AVC-REWARD: EVENT-CENTRIC AUDIO-VIDEO CONSISTENCY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As joint audio-video generation models become more capable, they can generate audio-visual scenes exhibiting three forms of complexity: **multi-event composition**, **multi-layer audio mixture**, and **asymmetric cross-modal relations**. However, current evaluations largely reuse metrics from earlier simple and unidirectional tasks. They typically evaluate an audio-video clip as a whole and rely on simple predefined relations, making them unsuitable for these increasingly complex scenes. We propose an **event-centric** evaluation framework that disentangles audio and visual events at the event level, analyzes each event independently, cross-validates relations across modalities, and aggregates the judgments into a **coarse-to-fine, three-level consistency metric**. The metric assesses global narrative harmony, event-level temporal synchronization, and fine-grained attribute alignment, providing a general formulation for audio-video consistency evaluation across diverse generation settings and cross-modal relations. We instantiate the metric as **AVC-Reward**, a trainable audio-video reward model. To evaluate the agreement between automatic judgments and human perception across generation paradigms and relation complexities, we construct AVC-Benchmark, which is primarily built from diverse positive and negative samples generated by various types of audio-video generation models. AVC-Reward achieves state-of-the-art alignment with human judgment among automatic methods, provides interpretable rationales that localize failures, and improves the audio-video consistency of a joint generation model when used as an online reinforcement-learning signal. Project page: `https://anonymous.4open.science/r/AVCReward-A743/`

1 INTRODUCTION

Audio-video generation has evolved from unidirectional driving toward joint generation with increasingly strong generative capabilities. Earlier audio-driven video generation (e.g., Wav2Lip (Prajwal et al., 2020)) and video-to-audio generation (e.g., HunyuanVideo-Foley (Shan et al., 2025)) were centered on more constrained consistency relations: the former mainly focused on lip synchronization (Chung & Zisserman, 2016), while the latter mainly focused on simple temporal alignment (Iashin et al., 2024). In contrast, joint generation removes the constraint of a fixed modality, with recent models such as Veo 3 (Google DeepMind, 2025), MiniMax H3 (MiniMax, 2026) and Seedance 2.0 (Seedance et al., 2026) extending this paradigm to richer and more complex scenes.

These richer audio-visual scenes exhibit three forms of complexity: **multi-event composition**, **multi-layer audio mixture**, and **asymmetric cross-modal relations**. First, a clip can contain multiple events that together form a complete scene, with several events potentially occurring at the same time. Second, its audio track can mix physical scene sounds with post-produced audio such as voice-over and background music, which instead support the video's narrative or atmosphere. Third, audio and visual events need not follow a simple one-to-one relation: valid audio events such as off-screen sounds may lack direct visible visual evidence.

However, **evaluation methodologies have not kept pace** with this evolution: the richer audio-visual content produced by joint models has made consistency evaluation more demanding. Current evaluations largely reuse metrics from earlier simple and unidirectional tasks, where one modality was fixed and only a limited relation such as lip–speech or sound–action timing needed to be checked; accordingly, they often evaluate an audio-video clip as a whole and rely on simple predefined re-

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.

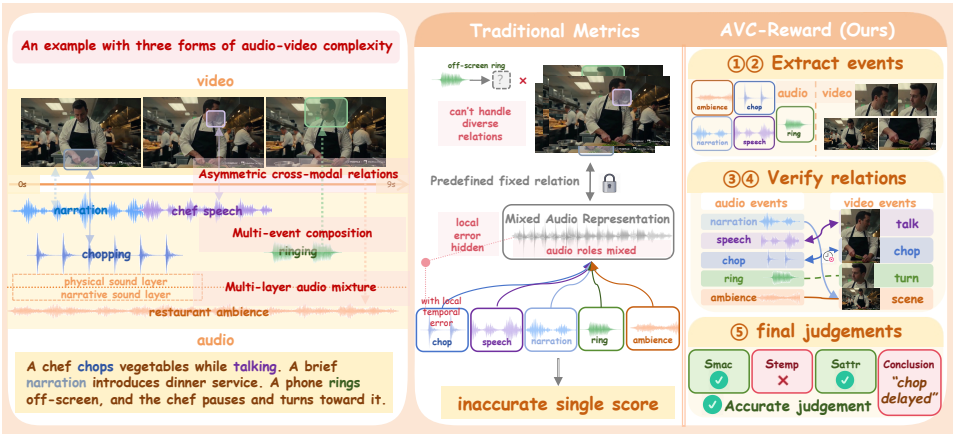

Figure 1: Overview of the proposed event-centric audio-video consistency evaluation framework.

lations such as semantic similarity or temporal synchronization. **Multi-event composition** makes such clip-level evaluation unreliable because different event relations are entangled; for example, ImageBind-based metrics (Girdhar et al., 2023; Mao et al., 2024; Liu et al., 2025) encode the audio and video within a clip or temporal window into a shared embedding space and measure their overall similarity, which can remain high when dominant semantics match despite a local event inconsistency. **Multi-layer audio mixture** further complicates evaluation: metrics such as Synchformer (Iashin et al., 2024) predict an audio–video temporal offset from the mixed soundtrack without separating its layers. When physical sounds are mixed with voice-over or background music, sounds that should align with visible actions are evaluated together with audio that does not require such temporal correspondence, which can lead to an inaccurate consistency score. **Complex cross-modal relations** further challenge predefined matching rules; for example, AV-Align (Yariv et al., 2024) matches audio onsets with visual-motion peaks and may therefore penalize a valid off-screen sound that lacks direct visual evidence.

Therefore, we need a more general audio-video consistency evaluator that can handle diverse event relations across different generation paradigms, particularly in modern joint generation. Moreover, we believe that such an evaluation model can serve as a training signal to directly improve generation models, closing the loop of **"generate–evaluate–optimize."**

To meet this need, we propose an **event-centric** evaluation perspective, in which the audio and video of each generated clip are first disentangled into individual events and analyzed independently before the final judgments. Under this perspective, we first separately extract and analyze audio and visual events, then cross-validate their relations across modalities, and finally aggregate the resulting judgments from coarse to fine. We formulate this final consistency assessment as **a coarse-to-fine, three-level consistency metric**. **Global narrative harmony** assesses whether the audio and video form a coherent scene and narrative. **Event-level temporal synchronization** establishes event relations, checks event presence and temporal alignment, and marks audio events without visual evidence. **Fine-grained attribute alignment** verifies the attributes of matched events and assesses whether marked audio is plausibly an off-screen sound, voice-over, or background music. By disentangling individual events, distinguishing different audio sources, and explicitly validating their cross-modal relations, this metric can separately assess overlapping events, handle audio layers with different roles, and accommodate valid asymmetric relations, as illustrated in Fig. 1.

Based on this event-level perspective, we implement the above three-level metric as a trainable audio-video reward model, named **AVC-Reward**, initialized from Qwen3-Omni-Thinking (Xu et al., 2025). AVC-Reward implements the metric through a five-stage structured reasoning chain consisting of audio(/video)-only analysis, cross-modal verification, and overall judgment. Its training data includes audio-video samples from different generation paradigms and models, with annotations from multiple expert MLLMs and human annotators, while its two-stage training procedure uses supervised fine-tuning (SFT) to cold-start the structured evaluation process and group relative policy optimization (GRPO) (Shao et al., 2024) to align the scoring with human preferences. Mean-

Table 1: Comparison of audio-video consistency metrics for challenges exposed by modern JAVG.

Method	Scoring Unit	Audio layers	Relation Evaluated
SyncNet (Chung & Zisserman, 2016)	Mouth–speech window	Mixed	Lip–speech synchrony
AV-Align (Yariv et al., 2024)	Audio / motion peaks	Mixed	Motion–sound timing
DeSync (Iashin et al., 2024)	AV temporal window	Mixed	Temporal offset
ImageBind (Girdhar et al., 2023)	Clip	Mixed	Global semantic similarity
AVHScore (Mao et al., 2024)	Frame $\times$ whole audio	Mixed	Frame–audio similarity
JavisScore (Liu et al., 2025)	Temporal window	Mixed	Window-level AV similarity
PEAVS (Goncalves et al., 2024)	Clip	Mixed	Predefined AV failures
AVBench (Yang et al., 2026a)	Clip	Mixed	Predefined AV failures
AGAV-Rater (Cao et al., 2025)	Clip	Mixed	Generative AV consistency
Beyond Time Shifts (Qian et al., 2026)	Clip	Mixed	Generative AV consistency
Ours	**Event**	**Decoupled**	**Event-dependent relation**

while, to measure the agreement between audio-video consistency metrics and human perception across generation paradigms and relation complexities, we construct an audio-video consistency benchmark, **AVC-Benchmark**. To our knowledge, it is the first audio-video consistency benchmark primarily built from diverse samples generated by various types of audio-video generation models.

We evaluate the reward model from two complementary perspectives. For **evaluation effectiveness**, it achieves **state-of-the-art alignment with human judgment** on AVC-Benchmark, and unlike a black-box scalar, it provides **interpretable, structured natural-language rationales** that localize where a generation fails (e.g., a temporally misaligned event or a hallucinated sound). For **usability as a reward**, applying it as a training signal for the online reinforcement-learning fine-tuning (Zhang et al., 2026a) of a joint generation model LTX-2 (HaCohen et al., 2026) measurably improves the audio-video consistency of generated outputs, demonstrating that it closes the optimization loop. Our contributions are summarized as follows:

- We formalize **audio-video consistency** across generation paradigms with an **event-centric, three-level metric** designed to handle complex generated scenes.
- We instantiate this metric as a trainable **audio-video reward model** built on the Qwen3-Omni-Thinking (Xu et al., 2025) model with rigorous event-level verification.
- We construct **AVC-Benchmark** for reproducible human-alignment evaluation, and validate the reward model both as an interpretable automatic evaluator and as an effective training signal that **improves the audio-video consistency of a joint generation model**.

2 Related Work

Joint Audio-Video Generation. Audio-video generation first developed through audio-driven video generation (Prajwal et al., 2020; Chen et al., 2025; Cui et al., 2025; Jiang et al., 2025) and video-to-audio generation (Zhang et al., 2026b; Cheng et al., 2025; Shan et al., 2025), where one modality is fixed. Joint audio-video generation (JAVG) removes this constraint: early works either learned a joint generative process (Ruan et al., 2023) or coordinated pretrained unimodal generators (Xing et al., 2024). Native JAVG then advanced toward open-ended generation, finer synchronization, and richer audio content, as represented by JavisDiT (Liu et al., 2025), Ovi (Low et al., 2025), UniVerse-1 (Wang et al., 2025), and LTX-2 (HaCohen et al., 2026). JAVG has further reached industrial-scale deployment through systems such as Veo 3 (Google DeepMind, 2025), Seedance 2.0 (Seedance et al., 2026), and MiniMax H3 (MiniMax, 2026), enabling the generation of more complex audio-visual scenes with multi-event composition, multi-layer audio mixtures, and complex cross-modal relations, which makes audio-video consistency evaluation more demanding.

Audio-Video Consistency Metrics. Existing audio-video consistency metrics and their limitations under modern joint audio-video generation are summarized in Table 1. Early metrics target specific consistency relations: SyncNet measures lip–speech synchrony, AV-Align matches motion and audio peaks, and DeSync predicts temporal offsets. ImageBind, AVHScore, and JavisScore measure cross-modal similarity at clip, frame–audio, or temporal-window granularity. PEAVS and AVBench learn human preferences from manually created errors. AGAV-Rater and Beyond Time Shifts further learn from real V2A outputs. Recent analysis further shows that existing metrics fo-

cus on different aspects of consistency and exhibit low cross-metric agreement (Sharma & Tapadiya, 2026). Together, these methods remain tied to specific tasks or relations, making them insufficient for general audio-video consistency evaluation under the complexity of modern joint generation.

3 METHOD

3.1 EVALUATION OVERVIEW

The complexities exposed by modern joint generation introduce three requirements for evaluation. Multi-event composition calls for **event disentanglement**, so that individual and overlapping events can be examined separately. Multi-layer audio mixture requires the evaluator to **distinguish the roles of different audio events**, as different audio sources can have different relations to the video. Asymmetric and diverse cross-modal relations further require **establishing and verifying event-dependent cross-modal relations** for individual events.

Meeting these requirements requires an evaluator capable of jointly understanding audio-visual content and reasoning over different events and their relations. MLLMs with audio-video understanding capabilities can jointly perform these tasks within a unified reasoning process, making them a suitable foundation for audio-video consistency evaluation.

Moreover, for more general and accurate evaluation, we argue that the judgments of the MLLM should be organized into an **event-centric** metric with **three coarse-to-fine levels**. **Global narrative harmony** first checks whether the audio and video are coherent across the full clip. **Event-level temporal synchronization** and **fine-grained attribute alignment** then move to the event level. They respectively assess the temporal alignment of each event and the consistency of audio-visual attributes within that event. Through event-level evaluation and detailed, independent verification of each event's cross-modal relations, the metric addresses the higher evaluation demands above.

With this evaluation design in place, Section 3.2 defines the event-centric protocol and three-level metric, Section 3.3 constructs the corresponding training supervision, and Section 3.4 learns AVC-Reward through SFT and GRPO.

3.2 EVENT-CENTRIC EVALUATION PROTOCOL

Following Section 3.1, we define three-level consistency scores. **Global narrative harmony** (S_{mac}) measures agreement in narrative content, entities, environment, and affective tone. **Event-level temporal synchronization** (S_{temp}) measures event presence and the alignment of onset, offset, and duration under the appropriate cross-modal relations. **Fine-grained attribute alignment** (S_{attr}) measures attribute compatibility within verified relations, including sound–material agreement, lip-shape agreement with speech, and plausible responses to off-screen sounds. All three scores lie in $[0, 1]$, with higher values indicating greater consistency.

To compute these scores, we guide the MLLM through a five-stage reasoning protocol that specifies the evidence and judgments required at each stage. Given an observed audio-video pair $\mathcal{P} = (\mathcal{A}, \mathcal{V})$, the first two stages direct the model to analyze each modality independently:

$$\mathcal{H}^A = (\mathcal{G}^A, \mathcal{E}^A), \qquad \mathcal{H}^V = (\mathcal{G}^V, \mathcal{E}^V), \tag{1}$$

where $\mathcal{H}$ denotes an analysis record, $\mathcal{G}$ denotes global scene and content information, and $\mathcal{E}$ denotes the event set. Each event records its temporal span and attributes; audio events additionally record their source role. The next two stages verify the extracted events against the other modality:

$$\mathcal{R}^{A\to V} = F_{A\to V}(\mathcal{H}^A, \mathcal{V}), \qquad \mathcal{R}^{V\to A} = F_{V\to A}(\mathcal{H}^V, \mathcal{A}). \tag{2}$$

Here, $F_{A\to V}$ checks visual evidence for audio events, and $F_{V\to A}$ checks expected sounds for visual events. The required evidence depends on each audio event's role. The output sets $\mathcal{R}^{A\to V}$ and $\mathcal{R}^{V\to A}$ contain verification records $r_k = (e_i^A, e_j^V, \rho_k)$. Here, e_i^A and e_j^V denote related audio and visual events, and ρ_k specifies their relation and verification evidence. Either event may be absent when evidence is missing or the relation is asymmetric.

Let $\mathcal{R} = \mathcal{R}^{A\to V} \cup \mathcal{R}^{V\to A}$. The final stage scores global consistency from the two global descriptions and event consistency from the verified evidence:

$$S_{\mathrm{mac}} = \Phi_{\mathrm{mac}}(\mathcal{G}^A, \mathcal{G}^V), \qquad S_{\mathrm{temp}} = \Phi_{\mathrm{temp}}(\mathcal{R}), \qquad S_{\mathrm{attr}} = \Phi_{\mathrm{attr}}(\mathcal{R}). \tag{3}$$

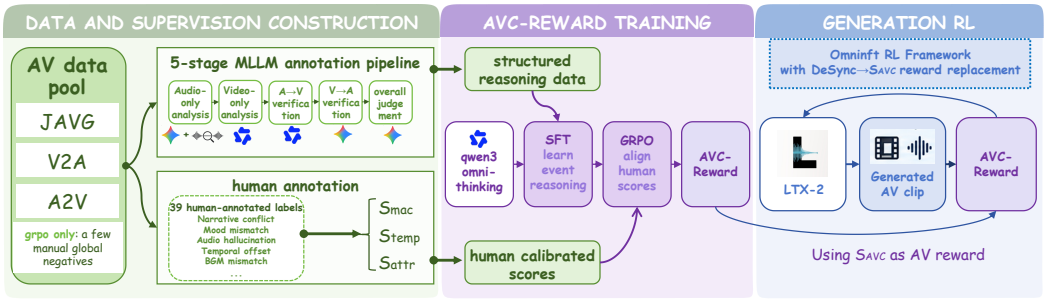

Figure 2: Overview of data construction, reward-model training, and generation-side optimization.

The Φ functions represent the MLLM's judgments under the three scoring criteria. Scoring proceeds from global semantics to event presence and timing, followed by attribute checks on relations that pass temporal verification. The three scores remain separate. When an overall score is required, we use their arithmetic mean as $S_{\text{AVC}} = \frac{1}{3}(S_{\text{mac}} + S_{\text{temp}} + S_{\text{attr}})$.

We use S_{AVC} as the default overall score in our experiments. Full event definitions and scoring criteria are given in App. A.

3.3 Training Data Construction and Annotation

Our training data includes outputs from existing joint audio-video, audio-driven video, and video-to-audio generation models. We further add a small number of manually edited global-narrative negatives in GRPO to balance the distribution of failure cases. The overall pipeline is illustrated in Fig. 2. We provide two complementary forms of supervision: structured reasoning data teaches the model to follow the event-centric evaluation protocol for SFT, while human-calibrated data aligns its scores with human judgment for GRPO. The complete data sources and calibration details are provided in App. B.

Structured reasoning data. Following the five-stage protocol in Section 3.2, the annotation organizes modality-specific event analysis, bidirectional cross-modal verification, and the final coarse-to-fine judgment into a structured reasoning process. We construct the annotations using Gemini 2.5 Pro (Comanici et al., 2025) and Qwen3-VL-Max (Bai et al., 2025) as complementary annotators, with a silence detector providing auxiliary temporal cues. Each stage produces one structured record. The five structured records are concatenated and rewritten by Qwen3-235B (Yang et al., 2025a) into a single SFT target, $\langle \texttt{think} \rangle \mathcal{C} \langle \texttt{/think} \rangle \{S_{\text{mac}}, S_{\text{temp}}, S_{\text{attr}}, \kappa\}$, where $\mathcal{C}$ preserves the five-stage reasoning order. After filtering, approximately 8k samples are retained for SFT. These samples provide explicit supervision for learning event-specific evidence and event-dependent cross-modal relations under the three-level metric. Complete reward-model prompts and input–output examples are provided in App. E.

Human annotation and score calibration. The structured annotation pipeline provides scalable reasoning supervision, but its scores are limited by the perceptual capability of the annotation models and can accumulate errors across stages. We therefore introduce human annotations to calibrate the scores toward human judgment.

We design a 39-label taxonomy over the three consistency dimensions, covering normal states and common failure types. Annotators assign applicable labels to each sample, grounding their judgments in explicit consistency evidence. We further construct a multi-annotator subset with both label annotations and independently assigned dimension scores. Based on this subset, Qwen3-235B (Yang et al., 2025a) maps each human-verified label record to $(S_{\text{mac}}, S_{\text{temp}}, S_{\text{attr}})$ and a diagnostic conclusion κ. This calibration produces approximately 10k human-calibrated samples for GRPO.

3.4 AVC-Reward Training

We initialize AVC-Reward from Qwen3-Omni-Thinking (Xu et al., 2025) and train it in two stages.

SFT cold start. We fully fine-tune the model on the approximately 8k structured reasoning samples from Section 3.3, using the complete five-stage reasoning and assessment as the training target.

GRPO for human alignment. After SFT cold start, we apply GRPO (Shao et al., 2024) on a 10k rollout set constructed from the human-calibrated data. The reward jointly preserves the required evaluation structure and aligns the three consistency scores with human judgment. Specifically, a structure reward r_{struct} enforces the five-stage reasoning and assessment format. A score distillation reward $r_{\text{distill}} = \frac{1}{3}\sum_{i=0}^{2} \exp[-w_i(\text{pred}_i - \text{ref}_i)^2]$ measures agreement with the human-calibrated scores, and the total reward is $r_{\text{total}} = (2r_{\text{struct}} + r_{\text{distill}})/3$, where $i = 0, 1, 2$ corresponds to $(S_{\text{mac}}, S_{\text{temp}}, S_{\text{attr}})$.

4 BENCHMARK

Motivation. We construct **AVC-Benchmark** to evaluate how well automatic audio-video consistency metrics align with human perception in realistic generation scenarios.

Benchmark composition and evaluation target. AVC-Benchmark is the first benchmark, to our knowledge, that primarily uses various types of model-generated audio-video data to evaluate the alignment between automatic consistency metrics and human perception. It contains samples from A2V, V2A, and JAVG models, together with a small set of controlled negatives added to balance the failure-case distribution and broaden the coverage of audio-video inconsistencies. Each sample is independently annotated by three annotators, providing human ground-truth scores on the three consistency dimensions defined in Section 3.2. We evaluate each metric against these human scores using pairwise accuracy under different score-difference thresholds, Spearman's rank correlation coefficient (SRCC), and Pearson's linear correlation coefficient (PLCC).

Statistics. Detailed benchmark statistics are provided in App. D.2. In general, most samples retain high global scene and narrative consistency, providing a meaningful basis for evaluating finer temporal and attribute-level relations; accordingly, the latter two dimensions span much broader score ranges. The annotated failures further cover global narrative, affective, and environmental conflicts; temporal errors across physical events, speech, and music; and fine-grained mismatches in material, morphology, and causal response.

5 EXPERIMENTS

We evaluate AVC-Reward as both an automatic evaluator and a training reward for joint generation. Section 5.1 evaluates its agreement with human judgments; Section 5.2 studies generation-side optimization; Section 5.3 reports human evaluation; Section 5.4 presents case studies; and Section 5.5 examines the main design choices.

5.1 REWARD MODEL EVALUATION

On AVC-Benchmark, we compare AVC-Reward with the audio-video metrics reviewed in Section 2 and with zero-shot MLLM judges, and use inter-annotator agreement under each measure as the human upper bound. The baselines include timing-based metrics (AV-Align (Yariv et al., 2024) and DeSync (Iashin et al., 2024)), embedding-based metrics (ImageBind (Girdhar et al., 2023), JavisScore (Liu et al., 2025), and AVHScore (Mao et al., 2024)), learned audio-video evaluators (PEAVS (Goncalves et al., 2024), AVSyncEvaluator (Qian et al., 2026), AV-Consistency Evaluator (Yang et al., 2026a), and AGAV-Rater (Cao et al., 2025)), and zero-shot MLLM judges (Gemini-3.1-Pro (Google DeepMind, 2026) and Qwen3-Omni (Xu et al., 2025)). For AVC-Reward, we use S_{AVC} as its overall prediction, while the human reference follows the overall score defined in App. D.1. We report Pairwise, Δ-Pairwise, PLCC (5PL), and SRCC. Since audio-video consistency judgments can be subjective for clips with very similar human scores, Δ-Pairwise computes pairwise accuracy only on pairs whose human-score difference is more than Δ. We use $\Delta = 0.2$ as the main setting and report other thresholds in Fig. 3.

Table 2: Main results on AVC-Benchmark (706 samples). Δ-Pairwise (Δ0.2) and PLCC are the primary metrics. **Best** automatic result in bold; *Human Upper Bound* is the inter-annotator ceiling.

Method	Pairwise ↑	Δ-Pairwise ↑	PLCC ↑	SRCC ↑
Timing-based synchronization				
AV-Align (Yariv et al., 2024)	0.577	0.674	0.301	0.215
DeSync (Iashin et al., 2024)	0.521	0.580	0.280	0.061
Embedding similarity				
ImageBind (Girdhar et al., 2023)	0.556	0.625	0.219	0.154
JavisScore (Liu et al., 2025)	0.554	0.611	0.180	0.147
AVHScore (Mao et al., 2024)	0.540	0.595	0.171	0.110
Learned cross-modal consistency evaluators				
PEAVS (Goncalves et al., 2024)	0.491	0.505	0.075	-0.026
AVSyncEvaluator (Qian et al., 2026)	0.502	0.554	0.108	0.007
AV-Consistency Evaluator (Yang et al., 2026a)	0.562	0.664	0.286	0.173
AGAV-Rater (Cao et al., 2025)	0.580	0.646	0.261	0.226
Zero-shot general MLLM judges				
Gemini-3.1-Pro (Google DeepMind, 2026)	0.542	0.616	0.246	0.118
Qwen3-Omni-Instruct (Xu et al., 2025)	0.548	0.589	0.257	0.136
Qwen3-Omni-Thinking (Xu et al., 2025) (base)	0.555	0.605	0.347	0.154
AVC-Reward (Ours)	**0.586**	**0.728**	**0.494**	**0.244**
Human Upper Bound	*0.756*	*0.920*	*0.781*	*0.497*

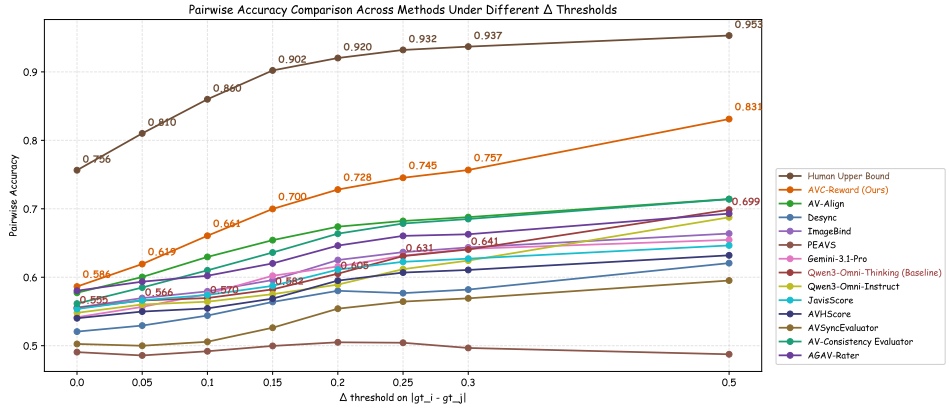

Figure 3: Pairwise accuracy under different margin thresholds Δ. Larger Δ keeps pairs with clearer human-score differences.

Overall, AVC-Reward agrees with human judgment better than all automatic baselines. Table 2 shows the best result on all four measures. Fig. 3 gives the same trend across thresholds: as Δ increases, the human-score differences become clearer and pairwise accuracy improves accordingly. This matches the expected behavior of a human-aligned evaluator: clearer human differences lead to more accurate pairwise judgments. These results support our AVC-Reward as **a more human-aligned measure of consistency in diverse generated audio-video scenes**.

5.2 Generation-side Optimization

Setup. We use OmniNFT (Zhang et al., 2026a), a strong recent online RL framework for LTX-2 (HaCohen et al., 2026), to test AVC-Reward in generation optimization. We keep the overall training framework unchanged and replace its original audio-video reward DeSync with AVC-Reward, using its overall score S_{AVC} as the optimization signal, with reward-specific learning-rate and weight adjustments for comparable optimization strength. Full training details are provided in App. C.3. For evaluation, we avoid JavisBench because part of its prompt set was used to collect training samples, and use AV-Phys-Bench (Cui et al., 2026), VABench-T2AV (Hua et al., 2026), and AVGen-Bench (Zhou et al., 2026) as prompt sets. We report the metric set used by OmniNFT on JavisBench and add some audio-video consistency metrics to show changes in this core aspect more clearly.

Results. AVC-Reward consistently improves external audio-video consistency metrics across all prompt sets: IB-AV by 20.65–28.35%, AV-Align by 18.78–59.47%, JavisScore by 20.47–28.74%,

Table 3: Generation results across three benchmarks. Gain rows report relative improvements over LTX-2. All other entries are absolute metric values. Best results within each benchmark are in bold.

(a) AV consistency

Model	AVC-Score↑	DeSync↓	IB-AV↑	AV-align↑	Javis↑	AVH↑	CAVP↑
AV-Phys-Bench							
LTX-2	0.756	0.376	0.234	0.158	0.189	0.226	0.812
OmniNFT (DeSync)	0.780	**0.073**	0.248	0.213	0.193	0.233	**0.819**
OmniNFT gain	+3.26%	**+80.60%**	+5.84%	+34.17%	+1.96%	+2.91%	**+0.88%**
Ours (AVC-Reward)	**0.819**	0.081	**0.288**	**0.246**	**0.228**	**0.276**	0.818
Ours gain	**+8.35%**	+78.44%	**+22.92%**	**+55.50%**	**+20.47%**	**+22.00%**	+0.71%
VABench							
LTX-2	0.764	0.421	0.216	0.185	0.177	0.207	0.805
OmniNFT (DeSync)	0.791	0.097	0.256	0.217	0.207	0.244	**0.815**
OmniNFT gain	+3.54%	+76.97%	+18.68%	+17.19%	+16.65%	+18.24%	**+1.31%**
Ours (AVC-Reward)	**0.850**	**0.063**	**0.277**	**0.295**	**0.223**	**0.262**	0.814
Ours gain	**+11.17%**	**+85.03%**	**+28.35%**	**+59.47%**	**+25.65%**	**+26.80%**	+1.15%
AVGenBench							
LTX-2	0.786	0.288	0.257	0.246	0.189	0.231	0.803
OmniNFT (DeSync)	0.808	**0.074**	0.295	0.225	0.217	0.262	**0.816**
OmniNFT gain	+2.76%	**+74.31%**	+14.86%	-8.37%	+14.92%	+13.20%	**+1.66%**
Ours (AVC-Reward)	**0.849**	0.089	**0.310**	**0.292**	**0.243**	**0.286**	0.814
Ours gain	**+8.02%**	+69.27%	**+20.65%**	**+18.78%**	**+28.74%**	**+23.98%**	+1.32%

(b) Text alignment and single-modality quality

Metric	AV-Phys-Bench			VABench			AVGenBench		
	LTX-2	OmniNFT	**Ours**	LTX-2	OmniNFT	**Ours**	LTX-2	OmniNFT	**Ours**
IB-TV↑	**0.324**	0.322	0.308	**0.248**	0.242	0.232	**0.158**	0.140	0.145
IB-TA↑	0.180	0.200	**0.207**	0.127	**0.144**	0.133	0.110	0.115	**0.119**
CLIP↑	**0.287**	0.282	0.272	**0.269**	0.264	0.255	**0.296**	0.292	0.289
CLAP↑	0.391	**0.402**	0.399	0.363	**0.380**	0.365	0.406	**0.437**	0.427
AQ↑	4.577	4.941	**4.970**	4.667	5.053	**5.087**	4.882	5.154	**5.265**
VQ↑	2.471	4.262	**4.298**	2.200	4.319	**4.421**	2.608	**4.157**	4.129

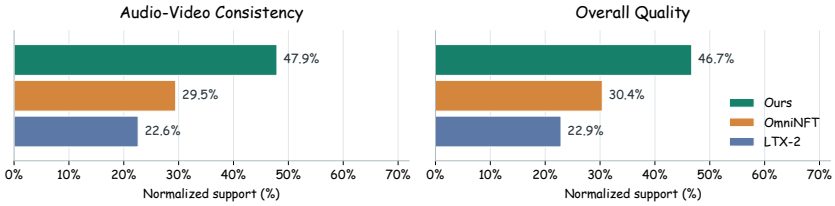

Figure 4: Overall human preference across the three benchmarks.

and AVHScore by 22.00–26.80%. Moreover, although OmniNFT directly optimizes DeSync while ours does not, our DeSync gains remain comparable. These gains demonstrate strong generalization to unseen audio-video consistency metrics. CAVP changes by less than 1.7% for both methods, indicating limited separation. As for other evaluate dimensions, because T-A, T-V, and A-V objectives are not perfectly aligned, stronger A-V optimization can slightly reduce text alignment. Within ImageBind, however, the AV gains exceed the observed TA/TV declines, while single-modality quality remains comparable at the OmniNFT level. These results show that **AVC-Reward is not only an accurate evaluator but can also serve as an effective reward signal for optimizing audio-video consistency in generation**.

5.3 Human Evaluation

We randomly sample 100 prompts from AVGenBench, AV-Phys-Bench, and VABench, and compare LTX-2, OmniNFT, and Ours with blinded randomized human judgments, as summarized in Fig. 4. We ask 10 annotators with undergraduate degrees or higher to evaluate the samples. Ours receives the highest preference for both audio-video consistency and overall quality. These preferences further **demonstrate the effectiveness of our generation-side optimization from a human perceptual perspective**. detailed rules and benchmark-wise results are reported in App. F.

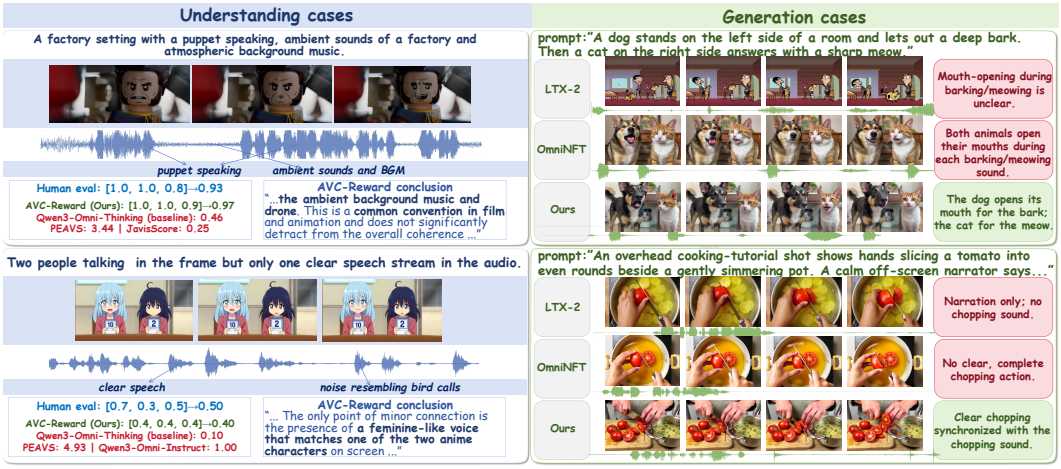

Figure 5: Representative cases. Left: AVC-Reward produces scores closer to human judgments and identifies specific consistency issues. Right: generation results of LTX-2, OmniNFT, and Ours.

Table 4: Ablations and aggregation analysis on AVC-Benchmark.

Training Strategy				**Score Aggregation**			
Variant	Δ-Pair↑	PLCC↑	SRCC↑	Variant	Δ-Pair↑	PLCC↑	SRCC↑
Base	0.605	0.347	0.154	S_{mac} only	0.688	0.467	**0.251**
GRPO only	0.630	0.330	0.161	S_{temp} only	0.717	0.479	0.235
SFT only	0.679	0.295	**0.244**	S_{attr} only	0.715	0.482	0.240
SFT+GRPO	**0.728**	**0.494**	**0.244**	**Mean** (S_{AVC})	**0.728**	**0.494**	0.244

5.4 Case Studies

Fig. 5 presents representative cases of AVC-Reward in evaluation and generation optimization. AVC-Reward handles multi-layer audio, diverse relations such as off-screen and ambient sounds, and multiple or concurrent events by distinguishing audio roles and evaluating event relations separately; as an RL reward, it also improves generation under these complexities. Moreover, bidirectional cross-modal verification enables finer checks of sound attributes and visible sound-producing actions, while the optimized generator produces more accurate sounds and clearer, more correct corresponding visual evidence. More cases are provided in App. E.

5.5 Training Ablation and Score-aggregation Analysis

We examine the training strategy and score aggregation in Table 4. Compared with GRPO only, SFT+GRPO achieves better human alignment, demonstrating the effectiveness of the proposed five-stage reasoning supervision, while the improvement from SFT only shows the benefit of human preference alignment through GRPO. For score aggregation, averaging all three dimensions achieves higher Δ-Pairwise than using single dimension alone, supporting their complementary use in S_{AVC}.

6 Conclusion

We study general audio-video consistency evaluation under the challenges of **multi-event composition**, **multi-layer audio mixture**, and **asymmetric cross-modal relations**. We propose an **event-centric coarse-to-fine three-level metric** and instantiate it as **AVC-Reward**, trained with structured SFT and human-aligned GRPO. We further introduce AVC-Benchmark to evaluate human agreement across diverse audio-video generation models. AVC-Reward achieves the strongest human alignment among evaluated methods and, when integrated into OmniNFT, effectively improves audio-video consistency. These results demonstrate the effectiveness of event-centric consistency evaluation as both a reliable metric and an optimization signal for joint audio-video generation.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat datasets, and support qualitative and thematic data analysis.

We have not used generative AI tools to develop theoretical models or conceptual frameworks or to interpret results, and tasks involving the formulation of mathematical claims, critical ingredients for proving mathematical claims, or assistance in writing proofs are not applicable to this work.

Additionally, we used generative AI tools to formulate questions for surveys or interviews, suggest experimental parameters, create or edit software code, summarize or analyze existing literature, brainstorm, source or search for information, edit the paper to improve readability, identify relevant literature, and propose titles or keywords for the paper. We have reviewed all AI-assisted work.

All AI-generated suggestions, outputs, and modifications were manually reviewed and verified by the authors. Any research design, analysis approach, experimental implementation, or manuscript revision adopted in this work was independently evaluated by the authors, with key decisions discussed and approved by multiple authors. The final manuscript was reviewed and approved by the author team.

We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work involves human annotation for benchmark construction. We assessed the annotation procedure and determined that formal IRB review was not required. The data used in this work comply with applicable licenses and terms of use and involve no material ethical, privacy, or copyright risks.

REPRODUCIBILITY STATEMENT

All results reported in this paper are reproducible. We provide complete details of the data sources, annotations, and processing pipeline in Appendix B, and the full training configurations, including frameworks, datasets, and hyperparameters, in Appendix C. The source code is available at the project page: `https://anonymous.4open.science/r/AVCReward-A743/`. The complete reward model weights, generative model weights, and benchmark datasets will be released after the ICLR anonymization period.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yuqin Cao, Xiongkuo Min, Yixuan Gao, Wei Sun, and Guangtao Zhai. Agav-rater: Adapting large multimodal model for ai-generated audio-visual quality assessment. *arXiv preprint arXiv:2501.18314*, 2025.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2403–2410, 2025.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28901–28911, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21086–21095, 2025.

Zijun Cui, Xiulong Liu, Hao Fang, Mingwei Xu, Jiageng Liu, Zexin Xu, Weiguo Pian, Shijian Deng, Feiyu Du, Chenming Ge, et al. Do joint audio-video generation models understand physics? *arXiv preprint arXiv:2605.07061*, 2026.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 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.

Google DeepMind. Veo 3. `https://deepmind.google/models/veo/`, 2025.

Google DeepMind. Gemini 3.1 pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026.

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.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23345–23355, 2026.

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.

Jianwen Jiang, Weihong Zeng, Zerong Zheng, Jiaqi Yang, Chao Liang, Wang Liao, Han Liang, Yuan Zhang, and Mingyuan Gao. Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. *arXiv preprint arXiv:2508.19209*, 2025.

Jiaqi Leng, Shuyuan Tu, Haidong Cao, Sicheng Xie, Daoguo Dong, Zuxuan Wu, and Yu-Gang Jiang. Preference score distillation: Leveraging 2d rewards to align text-to-3d generation with human preference. *arXiv preprint arXiv:2603.01594*, 2026.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, 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.

Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Tavgbench: Benchmarking text to audible-video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6607–6616, 2024.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities. MiniMax Research Blog, July 2026. URL `https://www.minimax.io/blog/minimax-h3`.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, et al. Omniweaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 484–492, 2020.

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 *European Conference on Computer Vision*, pp. 39–58. Springer, 2026.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10219–10228, 2023.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jai Kumar Sharma and Peeyush Tapadiya. What do audio-visual synchronization metrics actually measure? *arXiv preprint arXiv:2608.25157*, 2026.

Hamid R Sheikh. Image and video quality assessment research at live. *http://live. ece. utexas. edu/research/quality/*, 2003.

Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. In *ICCV*, 2023.

Shuyuan Tu, Qi Tian, Zihan Yang, Yue Wu, Xintong Han, Weijie Kong, Jiangfeng Xiong, Jian-Wei Zhang, Zhao Zhong, Liefeng Bo, et al. Baton: Explicit semantic blueprints for joint video-audio generation. *arXiv preprint arXiv:2605.25195*, 2026.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7161, 2024.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jialiang Yang, Bin Xia, Ruihang Chu, Dingdong Wang, Wanke Xia, Zhun Mou, Tianyang Zhong, Yiting Zhao, and Wenming Yang. Avbench: Human-aligned and automated evaluation benchmark for audio-video generative models. *arXiv preprint arXiv:2605.24652*, 2026a.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025b.

Zihan Yang, Shuyuan Tu, Licheng Zhang, Qi Dai, Yu-Gang Jiang, and Zuxuan Wu. Arcflow: Unleashing 2-step text-to-image generation via high-precision non-linear flow distillation. *arXiv preprint arXiv:2602.09014*, 2026b.

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.

Guohui Zhang, XiaoXiao Ma, Jie Huang, Hang Xu, Hu Yu, Siming Fu, Yuming Li, Zeyue Xue, Lin Song, Haoyang Huang, et al. Omninft: Modality-wise omni diffusion reinforcement for joint audio-video generation. *arXiv preprint arXiv:2605.12480*, 2026a.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134(1):46, 2026b.

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. *arXiv preprint arXiv:2604.08540*, 2026.

APPENDIX

In the Appendix, we present the following details:

- Complete formal metric definitions in Section A.
- Data construction and annotation details in Section B.
- Training details in Section C.
- AVC-Benchmark details in Section D.
- Additional evaluation and generation case studies in Section E.
- Human evaluation protocol and results in Section F.

A FORMAL METRIC DEFINITIONS

This section gives the complete mathematical form of the event-centric protocol in Section 3. The protocol evaluates an observed audio–video pair from coarse to fine: it first compares the two global descriptions, then verifies event existence and timing, and finally checks event attributes. The three scores remain separate throughout the evaluation; when one scalar is needed, we use their arithmetic mean.

A.1 CLIP AND MODALITY REPRESENTATIONS

Let an observed clip be

$$\mathcal{P} = (\mathcal{A}, \mathcal{V}), \qquad \mathcal{T} = [0, T], \tag{4}$$

where $\mathcal{A}$ and $\mathcal{V}$ denote the complete audio and video streams on the shared time interval $\mathcal{T}$. The protocol evaluates the observed pair directly, so the provenance of either stream is irrelevant. The first two stages convert the streams into

$$\mathcal{H}^A = (\mathcal{G}^A, \mathcal{E}^A), \qquad \mathcal{H}^V = (\mathcal{G}^V, \mathcal{E}^V), \tag{5}$$

where $\mathcal{G}^A$ and $\mathcal{G}^V$ are full-clip global scene and narrative descriptions, and $\mathcal{E}^A$ and $\mathcal{E}^V$ are event sets. Global descriptions may include the scene or environment, entities and their roles, and the affective or narrative tone. They summarize the whole clip across $\mathcal{T}$.

An audio event is represented as

$$e_i^A = \big(t_{i,s}^A, t_{i,e}^A, c_i^A, \lambda_i^A, \phi_i^A\big), \qquad 0 \le t_{i,s}^A \le t_{i,e}^A \le T, \tag{6}$$

where c_i^A is the event category, λ_i^A is its source role, and ϕ_i^A contains the observable acoustic attributes. The categories used by the annotation protocol are human speech, music or rhythm, and other physical sound events. The source role is

$$\lambda_i^A \in \{\text{physical}, \text{narrative}\}. \tag{7}$$

Physical events belong to the recorded world. Narrative events, such as background music or narration added for storytelling, are judged by their relation to the scene; visible source evidence is optional. For a physical event, the annotation also records whether the source is on screen or off screen when this distinction is observable.

A visual event is represented as

$$e_j^V = \big(t_{j,s}^V, t_{j,e}^V, c_j^V, \phi_j^V\big), \qquad 0 \le t_{j,s}^V \le t_{j,e}^V \le T, \tag{8}$$

where c_j^V identifies the audio-producing action (for example, speaking, playing an instrument, an object interaction, or an environmental action) and ϕ_j^V records its observable morphology, source location, and motion. Silent actions are retained as context; they create no required audio event.

A.2 FIVE-STAGE EVALUATION PROTOCOL

The five stages are formal interfaces between the evidence records and the final score. Stages 1 and 2 analyze one modality at a time:

$$\mathcal{H}^A = F_A(\mathcal{A}), \qquad \mathcal{H}^V = F_V(\mathcal{V}). \tag{9}$$

Stage 1 identifies the acoustic layers, global audio context, speaker or music information, and time-localized audio events. Stage 2 identifies the visual scene, people and actions, and the time-localized visual events that can produce sound.

Stages 3 and 4 verify the two directions separately:

$$\mathcal{R}^{A\to V} = F_{A\to V}(\mathcal{H}^A, \mathcal{V}), \qquad \mathcal{R}^{V\to A} = F_{V\to A}(\mathcal{H}^V, \mathcal{A}). \tag{10}$$

For every audio event, Stage 3 checks whether the video supplies the required evidence in the event's time interval. For a short event this is a direct check of the relevant action. For a continuous event it checks the minimal visible evidence throughout the interval. For every visual event, Stage 4 first decides whether the action should produce sound. Missing audio for a naturally silent action is reasonable silence; missing audio for a sound-producing action is an existence failure. Events that pass the existence judgment proceed to physical or attribute checks.

The output of either verification stage is a relation record

$$r_k = (e_i^A, e_j^V, \rho_k), \qquad e_i^A \in \mathcal{E}^A \cup \{\varnothing\}, \quad e_j^V \in \mathcal{E}^V \cup \{\varnothing\}, \tag{11}$$

where ρ_k specifies the relation being checked and stores the corresponding evidence. Either event can be absent because the relation is asymmetric or because the expected cross-modal evidence is missing. We write

$$\mathcal{R} = \mathcal{R}^{A\to V} \cup \mathcal{R}^{V\to A}. \tag{12}$$

Stage 5 uses the global descriptions and the verified relations to produce

$$\mathcal{S} = \big(S_{\text{mac}}, S_{\text{temp}}, S_{\text{attr}}\big) \in [0,1]^3, \tag{13}$$

with

$$S_{\text{mac}} = \Phi_{\text{mac}}(\mathcal{G}^A, \mathcal{G}^V), \qquad S_{\text{temp}} = \Phi_{\text{temp}}(\mathcal{R}), \qquad S_{\text{attr}} = \Phi_{\text{attr}}(\mathcal{R}). \tag{14}$$

A.3 THREE-LEVEL CONSISTENCY SCORES

Global narrative harmony. S_{mac} compares the audio and video global descriptions over the entire interval $\mathcal{T}$. The comparison covers narrative content and entity identity, scene or environment, and affective tone. A persistent conflict in any of these aspects lowers the score even when individual events happen to overlap in time. A voice-over or background music can be consistent without a visible source when its scene, narrative role, and affective relation are plausible.

Event-level temporal synchronization. For each relation r_k, let $x_k \in \{0,1\}$ denote the existence judgment and let $y_k \in [0,1]$ denote the onset–offset alignment judgment. The existence judgment is defined by the source role:

$$x_k = \begin{cases} 1, & \text{bidirectional evidence supports an on-screen physical event,} \\ 1, & \text{an off-screen physical event is plausible and its consequences are consistent,} \\ 1, & \text{a narrative event has an appropriate semantic or rhythmic role,} \\ 1, & \text{the visual action is naturally silent (reasonable silence),} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

For an on-screen physical event, bidirectional evidence means both directions are checked: a visible sound-producing action must have corresponding audio, and an audible on-screen source must have corresponding visual evidence. Each relation record contains an existence judgment $x_k \in \{0,1\}$ and a temporal-alignment judgment $y_k \in [0,1]$. The temporal judgment covers event onset, offset, duration, and the alignment of relevant states for rhythmic or continuous events. Narrative events use a meaningful rhythmic or temporal relation; reasonable-silence records receive full temporal evidence. The final LLM judge reads all temporal evidence in $\mathcal{R}$ and outputs

$$S_{\text{temp}} = \Phi_{\text{temp}}(\mathcal{R}) \in [0,1]. \tag{16}$$

The temporal score covers event existence and timing. Material, timbre, mouth shape, finger motion, and reaction plausibility belong to the attribute evidence.

Fine-grained attribute alignment. Attribute evaluation proceeds whenever temporal correspondence establishes a tentative audio–visual event pair within an acceptable tolerance. The verifier records applicable physical, morphological, and causal-response evidence for each relation. Physical evidence covers material or timbre, energy or loudness, and spatial or distance cues. Morphological evidence covers fine visual states such as mouth shape, finger motion, and the visible action that produces the sound. Causal-response evidence covers source localization, emotional response, and response magnitude for off-screen or narrative evidence. The final LLM judge reads all applicable attribute evidence in $\mathcal{R}$ and outputs

$$S_{\text{attr}} = \Phi_{\text{attr}}(\mathcal{R}) \in [0, 1]. \tag{17}$$

A.4 Overall AVC score

The final judgment follows the order $S_{\text{mac}} \to S_{\text{temp}} \to S_{\text{attr}}$. The three outputs remain separate diagnostic dimensions. When an overall score is required, we use the arithmetic mean

$$S_{\text{AVC}} = \frac{1}{3}\left(S_{\text{mac}} + S_{\text{temp}} + S_{\text{attr}}\right). \tag{18}$$

The overall score is the unweighted arithmetic mean of the three separately reported dimensions.

B Data Construction and Annotation

B.1 Training Data Sources and Statistics

We construct two data pools for AVC-Reward training, one for structured-reasoning SFT and the other for human calibration. Table 5 summarizes the sources and statistics of the training data used for SFT and GRPO.

SFT source pool. We first collect ground-truth audio-video clips from two categories: general scenes and human-centered scenes. We use Qwen3-Omni (Xu et al., 2025) to filter the clips for basic audio-video quality and clarity, salient event and temporal relations, and informative macro-semantic content, retaining 2.5k general clips and 1k human-centered clips. The four filtering questions are provided in the supplementary materials. Qwen3-Omni then generates audio and video captions for the retained clips, which are used to construct samples under different generation paradigms.

For the general subset, Ovi (Low et al., 2025) generates 2.5k JAVG samples from full audio-video captions, while HunyuanVideo-Foley (Shan et al., 2025) generates 2.5k V2A samples from the ground-truth videos and corresponding audio captions. For the human-centered subset, Ovi generates 1k JAVG samples, while InfiniteTalk (Yang et al., 2025b) and an internal model each generate 1k A2V samples using the ground-truth audio with visual conditioning. Together with the 3.5k ground-truth clips, this gives 11.5k candidate samples. We further filter the generated pool for basic audio-video quality and annotation completeness, retaining 8k clips for SFT.

Human-calibration source pool. We construct the human-calibration source pool from JavisBench (Liu et al., 2025). Its relatively large scale, paired ground-truth references, and broad content coverage provide a practical starting point for constructing diverse positive and negative examples when high-quality audio-video data remain limited. We sample 3.2k clips from each of five sources: JavisBench ground truth, Ovi-I2V, Ovi-T2V, HunyuanVideo-Foley, and an internal A2V model. The generated samples use the corresponding JavisBench inputs, covering JAVG, V2A, and A2V paradigms. These 16k clips form the initial candidate pool. We run the SFT model f_1 eight times on each candidate to estimate sample difficulty and prioritize cases for which the model has not yet formed stable judgments. To preserve content diversity, we cluster ImageBind (Girdhar et al., 2023) audio-video embeddings into 300 clusters and sample across clusters. We also construct global-narrative negatives by swapping audio tracks across clusters. Finally, we sample original candidates and constructed global negatives at a 3:1 ratio to obtain the 10k rollout set used for GRPO.

The resulting data pool covers different generation paradigms and provides the source samples for the two complementary supervision pipelines described below.

Subset	Source model	Paradigm	Raw clips	SFT	Human calib.
General	Ground truth	GT	2.5k	✓	
General	Ovi	JAVG	2.5k	✓	
General	HunyuanVideo-Foley	V2A	2.5k	✓	
Human	Ground truth	GT	1.0k	✓	
Human	Ovi	JAVG	1.0k	✓	
Human	InfiniteTalk	A2V	1.0k	✓	
Human	Internal model	A2V	1.0k	✓	
JavisBench	Ground truth	GT	3.2k		✓
JavisBench	Ovi-I2V	JAVG	3.2k		✓
JavisBench	Ovi-T2V	JAVG	3.2k		✓
JavisBench	HunyuanVideo-Foley	V2A	3.2k		✓
JavisBench	Internal model	A2V	3.2k		✓

Table 5: Sources and statistics of the training data. The SFT source pool contains 11.5k candidate clips, of which 8k are retained after supervision filtering. The human-calibration source pool contains 16k candidates, from which the 10k GRPO rollout set is constructed.

Stage	Input	Operation	Tool	Output
1. Audio analysis	Audio A + SED hints T_{SED}	Analyze the audio scene, layers, events, and timestamps.	Gemini 2.5 Pro + SED hints	$H_A = (G_A, E_A)$: audio context and timestamped audio events.
2. Video analysis	Video V	Analyze the visual scene and localize audio-relevant events.	Qwen3-VL-Max	$H_V = (G_V, E_V)$: visual context and timestamped visual events.
3. A→V verification	H_A from Stage 1 + V	Verify visual evidence, timing, and applicable event attributes for each audio event.	Qwen3-VL-Max	$R_{A\to V}$: audio-to-video verification records.
4. V→A verification	H_V from Stage 2 + A	Verify expected audio, reasonable silence, timing, and applicable event attributes for each visual event.	Gemini 2.5 Pro	$R_{V\to A}$: video-to-audio verification records.
5. Aggregation	$H_A, H_V, R_{A\to V}, R_{V\to A}$ + original AV	Re-examine the clip and aggregate the verified evidence.	Gemini 2.5 Pro	$(S_{\text{mac}}, S_{\text{temp}}, S_{\text{attr}}, \kappa)$.

Table 6: Summary of the five-stage automated annotation pipeline. Stages 1–2 produce modality-specific analysis records, Stages 3–4 verify their event relations in the opposite modality, and Stage 5 aggregates all records into the final three-level judgment.

B.2 Five-Stage Automated Annotation Pipeline

To construct structured supervision at scale, we implement a five-stage automated annotation pipeline that first analyzes audio and video independently, then verifies event relations in both directions, and finally aggregates the structured evidence into the three consistency levels. Table 6 summarizes the input, operation, tool, and output of each stage. Fig. 6 illustrates the complete data flow.

Pipeline details. Table 6 and Fig. 6 summarize the complete annotation procedure.

Stages 1 and 2 first parse the two modalities independently. The audio branch separates speech, music, and other sound events, and records their scene context, speaker information, acoustic attributes, and timestamps. The video branch extracts the global scene, people, emotions, and timestamped visual events that may be relevant to audio.

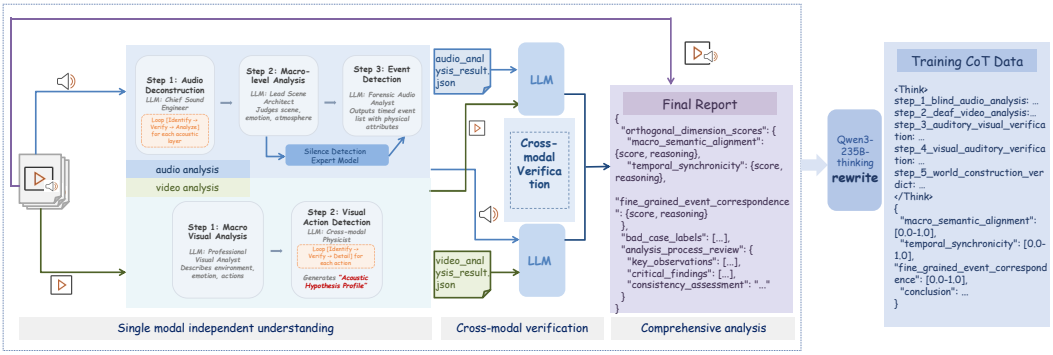

Figure 6: Overview of the five-stage automated annotation pipeline. Audio and video are first analyzed independently, followed by bidirectional event verification and final aggregation into three consistency scores and a diagnostic conclusion.

Stages 3 and 4 then perform bidirectional event verification. For each audio event, audio-to-video verification searches its temporal window for supporting visual evidence and, when applicable, examines temporal alignment and fine-grained physical attributes. The absence of a visible source is retained as evidence rather than directly treated as an inconsistency, allowing later reasoning over off-screen audio and other asymmetric relations. Conversely, video-to-audio verification first determines whether a visual event should naturally produce sound. Naturally silent actions are preserved as reasonable silence, while sound-producing actions are checked for corresponding acoustic evidence, temporal alignment, and physical compatibility.

Stage 5 aggregates the modality-specific analyses and the two directional verification records. The evaluator also re-examines the original audio-video clip instead of relying only on intermediate annotations. It then produces scores for global narrative harmony, event-level temporal synchronization, and fine-grained attribute alignment, together with a diagnostic conclusion that summarizes the main consistent or inconsistent events.

Harness-based process control. The three evaluation levels require several non-trivial judgments, so we do not rely on a single unconstrained model response to construct the supervision. Instead, the annotation harness introduces explicit process controls around the underlying expert models. First, audio and video are analyzed independently, preventing early cross-modal information from directly biasing event discovery. Second, each detected event is converted into explicit timestamps and evidence records, and later checks operate on these records rather than restarting from an unconstrained global judgment. Third, verification is conditional: existence is checked before temporal or attribute compatibility, and inapplicable cases such as reasonable silence are explicitly bypassed. Fourth, the same relation is checked in both audio-to-video and video-to-audio directions, providing a cross-check for missing and hallucinated events. Finally, the aggregation stage receives all intermediate evidence but also re-examines the original audio-video clip, forming a final feedback check against errors propagated from earlier stages.

These controlled verification records are then transferred to AVC-Reward through both training stages. During SFT, the complete five-stage annotation trajectory, including single-modal analyses, bidirectional event verification, three scores, and the diagnostic conclusion, is used as the full supervision target. The model is therefore trained to reproduce the same staged verification process rather than learning only a direct mapping from the input clip to final scores. During GRPO, the structure reward explicitly rewards the presence of all five reasoning stages, while score distillation aligns the resulting judgments with human-calibrated scores. Together, the two stages preserve the verification procedure established by the annotation harness while transferring it from the cascaded expert pipeline into a single reward model. Whether this learned procedure leads to correct judgments is further evaluated through human alignment, training ablations, and qualitative cases.

B.3 Human Annotation and Score Calibration

In human annotation, we found that score-only judgments showed noticeable subjective variation and made disagreements difficult to verify. We therefore design a 39-label taxonomy that asks annotators to record explicit consistency states and failure types before score calibration, reducing unconstrained subjectivity while preserving observable information about event presence, timing, audio roles, and cross-modal relations. The taxonomy follows the same three dimensions as AVC-Reward: global labels cover narrative, affective, and environmental consistency; temporal labels separately describe physical interactions, speaker events, and music/rhythmic events, including missing, hallucinated, off-screen, and narrative-layer cases; fine-grained labels describe material, intensity, spatial, motion, and causal-response attributes. Table 7 summarizes the complete annotation space.

Annotation protocol. Annotators judge the three dimensions independently. Within each dimension, all applicable failure labels can be selected, allowing a clip to contain multiple inconsistency types or multiple affected events. For temporal synchronization, annotators additionally distinguish physical-layer on-screen events, plausible off-screen events, and narrative-layer audio such as voice-over and background music. A severe global conflict can be marked as unjudgeable for a later event-level category when the corresponding relation cannot be meaningfully assessed, but this does not otherwise propagate the global judgment to the other dimensions.

For each annotated sample, three annotators independently provide the dimension labels. The training and test sets are strictly disjoint. The held-out test set is additionally assigned to annotators who do not participate in labeling the training set, preventing overlap in both samples and annotators between model training and final human evaluation.

Label-to-score calibration and score rewriting validation. Human annotations provide detailed categorical supervision, while AVC-Reward requires continuous targets for the three dimensions. Directly collecting continuous scores for all training samples introduces annotation subjectivity and variance, making large-scale score supervision difficult to verify and maintain consistently. Therefore, we investigate whether human-verified labels can serve as an intermediate representation for recovering continuous scores.

We first study the relationship between human-verified labels and continuous scores on AVC-Benchmark, where both label annotations and human score annotations are available. We analyze whether label configurations consistently correspond to similar human scores. Pairs sharing the same human-verified label configuration have an average absolute score difference of only 0.0321 (median 0.0111), compared with 0.2002 (median 0.1000) for pairs with different configurations (Table 8). This indicates that label configurations strongly constrain human scores and motivates label-to-score calibration for continuous reward construction.

Based on this observation, we further investigate whether a large language model can learn the label-to-score mapping through few-shot calibration. Specifically, we extract 10 representative label-to-score mappings from the AVC-Benchmark annotations and provide them to Qwen3-235B as few-shot demonstrations for score rewriting.

Given these demonstrations, Qwen3-235B maps each label record to $(S_{\mathrm{mac}}, S_{\mathrm{temp}}, S_{\mathrm{attr}})$ and a diagnostic conclusion. The prompt specifies label semantics, failure severity, multi-label handling, and output format; details are provided in the supplementary materials.

We evaluate the rewriting procedure on AVC-Benchmark samples with human scores, excluding the manually selected few-shot examples used for prompt construction. As shown in Table 9, the rewritten scores preserve human ordering across all three dimensions; the averaged score achieves PLCC 0.780 and Δ-Pairwise@0.2 of 0.905.

We further evaluate score rewriting in the score-distillation reward space used for GRPO (Table 10). Rewritten scores achieve $r_{\mathrm{distill}} = 0.853$, compared with 0.881 for human leave-one-annotator-out references. The corresponding reward-space discrepancies, measured by $1 - r_{\mathrm{distill}}$, are 0.147 and 0.119, respectively. These results suggest that rewriting errors are comparable in magnitude to the observed inter-annotator variability, with a modest additional discrepancy of 0.028 in the reward space directly relevant to GRPO supervision.

Dimension	Event type	Label	Annotation meaning
Global narrative harmony	Global	M-A	Globally consistent
		M-B	Narrative-content conflict
		M-C	Emotion or atmosphere mismatch
		M-D	Ambient sound or scene-setting conflict
		M-E	Other global semantic inconsistency
Event-level temporal synchronization	Physical interaction	T-P-A	On-screen event correctly matched
		T-P-B	On-screen event with missing audio
		T-P-C	Audio hallucination or implausible off-screen audio
		T-P-D	Temporal offset between action and sound
		T-P-E	Plausible off-screen audio
		T-P-F	No applicable event
		T-P-G	Other temporal mismatch
		T-P-H	Unjudgeable due to severe global inconsistency
	Speaker	T-S-A	Physical-layer speech correctly matched
		T-S-B	Missing speech audio
		T-S-C	Speech audio hallucination
		T-S-D	Speech-stream onset/offset mismatch
		T-S-E	Narrative-layer voice-over
		T-S-F	No applicable speaker event
		T-S-G	Other speaker temporal mismatch
		T-S-H	Unjudgeable due to severe global inconsistency
	Music / rhythmic	T-M-A	Physical-layer music correctly matched
		T-M-B	Missing music audio
		T-M-C	Music audio hallucination
		T-M-D	Physical-layer rhythmic misalignment
		T-M-E	Appropriate narrative-layer background music
		T-M-F	Narrative-layer rhythmic misalignment
		T-M-G	No applicable music or rhythmic event
		T-M-H	Other music or rhythmic temporal mismatch
		T-M-I	Unjudgeable due to severe global inconsistency
Fine-grained attribute alignment	Event attribute	F-A	Fully aligned
		F-B	Material or timbre mismatch
		F-C	Energy or intensity mismatch
		F-D	Spatial or distance mismatch
		F-E	Fine-grained state or motion mismatch
		F-F	Implausible causal reaction
		F-G	Event does not pass temporal verification
		F-H	Other fine-grained attribute mismatch
		F-I	Unjudgeable due to severe global inconsistency

Table 7: Human annotation taxonomy with 39 labels. Global narrative harmony evaluates clip-level narrative and scene coherence; event-level temporal synchronization separately covers physical interactions, speaker events, and music or rhythmic events; fine-grained attribute alignment examines the internal audio-visual attributes of temporally matched events.

The similar rewriting performance on the full and evaluation sets indicates that the few-shot label-to-score mapping generalizes beyond the manually selected examples. The remaining gap to human leave-one-out agreement reflects the approximation introduced by automatic score rewriting.

For GRPO training, we construct a separate label-to-score calibration prompt using only the GRPO training data. Specifically, we randomly sample 100 examples from the GRPO data pool and collect additional continuous score annotations from human annotators. From these examples, we extract another 10 representative label-to-score mappings and use them as few-shot demonstrations for Qwen3-235B to rewrite scores for GRPO training samples. This calibration process is completely

Label configuration	Pairs	Mean diff.	Median diff.
Same	6,717	**0.0321**	**0.0111**
Different	242,148	0.2002	0.1000

Table 8: Absolute differences between human scores for sample pairs with the same or different label configurations.

Score	PLCC	Δ-Pairwise@0.2
Global narrative harmony	0.591	0.712
Temporal synchronization	0.454	0.816
Fine-grained attribute alignment	0.601	0.798
Overall	**0.780**	**0.905**

Table 9: Agreement between rewritten and human scores after excluding the manually selected few-shot examples.

isolated from AVC-Benchmark, and no benchmark annotations are used for GRPO score rewriting or training.

The supplementary material provides the complete rewriting prompt.

C Training Details

C.1 SFT Cold Start

We initialize AVC-Reward from Qwen3-Omni-Thinking and perform full-parameter fine-tuning using the 8k structured annotation samples described in Sec. B.2. For each sample, the training target contains the complete five-stage reasoning sequence, followed by the three consistency scores $(S_{\mathrm{mac}}, S_{\mathrm{temp}}, S_{\mathrm{attr}})$ and the diagnostic conclusion. We train the model with the ms-swift framework for one epoch. The complete SFT configuration is reported in Table 11.

C.2 GRPO for Human Alignment

Starting from the SFT checkpoint, we further align AVC-Reward with the human-calibrated scores using GRPO. Training is implemented with ms-swift on the 10k rollout subset constructed from the human-calibration data.

The reward combines a structure reward r_{struct}, which enforces the required reasoning and output format, with a score-distillation reward r_{distill}, which encourages the three predicted scores to match their human-calibrated references.

Structure reward. The structure reward $r_{\mathrm{struct}} \in [0, 1]$ checks whether the model follows the required five-stage reasoning and output format. Its four components are summarized in Table 12.

The five-stage markers are:

- `step_1_blind_audio_analysis`
- `step_2_deaf_video_analysis`
- `step_3_auditory_visual_verification`
- `step_4_visual_auditory_verification`
- `step_5_world_construction_verdict`

We additionally use r_{struct} as a format gate: if $r_{\mathrm{struct}} < 0.4$, we set $r_{\mathrm{distill}} = 0$ and truncate the total reward to

$$r_{\mathrm{total}} = 0.5\, r_{\mathrm{struct}}.$$

This prevents malformed outputs from receiving a high reward solely from numerically accurate final scores.

Setting	Comparisons	r_{distill}	$1 - r_{\text{distill}}$
Rewrite, full evaluation set	706	0.853	0.147
Human leave-one-out	2,118	0.881	0.119

Table 10: Score-rewriting agreement measured with the score-distillation reward used during GRPO. Human leave-one-out results provide a reference for annotator variation.

Configuration	Value
Base model	Qwen3-Omni-Thinking
Training framework	MS-SWIFT
Fine-tuning	Full-parameter
Training samples	8k
Epochs	1
Global batch size	64
Peak learning rate	5×10^{-6}
LR scheduler	Constant with warmup
MoE weight	0.001

Table 11: Training configuration for the SFT cold-start stage.

Score-distillation reward. For structurally valid outputs, score agreement with the human-calibrated reference is measured using a Gaussian kernel:

$$r_{\text{distill}} = \frac{1}{3} \sum_{i=0}^{2} \exp\left(-w_i(\text{pred}_i - \text{ref}_i)^2\right),$$

where $(\text{pred}_0, \text{pred}_1, \text{pred}_2)$ are the predicted $(S_{\text{mac}}, S_{\text{temp}}, S_{\text{attr}})$ scores and $(\text{ref}_0, \text{ref}_1, \text{ref}_2)$ are their calibrated human references.

We use

$$w_0 = \max(1.0,\, 10 - 10\,\text{ref}_0), \qquad w_1 = w_2 = 5.0.$$

The adaptive w_0 increases the penalty for rare samples with low global narrative harmony, while the temporal and attribute dimensions use fixed weights of 5.0.

For outputs that pass the structure gate, the final reward is

$$r_{\text{total}} = \frac{2.0\, r_{\text{struct}} + 1.0\, r_{\text{distill}}}{3.0}.$$

This weighting keeps the structured event-centric reasoning process as a strong training constraint while simultaneously aligning the predicted scores with human judgments.

GRPO configuration. We train with group size $G = 8$ for 1,700 optimization steps. The detailed configuration is reported in Table 13.

C.3 Generation-Side RL with AVC-Reward

To evaluate AVC-Reward as an optimization signal, we build on the generation-side RL framework of OmniNFT (Zhang et al., 2026a). We keep the overall training framework unchanged and replace its original audio-video reward DeSync (Iashin et al., 2024) with AVC-Reward, using its overall score S_{AVC} as the optimization signal, with reward-specific learning-rate and weight adjustments for comparable optimization strength. This isolates the effect of the proposed reward while retaining the original generation training setup.

OmniNFT jointly optimizes audio quality, video quality, and audio-video consistency rewards; under the original weighting, AVC-Reward therefore has a weaker relative optimization effect than the original DeSync reward. We use a two-stage schedule to compensate for this difference. In Stage 1, we retain OmniNFT's original learning rate of 3×10^{-5}, set the AVC-Reward weight to 2.0, and

Condition	Reward
<think> content is present	+0.2
Each required stage marker is present	$+0.1 \times k$ ($k \leq 5$)
All three predicted scores are present	+0.2
Diagnostic conclusion is present	+0.1

Table 12: Components of the structure reward r_{struct}.

Configuration	Value
Training framework	MS-SWIFT
Rollout samples	10k
Training steps	1,700
Global batch size	32
Group size G	8
Learning rate	1×10^{-6}
LR scheduler	Cosine decay
Warmup ratio	0.05
KL coefficient β	0.05
Clipping parameter ϵ	0.2
Sampling temperature	0.9
Structure reward weight	2.0
Distillation reward weight	1.0

Table 13: Training configuration for GRPO human alignment.

train for 600 steps. In Stage 2, we reduce the learning rate to 1.5×10^{-5} and increase the other quality-reward weights, reducing the relative AVC-Reward weight to 1.2, and continue training for 275 steps. This setup provides generation-side RL validation of AVC-Reward while retaining a direct controlled comparison with the original OmniNFT training framework.

D AVC-Benchmark Details

D.1 Benchmark Construction and Evaluation Protocol

Sample selection. We construct AVC-Benchmark through three complementary steps. We first perform stratified sampling based on human label configurations to balance the coverage of major consistency and failure patterns, including global narrative conflicts, temporal offsets, missing or hallucinated audio, and fine-grained attribute mismatches. We then use ImageBind (Girdhar et al., 2023) to extract audio and video representations and cluster their joint embeddings, from which we sample across clusters to reduce semantic redundancy. Finally, we add explicitly constructed negative samples to strengthen the coverage of controlled failure cases that are less frequent in naturally generated data.

The resulting benchmark contains 706 samples across different generation paradigms and data sources: 172 ground truth samples, 243 JAVG samples, 100 A2V samples, 72 V2A samples, and 119 samples from the explicit-negative construction set. The JAVG subset contains Ovi-I2V, Ovi-T2V and Veo3 (Google DeepMind, 2025) outputs, the A2V subset is generated by an internal model, and the V2A subset is generated by HunyuanVideo-Foley.

Explicit negative construction. We construct additional controlled negatives from 50 Veo3 source clips. Among them, 10 contain salient emotion or global narrative cues, 20 contain clear temporal relations, and 20 contain salient audio attributes. We apply the edit operations in Table 14 to produce 119 negative samples. These edits complement naturally occurring generation failures with controlled semantic, temporal, event-existence, and fine-grained acoustic inconsistencies, balancing the failure-case distribution and broadening coverage beyond naturally occurring generation failures.

Edit operation	n
Swap audio tracks between paired clips	10
Overlay the audio track from another clip	10
Shift the audio by $+2$ s	20
Shift the audio by -2 s	20
Remove a specified audio event	20
Freeze the video for 1 s at an audio peak	19
Edit sound timbre	10
Gradually attenuate the audio	5
Insert audio from another clip into static video	5
Total	**119**

Table 14: Controlled edit operations used to construct explicit negative samples.

Human annotation. Each benchmark sample is independently evaluated by three annotators following the 39-label protocol in Appendix B.3. The benchmark annotators do not overlap with those used for the training set. Rather than directly assigning scores, annotators first inspect the three dimensions independently and select all applicable labels before providing the corresponding scores. This label-first protocol requires explicit checks of the relevant event existence, timing, attributes, and relation types, making the annotation process more traceable and reducing unconstrained score judgments. Annotators then assign a score from 0 to 1, with a step size of 0.1, for global narrative harmony, event-level temporal synchronization, and fine-grained attribute alignment.

We use agreement among the three annotators' label records within each sample as an additional quality-control signal before constructing the final benchmark reference. We further quantify inter-annotator agreement on the retained annotations, obtaining ICC(2,k) of 0.83, Krippendorff's α of 0.60, inter-annotator PLCC of 0.78, inter-annotator SRCC of 0.50, pairwise accuracy of 0.76, and Δ-Pairwise@0.2 of 0.92. The high ICC and margin-based pairwise agreement indicate stable aggregated human judgments, especially for samples with clear perceptual differences.

Human reference score. Let $s_{i,d}^{(r)}$ denote the score assigned by annotator r to sample i on dimension $d \in \{\mathrm{mac}, \mathrm{temp}, \mathrm{attr}\}$.

Since different annotators may use different scoring ranges across dimensions, we first normalize each annotator independently. Specifically, for each annotator r, we compute dimension-wise z-scores over the benchmark:

$$z_{i,d}^{(r)} = \frac{s_{i,d}^{(r)} - \mu_d^{(r)}}{\sigma_d^{(r)}},$$

where $\mu_d^{(r)}$ and $\sigma_d^{(r)}$ denote the mean and standard deviation of annotator r on dimension d computed over all benchmark samples.

To further align the scales among the three dimensions for each annotator, we apply min-max normalization over the complete three-dimensional z-score matrix of each annotator:

$$\tilde{z}_{i,d}^{(r)} = \frac{z_{i,d}^{(r)} - \min(z^{(r)})}{\max(z^{(r)}) - \min(z^{(r)})},$$

where $\min(z^{(r)})$ and $\max(z^{(r)})$ are computed over all samples and all three dimensions for annotator r.

The dimension-level human reference scores are obtained by averaging the aligned scores from three annotators:

Table 15: Sensitivity analysis of score normalization for Ours on AVC-Benchmark.

Setting	Δ-Pairwise ↑	PLCC ↑	SRCC ↑
Raw score aggregation	0.728	0.494	0.244
Z-score aggregation	0.729	0.492	0.245

$$GT_{i,d}^{human} = \frac{1}{3}\sum_{r=1}^{3} \tilde{z}_{i,d}^{(r)}.$$

The overall human reference score is then computed as:

$$S_i^{human} = \frac{GT_{i,\text{mac}}^{human} + GT_{i,\text{temp}}^{human} + GT_{i,\text{attr}}^{human}}{3}.$$

Normalization is applied only to human reference scores to calibrate subjective annotations across annotators. For automatic metrics, we do not introduce additional normalization beyond each metric's original formulation. Applying dimension-wise normalization to model outputs would additionally modify the score distributions and aggregation behaviors of individual metrics. Since different automatic metrics may have different output spaces, numbers of dimensions, and aggregation strategies, applying such transformations only for a specific evaluator would make the comparison protocol inconsistent.

For Ours, we further evaluate a z-score normalized variant as a sensitivity analysis. As shown in Table 15, the difference between raw-score and z-score aggregation is negligible, indicating that the reported results are not sensitive to this normalization choice.

The aggregation is used only for benchmark-level overall evaluation, while the three dimensions remain independently annotated and scored.

Evaluation metrics. We use Δ-Pairwise Accuracy as the primary metric for measuring metric-to-human agreement. For a margin Δ, the valid pair set is

$$\mathcal{P}_\Delta = \left\{(i,j) : \left|S_i^{\text{human}} - S_j^{\text{human}}\right| > \Delta\right\}.$$

For an automatic metric with scores m_i, we compute

$$\text{Acc}_\Delta = \frac{1}{|\mathcal{P}_\Delta|}\sum_{(i,j)\in\mathcal{P}_\Delta} \mathbb{I}\left[(m_i - m_j)(S_i^{\text{human}} - S_j^{\text{human}}) > 0\right].$$

We use $\Delta = 0.2$ as the primary setting to exclude near-tied human pairs and focus the evaluation on perceptually meaningful score differences. Standard pairwise accuracy, SRCC, and PLCC are also computed in the complete benchmark evaluation.

PLCC with five-parameter logistic mapping. Following common practice in subjective quality assessment (Sheikh, 2003), we apply a nonlinear logistic regression before computing PLCC. Human perceptual judgments are often not linearly related to objective prediction scores due to perceptual saturation and scale differences. Therefore, a mapping is first learned to align the prediction scale with the human score space.

Specifically, for each predictor and evaluation dimension, we fit a five-parameter logistic function on the same evaluation subset:

$$f(x) = b_1\left(0.5 - \frac{1}{1+\exp(b_2(x-b_3))}\right) + b_4 x + b_5, \tag{19}$$

where x denotes the predicted score and $f(x)$ is the calibrated prediction. The parameters $\{b_1, b_2, b_3, b_4, b_5\}$ are optimized by nonlinear least squares fitting between predictions and human ground-truth scores.

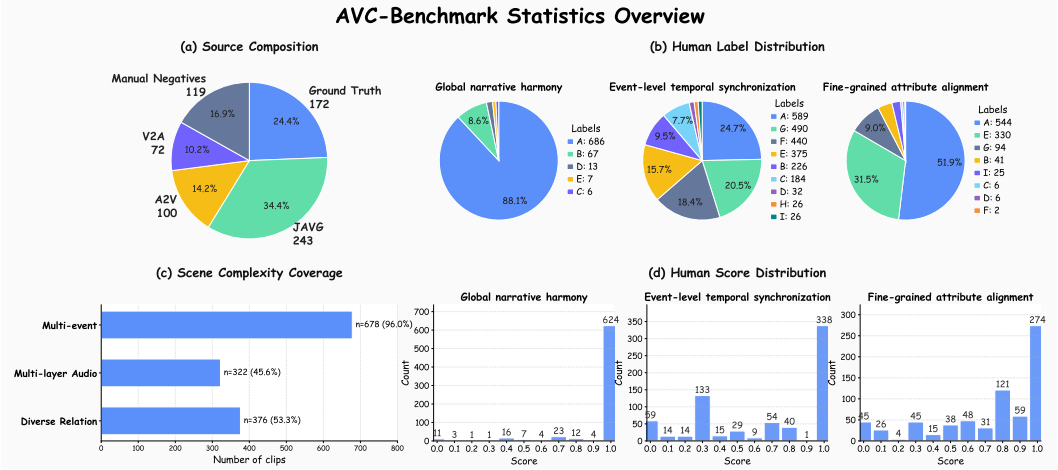

Figure 7: Statistics of AVC-Benchmark, including source composition, human-label distribution, scene-complexity coverage, and human-score distribution. The three complexity subsets are non-exclusive.

Δ	0.00	0.05	0.10	0.15	**0.20**	0.25	0.30	0.50
Valid pairs	227,819	162,008	113,124	84,407	**69,180**	59,371	52,559	30,831

Table 16: Number of valid sample pairs at different Δ-Pairwise margins using the z-score-normalized human reference.

The reported PLCC is then computed as the Pearson correlation coefficient between the calibrated prediction $f(x)$ and the human score y:

$$\mathrm{PLCC} = \frac{\mathrm{Cov}(f(x), y)}{\sigma_{f(x)}\sigma_y}. \tag{20}$$

If the nonlinear fitting fails, we fall back to Pearson correlation computed on the original prediction scores. Note that SRCC and Δ-Pairwise are computed using the original predictions without logistic calibration, since they only depend on score ordering.

D.2 Benchmark Statistics

Figure 7 summarizes the composition of AVC-Benchmark. The 706 samples cover ground truth, JAVG, A2V, V2A, and explicitly constructed negative data. Based on the human annotations, 678 samples (96.0%) exhibit multi-event composition, 322 (45.6%) contain multi-layer audio, and 376 (53.3%) contain diverse cross-modal relations. The three complexity subsets are non-exclusive, allowing a sample to exhibit multiple forms of complexity simultaneously. Together with the 39-label and human-score distributions, these statistics show that AVC-Benchmark covers different generation settings, failure patterns, and relation complexities.

Pair counts under different margins. Table 16 reports the number of valid sample pairs under different margins after constructing the z-score-normalized human reference. At the primary threshold $\Delta = 0.2$, 69,180 valid pairs remain for evaluation.

D.3 Fine-Grained Results by Scene Complexity

We further evaluate metric-to-human agreement on the three scene-complexity subsets derived from the human annotations. The subsets are non-exclusive and correspond to the three properties motivating our evaluation design: multi-event composition, multi-layer audio, and diverse cross-modal relations.

Method	Multi-event (678)	Multi-layer Audio (322)	Diverse Relation (376)
AV-Align	0.654	0.681	0.670
DeSync	0.587	0.582	0.567
ImageBind	0.626	0.658	0.632
JavisScore	0.615	0.659	0.636
AVHScore	0.597	0.635	0.611
PEAVS	0.504	0.590	0.564
AVSyncEvaluator	0.551	0.457	0.455
AV-Consistency Evaluator	0.659	0.688	0.653
AGAV-Rater	<u>0.661</u>	0.629	0.641
Gemini-3.1-Pro	0.612	0.622	0.608
Qwen3-Omni-Instruct	0.591	0.620	0.593
Qwen3-Omni-Thinking	0.597	<u>0.696</u>	<u>0.698</u>
AVC-Reward	**0.715**	**0.747**	**0.703**
Human Upper Bound	0.923	0.944	0.934

Table 17: Δ-Pairwise@0.2 on the three non-exclusive scene-complexity subsets. The best and second-best automatic metrics are shown in bold and underlined, respectively. Human Upper Bound is excluded from ranking.

Table 18: Cross-dimensional permutation analysis for validating dimension correspondence. Values denote AUC decrease after permuting each predicted dimension.

Human target	Δ Macro	Δ Temporal	Δ Fine-grained
Macro	**0.367**	0.026	-0.032
Temporal	-0.009	**0.078**	0.038
Fine-grained	-0.004	0.039	**0.102**

Table 17 shows that AVC-Reward achieves the highest Δ-Pairwise@0.2 on all three complexity subsets. The margins are clearer on multi-event and multi-layer-audio samples, where AVC-Reward improves from the second-best results of 0.661 and 0.696 to 0.715 and 0.747, respectively. On diverse-relation samples, AVC-Reward also gives the best pairwise agreement, although the margin over Qwen3-Omni-Thinking is smaller (0.703 vs. 0.698). These results show that the event-centric evaluation is particularly effective for samples involving multiple events and layered audio, while retaining its advantage when audio and video are related through off-screen, narrative, or other non-trivial relations.

D.4 Dimension Correspondence Analysis

To verify whether the three predicted scores correspond to their intended consistency dimensions, we perform a dimension correspondence analysis based on human-annotated issue labels.

Instead of directly measuring the correlation between predicted dimension scores and human dimension scores, we formulate this analysis as a dimension-specific issue classification task. Human dimension scores are mainly designed for overall perceptual evaluation, and the three dimensions may have different levels of granularity. Therefore, compared with continuous scores, human issue labels provide a more direct signal for evaluating whether each predicted dimension corresponds to its intended consistency dimension.

Specifically, we construct dimension-specific binary targets from human-annotated issue labels in the benchmark, where samples are separated into positive and negative cases according to whether they contain the corresponding types of consistency issues. We then use the three predicted scores from Ours as input features and evaluate a classifier with cross-validated AUC.

To measure the contribution of each predicted dimension, we randomly permute one prediction dimension while keeping the remaining dimensions unchanged. The performance decrease is calculated as:

$$\Delta AUC = AUC_{\text{original}} - AUC_{\text{permuted}}.$$

A larger ΔAUC indicates that the corresponding predicted dimension provides more information for identifying the target consistency issue category. Results are averaged over repeated cross-validation runs. As shown in Table 18, the largest performance drop is consistently observed when permuting the corresponding dimension, supporting the correspondence between model outputs and human-defined consistency dimensions.

E More Case Studies

Full AVC-Reward input–output. Fig. 8 shows a complete input–output example of AVC-Reward. The clip contains a coherent kitchen scene with several events, including chef speech, chopping, and an off-screen phone ring. The phone ring can plausibly originate outside the frame, but the visible chef shows no reaction or other visual evidence supporting its interaction with the scene. AVC-Reward first analyzes the audio and video independently to extract their scenes and major events, and then performs audio-to-video and video-to-audio verification for the extracted events. The final stage combines the scene-level information and cross-modal evidence to judge the three consistency dimensions. The complete reasoning process is contained in the `<think>` block, followed by the three scores and a diagnostic conclusion.

Additional evaluation cases. Fig. 9, together with the evaluation cases in Fig. 5, further shows how AVC-Reward handles different forms of complexity in generated audio-video content. The two-person conversation case in Fig. 5 tests multi-event disentanglement: two people are visible, but the audio provides clear speech evidence for only one of them. The puppet case tests multi-layer audio, where character speech, ambient factory sounds, and background music play different roles in the scene.

The additional cases extend this evaluation to several other situations. The band case contains multiple instruments and vocals at the same time, testing concurrent event verification. The choir case changes from a conductor to the singers and requires the evaluator to verify different valid audio-video relations across the clip. The hammer case tests precise event-level temporal verification by introducing a clear delay between the visible impact and its sound. The concert-ending case tests fine-grained visual evidence for sound production: the performance has ended and the orchestra begins packing up and bowing, although the performance audio continues. Across these cases, AVC-Reward remains close to human judgments and provides conclusions that identify the relevant event or relation. The positive choir and band examples further show that complex but valid relations can still receive high consistency judgments.

Additional generation cases. Fig. 10, together with the generation cases in Fig. 5, shows how the same event-level evaluation capabilities are reflected in generation after using AVC-Reward for optimization. In the dog–cat case in Fig. 5, the optimized model better preserves sequential multi-event composition, with the dog and cat opening their mouths for their respective sounds. In the cooking case, it better handles multi-layer audio by retaining narration while also generating visible chopping with synchronized chopping sounds.

The additional cases cover further generation complexities. In the waterfall case, the optimized model preserves concurrent waterfall sound and human speech. In the election-ad case, visual cuts, sound effects, footsteps, and narration form different event-dependent relations that are more consistently generated. The water-pouring case shows finer audio attributes, where the water-flow sound changes with the evolving visual event. The piano case shows finer visual evidence for sound production, with finger presses, key motion, and pitch changes forming a more coherent relation. Together, these examples qualitatively mirror the capabilities examined by AVC-Reward: finer checks of individual events, their cross-modal relations, audio attributes, and visible sound-producing evidence lead to more complete event generation and more consistent audio-video relations.

Together, these cases cover the main capabilities examined on both the evaluation and generation sides, including multi-event composition, concurrent events, complex cross-modal relations, multi-layer audio, precise temporal verification, and fine-grained audio or visual evidence.

INPUT

System Prompt

You are a professional Audio-Video Consistency Evaluation Expert. Your core evaluation principle is based on 'World Construction': Consider audio as the 3D ground truth of the world, and video as a limited 2D observation window. Your task is to assess the quality of audio-video generation by strictly following this 5-layer reasoning pipeline:
1. Blind Audio Analysis: Establish the auditory world cognition independently without visual cues.
2. Deaf Video Analysis: Analyze visual events independently without auditory cues.
3. Auditory-Visual Verification (A2V): Using audio as ground truth, verify if visual events align physically and temporally.
4. Visual-Auditory Verification (V2A): Using video as an anchor, check for missing or physically incorrect auditory feedback.
5. World Construction Verdict: Synthesize findings. Crucially, distinguish between:
- In-Window (Diegetic): Sounds from visible sources (must sync perfectly).
- Off-Screen (Diegetic): Sounds from the implied 3D world outside the frame (must match spatial logic).
- Narrative (Non-Diegetic): BGM or Voiceover (must match mood/semantic context).
Finally, output scores for Macro-Semantic Alignment, Temporal Synchronization, and Fine-Grained Event Correspondence.

User Input

<image>

<video><audio>Please analyze the audio-video consistency of this material and provide a rating.

A chef chops vegetables while talking about his dish. After a while, he reaches for food from another bowl as an off-screen phone rings, but he shows no reaction to the ringing.

OUTPUT

<think>

```
{
  "step_1_blind_audio_analysis": {
    "detected_events": [
      {
        "id": 1,
        "timestamp": "[0.00s - 4.45s]",
        "type": "Human Speech",
        "description": "Masculine-like voice (S1): I think the key ..."
      },
      {
        "id": 2,
        "timestamp": "[5.19s - 6.83s]",
        "type": "Human Speech",
        "description": "Masculine-like voice (S1): Oh, hang on, I ..."
      },
      {
        "id": 3,
        "timestamp": "[0.00s - 7.55s]",
        "type": "Other Sound Events",
        "description": "sustained, multi-toned, pitched electronic tone"
      }
    ],
    "scene_hypothesis": "A professional cooking show or a high-end restaurant kitchen."
  },
  "step_2_deaf_video_analysis": {
    "scene_overview": "A professional kitchen with stainless steel surfaces, cooking equipment, and a chef in a white coat at a wooden cutting board.",
    "key_visual_events": [
      {
        "id": 1,
        "timestamp": "[0.00s - 3.00s]",
        "event": "Chef chopping red peppers on a wooden cutting board"
      },
      {
        "id": 2,
        "timestamp": "[0.00s - 7.50s]",
        "event": "Chef speaking while cutting vegetables"
      }
    ]
  },
  "step_3_auditory_visual_verification": {
    "process": [
      {
        "event_id": 1,
        "audio_source": "Masculine-like voice (S1): I think the key ...",
        "timestamp": "[0.00s - 4.45s]",
        "verification_result": "PASS",
        "analysis": "Existence Check: Pass; Physical Consistency: Pass",
        "detailed_evidence": [
          {
            "evidence_type": "Visual evidence for transient event 'Masculine-like voice (S1): I think the key ...' in time window 0.00s-4.45s",
            "reasoning": "Timbre Match: The visual evidence within the specified time window shows a chef actively speaking. His mouth is visibly moving in a manner consistent with articulate speech. The timbre of a human voice is produced by the human vocal apparatus, which is exactly what is being observed. There is a direct and unambiguous causal link between the visual action of speaking and the resulting audio of a human voice. Therefore, the visual actions are perfectly compatible with the detected audio event; Temporal Match: The synchronization between the audio and video is excellent. The chef's lip movements and facial expressions align perfectly with the words being spoken. The timing of his speech, starting at 0.00s and continuing until 4.45s, matches the visual of him talking while chopping the peppers."
          }
        ]
      },
      {
        "event_id": 2,
        "audio_source": "Masculine-like voice (S1): Oh, hang on, I ...",
        "timestamp": "[5.19s - 6.83s]",
        "verification_result": "PASS",
        "analysis": "Existence Check: Pass; Physical Consistency: Pass",
        "detailed_evidence": [
          {
            "evidence_type": "Visual evidence for transient event 'Masculine-like voice (S1): Oh, hang on, I ...' in time window 5.19s-6.83s",
            "reasoning": "Timbre Match: Within the specified time window, the chef is visually observed speaking. His mouth opens and moves in a manner consistent with the articulation of the words heard in the audio. The visual action of a human speaking is the direct cause of the sound of a human voice. Therefore, the visual evidence perfectly matches and explains the timbre of the detected audio event.; Temporal Match: The synchronization between the audio and video is excellent. The chef's lip movements and facial expressions precisely match the timing and cadence of the spoken words, \"Oh, hang on, I...\". The visual action of speaking is the direct cause of the audio, and there is no discernible delay or mismatch."
          }
        ]
      },
      {
        "event_id": 3,
        "audio_source": "sustained, multi-toned, pitched electronic tone",
        "timestamp": "[0.00s - 7.55s]",
        "verification_result": "FAIL",
        "analysis": "Existence Check: Fail; Physical Consistency: Fail",
        "detailed_evidence": [
          {
            "evidence_type": "Visual evidence for transient event 'sustained, multi-toned, pitched electronic tone' in time window 0.00s-7.55s",
            "reasoning": ""
          }
        ]
      }
    ]
  },
  "step_4_visual_auditory_verification": {
    "process": [
      {
        "action_id": 1,
        "video_action": "Chef chopping red peppers on a wooden cutting board",
        "timestamp": "[0.00s - 3.00s]",
        "verification_result": "PASS",
        "detailed_analysis": {
          "timbre": "The audio signal in this interval consists of two distinct layered components. The primary sound is a series of sharp, percussive impacts, which is highly characteristic of a knife hitting a dense, solid surface like a wooden cutting board. The physical properties of this sound—its high-frequency content (sharp 'chop' or 'thwack') and rapid decay—align perfectly with the physics of a blade striking wood. The secondary sound is a faint, high-pitched frictional noise, consistent with the sound of a knife blade scraping against the surface of the red peppers. The combination of these two sounds—the impact on wood and the friction on the vegetable—creates the exact acoustic signature expected from this specific cooking action. The timing and rhythm of the impacts also match the cadence of chopping.",
          "temporal": "The analysis confirms a perfect temporal alignment. The first visual impact of the knife on the peppers occurs at approximately 0.10s. This moment coincides precisely with the first audible chopping sound. This one-to-one correspondence between the visual chop and the audio chop continues consistently throughout the 0.00s to 3.00s interval. There is no perceptible delay or lead between the audio and video events, indicating that the audio is synchronized with the video."
        }
      },
      {
        "action_id": 2,
        "video_action": "Chef speaking while cutting vegetables",
        "timestamp": "[0.00s - 7.50s]",
        "verification_result": "PASS",
        "detailed_analysis": {
          "timbre": "The audio analysis strongly supports the visual action. The physical properties of the sound are highly consistent with a human speaking. The frequency spectrum contains a clear fundamental frequency and harmonics characteristic of a male voice (S1), with energy concentrated in the typical 100 Hz to 4 kHz vocal range. The amplitude dynamics show distinct modulation, with peaks and troughs corresponding to the syllables and words of the spoken phrase. This is the expected acoustic signature for speech. The lack of other prominent sounds rules out significant concurrent events, making the audio a direct match for the visual of a person speaking.",
          "temporal": "The audio is perfectly synchronized with the video. The chef's lip movements align precisely with the spoken words, forming a clear and direct match between the visual action of speaking and the corresponding audio track. There is no discernible lag or offset."
        }
      }
    ]
  },
  "step_5_world_construction_verdict": {
    "dimension_reasoning": {
      "macro_semantic": "The macro-semantic alignment is perfect. The audio hypothesis of a 'professional cooking show' is visually confirmed. The chef's 'Enthusiastic' emotion, as described in the audio analysis, is mirrored by his animated and happy visual expressions. The scene's atmosphere is cohesive, with the continuous background hum from the kitchen equipment providing a realistic and consistent soundscape for the visual environment.",
      "temporal": "Temporal synchronization is excellent. The speaker's lip movements are perfectly in sync with their speech throughout the clip. The timing of the chef's hand gestures and movements also aligns naturally with the cadence of his words and actions. There are no noticeable delays or mismatches.",
      "fine_grained": "The correspondence is mixed. The primary audio event, the chef speaking, is perfectly matched with his lip movements and expressions. However, there is a significant fine-grained mismatch: the prominent audio event of an electronic phone ringing (the 'sustained, multi-toned, pitched electronic tone') has no corresponding visual cause. The chef is the only person visible, and he is not reacting to or interacting with any phone, making the sound appear non-diegetic and breaking the cause-and-effect link between the audio and video events."
    }
  }
}
```

</think>

```
{
  "macro_semantic_alignment": 1.0,
  "temporal_synchronicity": 0.95,
  "fine_grained_event_correspondence": 0.8,
  "conclusion": "Overall, the video exhibits high consistency in its macro-semantic and temporal dimensions, creating a believable and engaging scene. However, a critical flaw in fine-grained event correspondence—a non-diegetic electronic ring—prevents it from being perfectly consistent. This makes the audio feel intrusive and out of place within the visual context, resulting in a lower score for fine-grained correspondence."
}
```


Figure 8: Full input–output example of AVC-Reward. For visualization purposes, the semantically empty lines in Stage 4 and Stage 5 are compressed due to space limitations.

Terminology note. The original prompts and outputs above retain the terminology used during data construction. Macro-Semantic Alignment, Temporal Synchronicity, and Fine-Grained Event Correspondence correspond to Global Narrative Harmony, Event-Level Temporal Synchronization, and Fine-Grained Attribute Alignment in the main text, respectively.

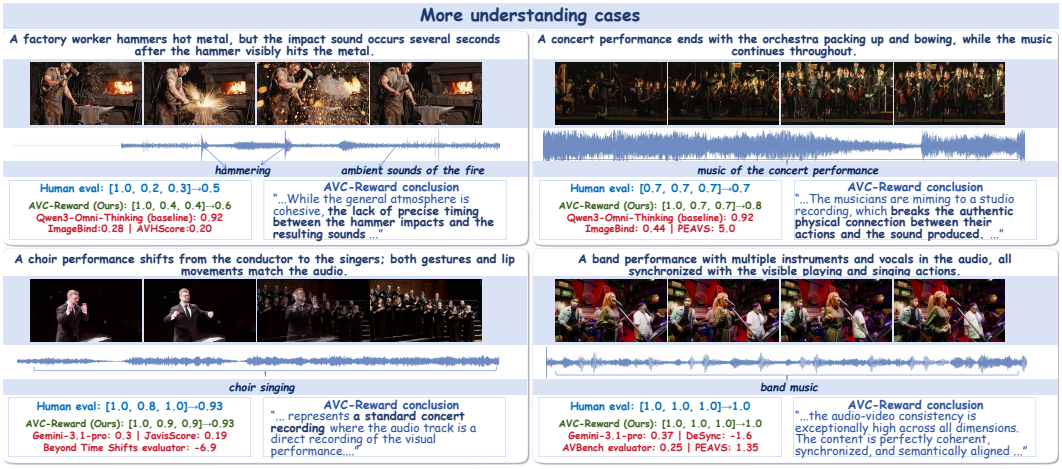

Figure 9: Additional evaluation cases of AVC-Reward.

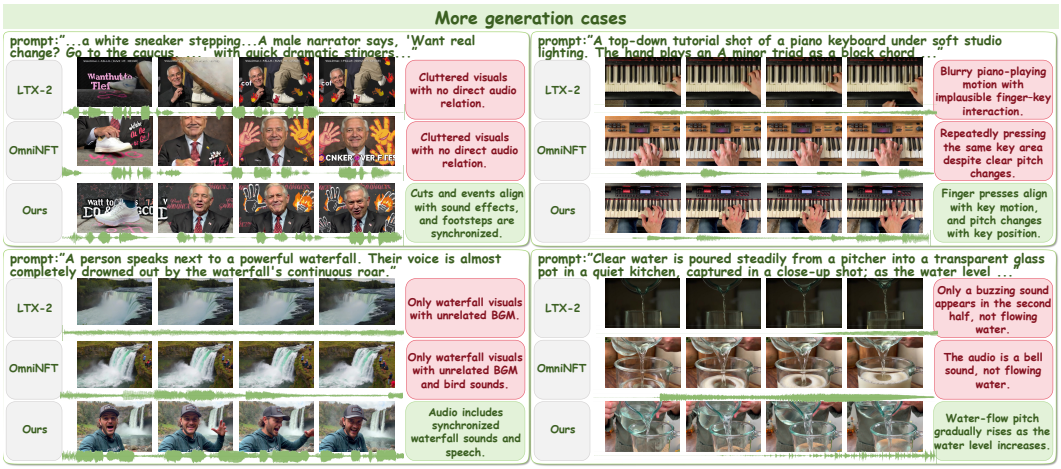

Figure 10: Additional generation cases after AVC-Reward optimization.

F Human Evaluation Protocol and Results

F.1 Sampling and Annotation Protocol

We randomly sample 100 prompts from the three generation benchmarks with fixed coverage: 30 prompts from AVGenBench, 30 from AV-Phys-Bench, and 40 from VABench. For each prompt, we generate or collect one sample from each of LTX-2, OmniNFT, and Ours. Each case therefore contains three videos with the same English prompt. The three videos are shown as anonymous options A, B, and C in a randomized order, and model identities are hidden from annotators.

Ten annotators with at least a bachelor's degree independently evaluate 30 cases each. The assignment is balanced so that every case is evaluated by exactly three annotators. Annotators answer two questions:

1. **Audio-video consistency.** Which sample better matches sound events to the visible action in timing and attributes? The judgment also considers whether narration, off-screen sound, ambient sound, and background music are appropriate for the scene.
2. **Overall quality.** Which sample is better when audio quality, visual quality, and audio-video consistency are considered together?

Table 19: Normalized human support rates and pairwise win rates. Pairwise win rates exclude double selections containing both compared systems and Tie responses.

Benchmark	Question	Ours	OmniNFT	LTX-2	Ours>OmniNFT	Ours>LTX-2
AVGenBench	Consistency	47.8	24.4	27.8	69.1	65.5
AVGenBench	Overall	46.2	27.5	26.4	66.0	64.5
AV-Phys-Bench	Consistency	50.0	27.8	22.2	73.1	71.4
AV-Phys-Bench	Overall	49.5	26.2	24.3	72.2	68.6
VABench	Consistency	46.4	34.1	19.6	63.5	72.3
VABench	Overall	44.8	36.0	19.2	57.7	70.0

Table 20: Normalized support rates and pairwise win rates on the fully consistent subset.

Question	Ours	OmniNFT	LTX-2	Ours>OmniNFT / Ours>LTX-2
Audio-video consistency	64.3	7.1	28.6	90.0 / 72.7
Overall quality	56.3	18.8	25.0	75.0 / 69.2

The default response is one selected sample. An annotator may select two samples only when they are both clearly better than the third and are difficult to distinguish from each other. `Tie` is selected when the three samples are approximately equivalent. A selection of two samples is retained as a two-element set; `Tie` is retained as a separate category.

F.2 Aggregation Rules

For a model m, let S_m be the number of responses in which m is selected. A two-element response contributes one support to each selected model, while a `Tie` contributes no support. We report the normalized support rate

$$R_m = \frac{S_m}{\sum_{m' \in \mathcal{M}} S_{m'}}, \qquad \mathcal{M} = \{\text{Ours}, \text{OmniNFT}, \text{LTX-2}\}. \tag{21}$$

For the pairwise comparison between Ours and a baseline b, a response counts as an Ours win when Ours is selected and b is not selected. It counts as a baseline win when b is selected and Ours is not selected. Responses selecting both systems or `Tie` are excluded from that pairwise denominator. The pairwise win rate is therefore

$$P_{\text{Ours}>b} = \frac{W_{\text{Ours}>b}}{W_{\text{Ours}>b} + W_{b>\text{Ours}}}. \tag{22}$$

The mean pairwise Jaccard agreement is 42.3% for audio-video consistency and 44.3% for overall quality. This medium overlap is acceptable for a comparison task with substantial subjectivity and allowed near-ties. Because the task is subjective, we additionally construct a fully consistent subset in which all three annotators select the same set for the same prompt and question, and report its model support rates below.

F.3 Benchmark-wise Results

Table 19 reports support rates and the pairwise win rates of Ours against each baseline. The results use all 300 completed judgments. As shown in Fig. 11, Ours has an advantage on both questions in every benchmark, with the clearest advantage on AV-Phys-Bench.

F.4 Fully Consistent Subset Results

As shown in Table 20, on the fully consistent subset, Ours obtains 64.3% support for audio-video consistency versus 7.1% for OmniNFT and 28.6% for LTX-2, with pairwise win rates of 90.0% and 72.7%. For overall quality, Ours obtains 56.3% support versus 18.8% and 25.0%, with pairwise

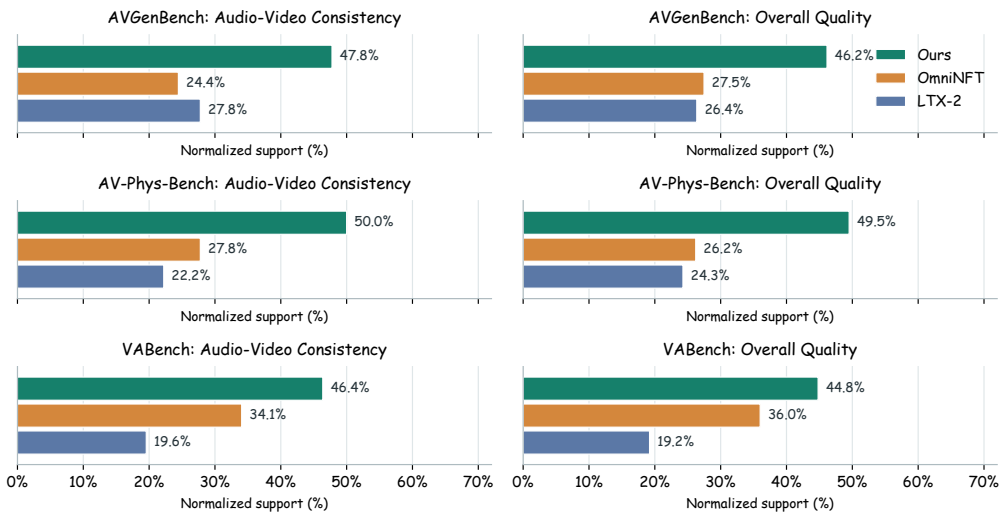

Figure 11: Benchmark-wise normalized human support rates. Each row corresponds to one benchmark, and the two columns report audio-video consistency and overall quality.

win rates of 75.0% and 69.2%. These results show that the preference advantage of Ours becomes clearer on the high-agreement subset.