-RWR (Liu et al., 2025b) extend preference optimization and reward-weighted regression to rectified-flow models. DanceGRPO (Xue et al., 2025) and Flow-GRPO (Liu et al., 2025a) further introduce online reinforcement learning with group-relative rewards, enabling optimization using samples generated by the current policy. More recently, DiffusionNFT (Zheng et al., 2026) directly incorporates reward optimization into the native diffusion or flow-matching objective by contrasting high- and low-reward samples, avoiding explicit likelihood estimation over reverse-generation trajectories.

B SyncReward-Data-80K Details

B.1 Sound-Event Taxonomy

We organize on-screen sound events into a three-level hierarchy with four event families, 16 intermediate categories, and 58 fine-grained subcategories (Table 4). The hierarchy specifies the sound event targeted during prompt construction. Categories therefore describe the intended event rather than whether a generated clip successfully depicts it or achieves synchronization.

Event families. *Speech* covers speaking and singing by visible human or animated characters, including stylized characters whose articulation may differ from human lip motion. *Onset* covers action-linked sounds and discrete acoustic events, such as footsteps, impacts, animal calls, and device sounds. Here, "Onset" is a category name rather than a restriction to instantaneous sounds. *Instrument* covers sounds produced by visible musical performances. *Ambient* covers environmental sounds associated with visible sources or processes, including rain, traffic, and indoor activity. Across all four families, the target sound must be grounded in the visible scene; off-screen narration and background music are excluded.

Language tags. Speech prompts additionally receive one of six language tags: English, Mandarin Chinese, Japanese, Korean, Cantonese, or Other. Language is an independent attribute rather than an additional level of the hierarchy. For example, a Japanese-anime dialogue can be tagged as English, while a monologue can be tagged as Cantonese. These tags do not increase subcategories.

Category interpretation. The taxonomy combines event type with selected scene and character attributes to support prompt coverage. Its categories are not intended to be mutually exclusive descriptions of every sound in a clip. A street scene may contain both footsteps and traffic, for

Table 4: **The SyncReward sound-event taxonomy.** The three levels comprise four event families, 16 intermediate categories, and 58 fine-grained subcategories. Numbers in parentheses indicate the number of subcategories in each category. Language tags are recorded separately for Speech.

Family	Category	Subcategories
Speech	Conversation (2)	Monologue; multi-speaker conversation.
	Vocal performance (3)	Singing; expressive speech; challenging tongue twisters.
	Stylized characters (5)	Japanese anime; Western cartoons; children's animation; 3D film/game characters; puppets and robots without conventional lips.
Onset	Human actions (4)	Body contact; eating and chewing; breathing actions; footsteps.
	Object interactions (5)	Impacts and breakage; object manipulation; material processing; liquid interactions; tool use.
	Mechanical and electronic events (4)	Device startup; electronic notification sounds; vehicle operation; rhythmic mechanisms.
	Animals (2)	Animal vocalizations; animal movements.
	Sports (5)	Ball sports; combat sports and martial arts; athletics and gymnastics; water sports; ice and snow sports.
	Fictional events (4)	Cartoon character actions; fantastical creatures; magical and energy effects; in-world game feedback.
Instrument	Strings (2)	Plucked strings; bowed strings.
	Winds (3)	Woodwinds; brass; free-reed instruments.
	Keyboards (3)	Piano; organ; electronic keyboards.
	Percussion (2)	Drums; pitched percussion.
Ambient	Natural processes (5)	Rain; thunderstorms; wind; fire; flowing water.
	Urban and harbor environments (6)	Street traffic; crowds; construction; aircraft; ships and boats; trains and subways.
	Indoor environments (3)	Household appliances; kitchen activity; indoor crowds.

example. The category used for evaluation is the target event selected during prompt construction: footsteps belong to Onset, whereas the surrounding traffic belongs to Ambient. Similarly, Speech targets vocal articulation, while Instrument targets the visible act of playing an instrument. Fictional events remain subject to the same visible-source requirement; an audible effect must correspond to an on-screen action or event.

B.1.1 Human Annotation

Human annotation guidelines for audio-visual synchronization

This task evaluates audio-visual temporal synchronization. For each generated clip, annotate the audio and video in two steps.

Step 1: Is the audio related to the video?

- 0 = Unrelated: the audio is completely unrelated to the visual content (e.g., the video shows one kind of content while the audio contains an irrelevant sound).
- 1 = Related: the audio semantically corresponds to the visual content, so its synchronization can be assessed.

Step 2: Synchronization score (only if Step 1 = 1):

- 0 = Not synchronized: lip motion or actions clearly do not match the sound.
- 1 = Partially synchronized: audio and video are partly consistent but not aligned, with noticeable delay, lead, or unstable timing.
- 2 = Perfectly synchronized: the visible sound source and the audio are highly consistent in both content and timing.

Notes:

- If Step 1 = 0, no synchronization score is given and the clip is submitted directly.
- A free-text comment is optional.
- Wear headphones and turn the volume up.

B.2 Prompt Construction

This section describes how generation prompts are derived from source captions: taxonomy assignment, category-balanced sampling, and synchronization-oriented rewriting. All three steps use Qwen3.6-35B-A3B served with vLLM, with thinking mode disabled. The full instructions are reproduced below.

Source captions and taxonomy assignment. Source captions describe real video clips and the sounds they contain. We classify every sound mentioned in a caption into the taxonomy of Appendix B.1 with greedy decoding (temperature 0), processing eight captions per request. The system prompt contains the complete taxonomy guideline (Box B.2). For each sound, the model outputs its three taxonomy levels, whether it is on screen, a short piece of textual evidence, and a confidence level. Speech sounds additionally receive one of the six language tags. A caption may contain several independent sounds, and the model is instructed not to merge multiple sources into one class.

Category-balanced sampling. The caption pool is dominated by a few frequent categories, such as conversation and street ambience. We therefore resample captions with per-subcategory quotas that trade off taxonomy coverage against relevance for synchronization evaluation. This yields three disjoint sets of 10K captions. Each sampled caption keeps its target subcategory, which serves as the evaluation category of every clip generated from it (Appendix B.1).

Synchronization-oriented rewriting. Source captions often attribute sounds to off-screen sources, for example narration, off-screen voices, or background music. Such sounds cannot be evaluated for audio–visual timing. We rewrite each sampled caption with the instruction in Box B.2 (temperature 0.7, top-p 0.8, at most 1,024 output tokens). The instruction asks the model to (i) enumerate every sound in the caption; (ii) keep on-screen sounds and state explicitly how the visible source's action corresponds to the sound, e.g., lips moving with speech or a spoon stirring with the stirring sound; (iii) reassign off-screen speech to a character already visible in the frame, adding a visible speaker only when no suitable character exists; (iv) delete sounds that cannot have a visible source, such as background music; and (v) preserve the original camera, setting, and character descriptions. Rewritten prompts are written in English, while spoken lines keep their original language. This way the language tag of Speech prompts is preserved.

Scope. The prompts specify the *intended* audio–visual correspondence; they do not guarantee that it is realized. Generated clips may still contain the synchronization errors that SyncReward is designed to detect. Moreover, the target subcategory records the event requested in the prompt, not the event that a generated clip actually depicts.

Taxonomy assignment prompt

System:
You are a sound-scene classifier for audio-visual captions. Classify the sounds in the caption strictly following the taxonomy guideline below.

Requirements:

- Output a single valid JSON object only; no Markdown, no explanation.
- If the caption contains several independent sounds, output one entry per sound; if no audio information can be inferred, output {"sounds": []}.

- level1/level2/level3 must preferably use existing taxonomy nodes; if a new node is needed, set ontree=false and explain in reason.
- Do not merge multiple sources into a compound class, and do not repeat the caption text.
- Keep evidence within 12 characters and reason within 18 characters; output only the required fields.
- If level1 = Speech: level2/level3 must come from the Speech subtree, and a language field must be output, one of English / Mandarin Chinese / Japanese / Korean / Cantonese / Other. Default to English if no language is stated and no non-English line is quoted. Singing, tongue twisters, and animated characters also receive a language tag.
- Non-speech sounds must not have a language field.

{full taxonomy guideline}

Output schema: {"sounds": [{"sound_id", "level1", "level2", "level3", "ontree", "on_screen", "evidence", "confidence", "reason"}]}

User:
Classify the following caption. Output only the JSON object.
CAPTION: {caption}
JSON:

Synchronization-oriented rewriting prompt

You are an audio-visual caption rewriting assistant. You will be given a caption describing a video clip's visuals and the sounds in it. Rewrite it according to the rules below and output in English.

Concepts (understand first):

- On-screen sound: the person/object making the sound is visible in the frame AND its visible action matches the sound (e.g., you see lips moving while hearing speech, or see a dog panting while hearing it).
- Off-screen sound: the source is NOT visible (e.g., ``an off-screen voice says...'', ``a narrator says...'', background music with no visible source).
- Audio-visual consistency: every sound traces to a visible source whose action matches the sound.

Goal: After rewriting, every sound described must be on-screen and audio-visually consistent: wherever there is a sound, a visible person/object in the frame must be its source, with its action matching the sound.

Steps (in order):

- Step 1: List every sound mentioned (speech, dialogue, animal sounds, ambient sounds, music, etc.).
- Step 2: For each sound, decide if its source is visible (on-screen) or not (off-screen). On-screen with matching action: keep it, and make the ``visible source + action matches sound'' relationship explicit (e.g., ``his lips are moving'', ``the boom mic points at the speaker'', ``the dog pants as it is petted''). Off-screen: rewrite using the partner method (Step 3). Off-screen ambient/music with no possible visible source (e.g., background music): DELETE it entirely.
- Step 3: Partner method --- reassign the line/sound to a person already visible in the frame, so they become its on-screen source, and state that this person is producing it (lips moving

/ facing the other / action matching). If no suitable person exists, you may add a visible speaker, describing their appearance, position, and speaking action.
- Step 4: You may reword the caption and add reasonable details (appearance, position, action, sound direction) as long as all sounds end up on-screen and consistent. Do not invent sounds unrelated to the visuals.

Hard rules:

- Output in English, but keep all spoken dialogue in its ORIGINAL language --- do not translate the words people say.
- The caption description must be in English even though the input caption is in Chinese; only the spoken dialogue stays in its original language.
- No off-screen/voice-over/invisible-source sounds may remain; delete unfixable ones (e.g., background music).
- Every remaining sound must map to a visible source with matching action.
- Keep the original descriptions of camera angle, movement, setting, and characters.
- Write a single, smooth, coherent paragraph. Output ONLY the rewritten caption --- no explanations or steps.

Example:

Input (originally in Chinese): In a medium shot, a man in a baseball cap is petting a large dog. Soft background music plays. An off-screen voice asks, ''Is Simon home?'', and a woman at the door answers, ''He's at work.''

Output: A medium shot shows a man in a baseball cap petting a large dog. On screen, the man in the baseball cap, his lips moving, asks, ''Is Simon home?''† A woman in the doorway, facing him on camera, replies, ''He's at work.''† A boom mic held by a crew member points toward both speakers, and the dog pants audibly as it is petted, its panting matching its visible movement.

(Note: the off-screen voice was reassigned to the visible man --- the partner method --- the dialogue stays in its original language, and the background music was deleted.)

† Spoken in Chinese in the actual prompt; translated here for readability.

Now begin:

Rewrite this caption:
{caption}

C SyncReward Details

C.1 Detailed Architecture of Syncreward

The encoder is Meta's Perception Encoder AV (`pe-av-large`), run in bfloat16. Each clip is divided into $S{=}18$ segments of 0.5 s with a stride of 0.25 s (half overlap), so the interaction module receives $2S{+}2 = 38$ tokens, $[\texttt{REW}; v_{1:S}; \texttt{MOD}; a_{1:S}]$. Audio is resampled to 48 kHz. Segment embeddings are the mean-pooled last hidden states of the encoder, with its projection heads retained. Audio and video embeddings are linearly projected to width 768 and processed by a 3-layer Pre-LN Transformer (8 heads, FFN ratio 4, dropout 0.1); the reward is read out from the `REW` token by a two-layer MLP followed by a sigmoid rescaled to $[0, 2]$.

C.2 Stage 1: temporal contrastive pretraining

Stage 1 is trained on real videos from AudioSet and VGGSound. We score every clip with ImageBind audio–visual similarity and discard clips scoring below 0.2, which removes most videos whose soundtrack is unrelated to the visual content (e.g., background music or voice-over). We optimize

Table 5: **SyncReward training hyperparameters.** Settings for temporal pretraining (Stage 1) and human alignment (Stage 2).

Setting	Stage 1	Stage 2
Data and inputs		
Training data	AudioSet + VGGSound	SyncReward-Data-80K
Filtering / labels	ImageBind similarity ≥ 0.2	Human ratings: $0/1/2$
Clip duration	4.75 s	4.75 s
Segmentation	18×0.5 s	18×0.5 s
Segment stride	0.25 s	0.25 s
Audio sample rate	48 kHz	48 kHz
Model		
Encoder	`pe-av-large`	`pe-av-large`
Encoder updates	Trainable	Frozen
Trainable modules	Encoders + projection heads	Interaction + score head (23.5M)
Interaction Transformer	–	3 layers, $d = 768$, 8 heads
Training objective		
Loss	Symmetric InfoNCE	Smooth L_1 + 0.1 $\mathcal{L}_{\text{rank}}$
Temperature	Learned, initialized at 0.07	–
Smooth L_1 parameter	–	$\beta = 0.1$
Ranking pair threshold	–	$\lvert\Delta r\rvert \geq 0.75$
Optimization		
Optimizer	AdamW	AdamW
Learning rate	1×10^{-5}	1×10^{-5}
Cosine decay	–	Enabled
Warmup steps	1,000	100
Gradient clipping	1.0	–
Batch size per GPU	2 (sub-batch: 6)	8
Hardware	32× H20	16× H20
Global batch size	64	128
Training duration	5,500 steps	3 epochs
Checkpoint selection	Final step	Best validation Spearman ρ

a symmetric InfoNCE loss over the flattened $[B{\times}S,\ D]$ audio–visual similarity matrix, where co-occurring segments are positives and all other segments on the same GPU serve as negatives; the temperature is initialized at 0.07 and learned. We use AdamW with learning rate $1{\times}10^{-5}$, 1,000 warmup steps, gradient clipping at 1.0, a per-GPU batch of 2 clips, and an encoding sub-batch of 6. We report the step-5,500 checkpoint.

C.3 Stage 2: human alignment (reported model)

We initialize the encoder from Stage 1 and keep it frozen; the interaction module and reward head are randomly initialized and trained for 3 epochs on `SyncReward-Data-80K`. The validation and test splits are disjoint from training. Following Section 3, the objective is

$$\mathcal{L} = \text{SmoothL1}(\hat{r}, r;\ \beta{=}0.1) + \lambda\, \mathcal{L}_{\text{rank}}, \qquad \lambda{=}0.1, \tag{4}$$

where $r \in [0, 2]$ is the human synchronization score and $\mathcal{L}_{\text{rank}}$ is a softplus pairwise loss over in-batch pairs whose labels differ by at least 0.75. We use AdamW with learning rate $1{\times}10^{-5}$, 100 warmup steps, cosine decay, and a per-GPU batch of 8 on $2{\times}8$ H20 GPUs. The released checkpoint is selected by validation Spearman correlation.

D Details of SyncReward-Bench

D.1 Benchmark Construction and Annotation

Benchmark composition. SyncReward-Bench contains 1,000 generated clips from eight generators, with balanced representation across generators: 125 clips each. The benchmark covers four sound-event families: Speech (321), Onset (276), Instrument (169), and Ambient (234).

Human annotation. Each benchmark clip is independently assessed by three professional annotators following the same annotation guidelines. Annotators first assess audio–visual relatedness and then assign a synchronization score in $\{0, 1, 2\}$ to related clips. The evaluation target is the mean of the valid synchronization scores, without annotator-bias correction.

Inter-annotator agreement. On the 1000 clips with three valid synchronization scores, Fleiss' κ is 0.23. Pairwise annotator comparisons yield Spearman correlations of 0.45–0.50 and exact agreement rates of 52–55%. Opposite ratings (0 versus 2) account for 1.3–2.3% of paired judgments.

Evaluation metrics. We evaluate agreement with human ratings using Spearman's rank correlation and pairwise accuracy, accounting for each metric's score direction. Spearman's correlation compares metric scores with the aggregated human ratings. For pairwise accuracy, we consider all unordered pairs of distinct benchmark clips whose human ratings differ by at least 0.75, regardless of generator or prompt. Each pair receives 1 for a correct ordering, 0 for an incorrect ordering, and 0.5 for tied predictions. Pairwise accuracy is the average over all eligible pairs.

D.2 Pairwise judge prompt for Gemini

Pairwise judge prompt for Gemini-2.5-Pro

As a professional ``Text-to-Audible-Video'' quality inspector, your task is to assess the quality of two videos ([[RESPONSE A]] and [[RESPONSE B]]) generated from the same prompt ([[PROMPT]]). Each video carries its own generated audio track. Rate the quality of the AI assistant's response(s) according to the following criteria:

[Audio-Visual Synchronization] Judge ONLY how well the audio track of each video is synchronized with, and appropriate to, the visual content of that same video:

- (1) Temporal alignment. Does each sound event start, peak and stop at the same moment as the on-screen action that would physically produce it (impacts, footsteps, speech and lip movement, collisions, mechanical clicks, door closes, water splashes, instrument strikes, etc.)? Penalize audio that leads or lags the corresponding visual event, drifts out of phase over the clip, or is smeared across the wrong frames.
- (2) Semantic correspondence. Do the sounds that are present match the actions, objects and materials that are visible? Penalize sounds with no visible cause, visibly sound-producing actions that stay silent, and sounds of the wrong type, wrong material or wrong number of events.
- (3) Realization of the sound events described in [[PROMPT]]. The prompt states which sounds should occur and at what point; prefer the video whose audio actually produces those events at the right moments.

Strict constraints: You MUST ignore all of the following; none of them may influence your verdict: image quality, sharpness, resolution, bitrate, compression or codec artifacts, colour, exposure, lighting, aesthetics, motion smoothness, visual realism, artifacts or distortions in the picture, text rendering, audio loudness or audio bitrate, and how faithfully the *visual* content follows the prompt. The two videos may differ greatly in visual quality while being

equally well synchronized; in that case they are equal on this criterion. Judge audio-visual synchronization and nothing else.

[[PROMPT]]: {prompt}

Output: Your score should reflect the quality of the AI assistant's response(s) with respect to the specific criteria above, ignoring other aspects of the answer (such as overall quality), and should agree with the score provided by a reasonable human evaluator. The order of the responses is random, and you must avoid letting the order bias your answer. Be as objective as possible in your evaluation. Begin your evaluation by carefully analyzing the evaluation criteria and the response. After providing your explanation, output your final verdict by strictly following this format: ``[[A]]'' if response A is better, ``[[B]]'' if response B is better, ``[[C]]'' means you cannot decide which one is better (or they are equal). However, please try to avoid giving a ``tie'' preference and be as decisive as possible.

Return JSON only, with key ``reasoning'' (your analysis, citing the concrete sound events and the timestamps at which they occur relative to the visuals) and key ``verdict'' set to ``A'', ``B'', or ``tie'' (``tie'' corresponds to ``[[C]]'').

D.3 Human preference study instructions

Human preference study instructions

The two videos are generated from the same text prompt, each with its own generated audio track. Your task is to decide which video has **better audio-visual synchronization**.

Criteria (judge synchronization only):

- Temporal alignment: does each sound event start, peak and stop at the same moment as the on-screen action that physically produces it (impacts, footsteps, speech and lip movement, collisions, mechanical clicks, door closes, water splashes, instrument strikes, etc.)? Penalize sounds that lead or lag the visual event, drift out of phase, or are smeared across the wrong frames.
- Semantic correspondence: do the sounds match the visible actions, objects and materials? Penalize sounds with no visible source, visibly sound-producing actions that stay silent, and sounds of the wrong type, material, or number of events.
- Realization of the prompt: the prompt (shown above the videos) states which sounds should occur and when; prefer the video whose audio produces these events at the right moments.

Ignore: Image clarity, resolution, bitrate, compression artifacts, colour, exposure, lighting, aesthetics, motion smoothness, visual artifacts and distortions, text rendering, loudness, and how faithfully the *visual* content follows the prompt. Two videos may differ greatly in visual quality yet be equally well synchronized; in that case they are equal under this criterion.

Procedure:

- Wear headphones and make sure the volume is on; the task cannot be completed without sound.
- The two videos never play simultaneously: use ``Play in sequence'' to hear A and then B, and replay either video as often as needed.
- Choose ``A is better'', ``B is better'', or ``Tie''. Be as decisive as possible and select Tie only when the two cannot be distinguished.

- You may revisit and revise earlier answers; a revision overwrites the previous answer.

E GENERATOR POST-TRAINING

We post-train LTX-2.3 with DiffusionNFT (Zheng et al., 2026), an online reinforcement learning method that optimizes a diffusion or flow model directly on the forward (noising) process. It requires neither likelihood estimation along the reverse trajectory nor a specific sampler. This subsection summarizes the objective, the policy and sampler, the prompt pool, and the reward arms; Table 6 lists the recipe.

DiffusionNFT objective. For a prompt c, the sampling policy generates a group of k rollouts $\{x_0^{(i)}\}_{i=1}^{k}$ that are scored by the reward model R. Group-relative advantages are converted into optimality probabilities

$$A_i = \frac{R_i - \text{mean}_{j \in \mathcal{G}(c)} R_j}{\text{std}(R)}, \qquad r_i = \tfrac{1}{2} + \tfrac{1}{2}\,\text{clip}\big(A_i / A_{\max}, -1, 1\big) \in [0,1], \tag{5}$$

where the mean is taken within the group of prompt c, the standard deviation over all rewards in the sampling round, and $A_{\max} = 5$. Each rollout is re-noised along the forward process, $x_t = (1-\sigma)x_0 + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, and the velocity prediction is converted into a clean-sample estimate $\hat{x}_\theta = x_t - \sigma v_\theta(x_t, \sigma, c)$. DiffusionNFT defines implicit positive and negative policies around a frozen reference $\hat{x}_{\text{old}}$:

$$x_\theta^+ = (1-\beta)\,\hat{x}_{\text{old}} + \beta\,\hat{x}_\theta, \qquad x_\theta^- = (1+\beta)\,\hat{x}_{\text{old}} - \beta\,\hat{x}_\theta, \tag{6}$$

and minimizes

$$\mathcal{L}_{\text{NFT}}(\theta) = \mathbb{E}_{c,\,x_0,\,\sigma,\,\epsilon}\left[r\,\frac{\|x_\theta^+ - x_0\|_2^2}{w^+} + (1-r)\,\frac{\|x_\theta^- - x_0\|_2^2}{w^-} \right], \tag{7}$$

where $w^\pm$ is the per-sample mean absolute error of the corresponding prediction, which acts as an adaptive weight. High-reward rollouts pull the positive policy toward themselves. Low-reward rollouts are fitted by the negative policy, which, by the mirror symmetry of Eq. equation 6, pushes $\hat{x}_\theta$ away from them. At the optimum, the learned policy moves from v_{old} along the difference between the positive and negative data distributions. This internalizes a guidance direction, so CFG is unnecessary at inference. We use $\beta = 1$, which gives $x_\theta^+ = \hat{x}_\theta$ and $x_\theta^- = 2\hat{x}_{\text{old}} - \hat{x}_\theta$, and weight the video and audio latents equally. The reference policy is updated once per sampling round by an exponential moving average, $\theta_{\text{old}} \leftarrow \eta_t\,\theta_{\text{old}} + (1-\eta_t)\,\theta$ with $\eta_t = \min(0.001\,t,\, 0.5)$. No KL penalty is used.

Policy and sampler. We start from the released distilled checkpoint and its 8-step Euler schedule $\sigma \in \{1, .994, .988, .981, .975, .909, .725, .422, 0\}$, with classifier-free guidance and STG disabled for both video and audio. The transformer, Gemma-3 text encoder, video VAE, audio VAE and vocoder are frozen. A trainable LoRA (r=32, α=64, Gaussian init, no dropout) is attached to `to_q`/`to_k`/`to_v`/`to_out.0` of all six attention types in each of the 48 blocks: video and audio self- and text-attention, and both audio–video cross-attentions (214M parameters). Both the video and the audio stream are therefore updated. Rollouts are 512×768, 121 frames at 24 fps (5.04 s) with 48 kHz audio. The loss in Eq. equation 7 is evaluated only at the four highest-noise levels $\sigma \in \{1, .994, .988, .981\}$, with fresh noise for every level.

Grouping and optimization. Each sampling round draws 2 prompts with k=32 rollouts each (16 GPUs × 4 rollouts, distinct seeds), and advantages are clipped to $[-5, 5]$. Each round is consumed by 2 optimizer steps (per-GPU micro-batch 2, i.e., 32 rollouts per step), and each rollout is used once. 2,000 steps therefore correspond to 64,000 rollouts, or two passes over the 1,000 training prompts. We use AdamW (β_1=0.9, β_2=0.999, weight decay 10^{-4}, ϵ=10^{-8}) with a constant learning rate of 1×10^{-5} and no warmup, gradient clipping at 1.0, bf16 autocast with fp32 master weights, activation checkpointing, and FSDP full sharding on 2×8 H20 GPUs. A run takes 100–162 wall-clock hours (180–291 s per step, ≈1.6–2.6k GPU-hours per arm). We evaluate the final checkpoint (step 2,000) of every arm without checkpoint selection.

Table 6: **DiffusionNFT post-training recipe.**

Generator	LTX-2.3 22B, distilled-1.1, 8-step Euler, CFG/STG off
Resolution / length	512×768, 121 frames @ 24 fps, 48 kHz audio
Policy	LoRA r=32, α=64, all attn. q/k/v/o (video+audio), 214M
Group size k / prompts per round	32 / 2 (64 rollouts)
Micro-batch / rollouts per step	2 per GPU / 32
Advantage	group mean, batch std, clip ±5, $A_{\max}$=5
Trained σ	4 highest of 8 ($\sigma\geq0.98$)
NFT β / reference EMA η_t	1.0 / $\min(0.001t, 0.5)$ per round
Optimizer	AdamW, lr 10^{-5} const., wd 10^{-4}, clip 1.0
Precision / parallelism	bf16, FSDP full shard, activation ckpt.
Steps / rollouts	2,000 / 64,000 (2 passes)
Train prompts	1,000 (four on-screen categories)
Monitoring prompts	120 (disjoint, same construction)
Hardware	16×H20

Prompt pool. Starting from $80k$ captions, a rule-based filter keeps only on-screen, causally visible sound. It rejects background music, narration, off-screen or ambient audio, stylized or animated content, applause, montage cuts and very dark scenes; it also requires the source to be visible and caps quoted speech so that it fits in 5 s. Captions are assigned to four categories: multi-speaker dialogue, difficult speech (singing, rapping, tongue-twisters), on-screen percussion, and sudden impacts (glass, gunshots, hammer blows). Using human 0–2 synchronization ratings, we drop prompts on which LTX-2.3 already scores above 1.5. We prioritize prompts with evidence of failure: LTX-2.3 rated at most 1.5 or judged unrelated, or another generator rated at most 1.0. Such evidence exists for 99% of the training prompts. With an equal per-category quota, 1,120 prompts are selected and every ninth prompt in the ranking is held out. This gives 1,000 training prompts (255/252/246/247 per category) and 120 monitoring prompts (25/28/34/33). Prompts from SyncReward-Bench and from the reward-model validation and calibration splits are excluded. The monitoring prompts are used only to track each arm's own reward during training.