VA-JUDGER: REWARD MODELING FROM HUMAN PREFERENCE FEEDBACK FOR JOINT VIDEO-AUDIO GENERATION

Anonymous authors
Paper under double-blind review

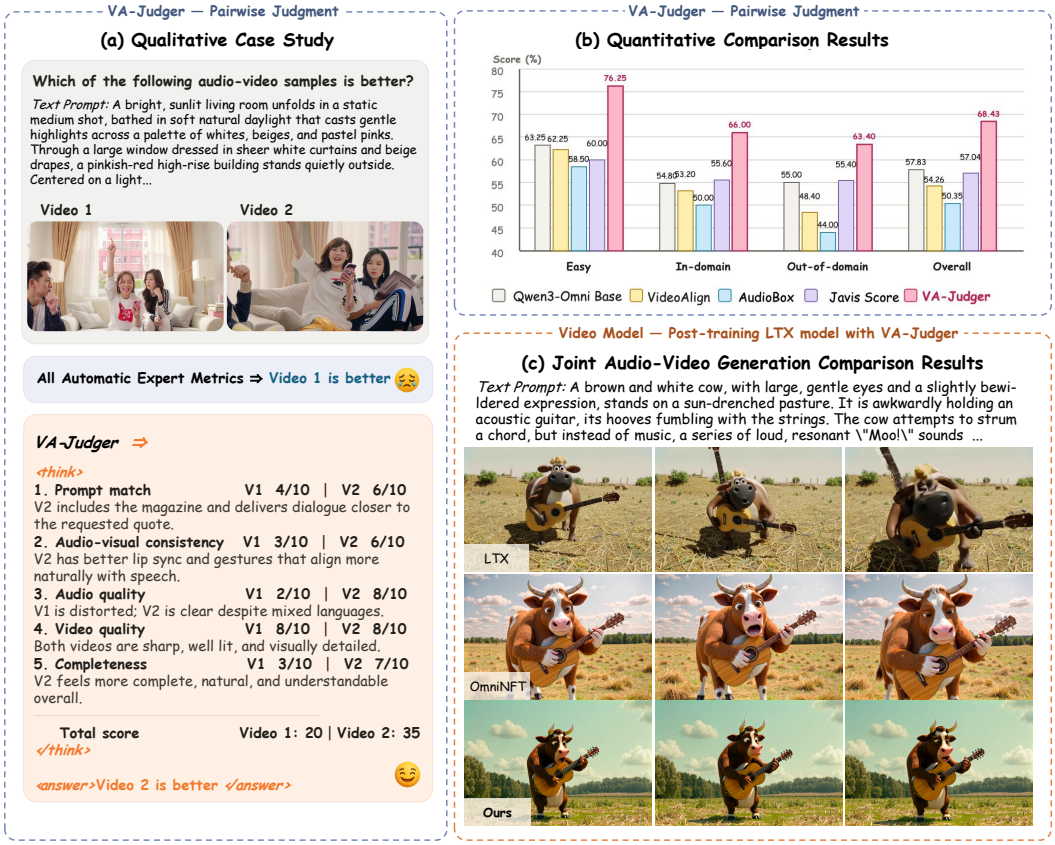


Figure 1: Evaluation of VA-Judger. (a) VA-Judger selects the preferred clip through structured dimension-wise reasoning, whereas separate evaluation metrics disagree with human judgment. (b) VA-Judger achieves the highest agreement with human preferences. (c) VA-Judger provides the reward for post-training LTX-2, improving the quality over the base model and OmniNFT.

ABSTRACT

Using reinforcement learning to post-train joint video-audio generation models requires a reward signal. Existing methods construct this reward by combining metrics for individual quality dimensions, including audio quality, visual fidelity, and synchronization. However, these metrics evaluate perceptual dimensions separately and fail to capture the overall semantic and temporal coherence among the text prompt, video, and audio that shapes human preferences. More critically, they are poorly aligned with actual human judgments. Optimizing models against these metrics encourages reward hacking, generating video-audio content that achieves high scores on these metrics yet appears incoherent or unfaithful to human viewers. To address this problem, we first construct **VAPref-10K** for joint video-audio generation, comprising 9K prompts and 10.3K fine-grained paired comparisons

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.

from generation models. We also introduce the **VA-Judger-Bench** benchmark with both in-domain and out-of-domain model comparisons to evaluate whether reward models truly align with human preferences. We further propose **VA-Judger**, a chain-of-thought omni-reward model for joint video-audio generation. In particular, VA-Judger first learns from pairs with clear quality gaps to establish structured output and coarse preference discrimination, then distills reliable preference explanations for harder near-quality comparisons via rejection sampling verified against human annotations, and finally performs dimension-wise reinforcement learning that decomposes human feedback into individual quality dimensions for denser reward signals than a single binary preference label. Experiments show that VA-Judger outperforms metric baselines in predicting human preferences on both in-domain and out-of-domain evaluations. Using its human-aligned rewards to post-train an audio-video generation model also improves most automatic metrics and human preference.

1 INTRODUCTION

Recent advancements in joint video-audio generation (OpenAI, 2025; Google DeepMind, 2025; Seedance et al., 2026; Alibaba Cloud, 2026b; Google DeepMind, 2026b; Wang et al., 2025; Low et al., 2025; Liu et al., 2026a; Team et al., 2025; 2026; Liu et al., 2026b; Chern et al., 2026; HaCohen et al., 2026; Zhou et al., 2026) have made it possible to generate increasingly realistic visual content and synchronized sound. These inspire researchers to explore post-training via reinforcement learning to align the synthesized content with perceptual quality and user intent. However, the reward signal used during reinforcement learning remains a critical bottleneck for joint video-audio generation. Current approaches such as OmniNFT (Zhang et al., 2026) construct rewards from multiple evaluation metrics, including VideoAlign (Liu et al., 2025b) and HPSv3 (Ma et al., 2025) for video quality, Audiobox Aesthetics (Tjandra et al., 2025) for audio quality, CLAP (Wu et al., 2023) for audio-text alignment, and DeSync (Iashin et al., 2024) for audio-visual synchronization. These metrics score perceptual dimensions separately and may miss semantic and temporal relations among the prompt, video, and audio. More fundamentally, such evaluation metrics are poorly aligned with actual human judgments (Liu et al., 2025b; Wang et al., 2024b). A sample may receive a high synchronization score while depicting the wrong event, obtain a strong visual score while carrying emotionally incompatible sound, or satisfy several metric criteria while failing as a coherent video-audio experience (Figure 1(a,b)). Optimizing models against such rewards encourages overfitting to narrow metric targets, synthesizing video-audio content that scores well on these metrics yet appears unfaithful to human viewers, undermining the practical value of post-training (Figure 1(c)).

This issue has been addressed in other generation domains through human preference learning. In language modeling (Li et al., 2024), preference-based optimization and reward modeling have enabled instruction-following generation (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023). Image generation has adopted human-preference datasets and reward models such as Pick-a-Pic (Kirstain et al., 2023), ImageReward (Xu et al., 2023), HPSv3 (Ma et al., 2025), and MPS (Zhang et al., 2024). Video generation has similarly begun to leverage human feedback through InstructVideo (Yuan et al., 2024), LiFT (Wang et al., 2024b), VADER (Prabhudesai et al., 2024), and VideoReward (Liu et al., 2025b). However, none of these efforts address joint video-audio generation, and their reward models cannot be directly transferred to this setting. The fundamental challenge is that human preference over video-audio content is not decomposable into independent per-modality judgments. Text-to-image/video reward models do not observe the audio stream and thus cannot assess whether sound matches the depicted event, while audio-only evaluation lacks the visual context that shapes human perception of sound plausibility. Even combining such single-modality evaluators cannot recover the cross-modal interactions that govern human preference. Addressing this problem requires both joint audio-video assessment and preference data derived from holistic human judgments of the complete text-video-audio experience, neither of which currently exists for joint video-audio generation.

In light of this, we first address the scarcity of human preference data by constructing **VAPref-10K**, a large-scale preference dataset tailored for joint video-audio generation. VAPref-10K comprises approximately 9K prompts derived from captioned web videos across seven broad video-audio categories, with 10.3K fine-grained paired comparisons generated by state-of-the-art generation models.

Rather than absolute scalar scoring for each quality dimension, which can vary substantially across human experts due to subjective calibration, annotators compare two generated clips from the same prompt, select the preferred one, and identify the quality dimensions that justify their decision using the interface in Appendix Figure 9. From VAPref-10K, we introduce **VA-Judger-Bench** as a benchmark for evaluating whether reward models align with human video-audio preferences. Beyond in-domain evaluation, VA-Judger-Bench includes an out-of-domain split built from AVGen-Bench (Zhou et al., 2026), pairing outputs from closed-source models (Kling 2.6 (Kuaishou Technology, 2025), Wan 2.6 (Alibaba Cloud, 2026a), Veo 3.1 Fast, Veo 3.1 Quality (Google DeepMind, 2025), and Sora 2 (OpenAI, 2025)). They are excluded from training.

Furthermore, we propose **VA-Judger**, a chain-of-thought omni-reward model for joint video-audio generation. Video-audio preference judgment is a complex reasoning task because it must simultaneously acquire a structured comparison format, discriminate subtle cross-modal failures, and provide dimension-level feedback. Thus, we adopt a progressive training strategy. We begin with pairs that have clear quality gaps, for which Gemini 3.1 (Google DeepMind, 2026a) generates rubric-format CoT answers containing dimension-wise scores, explanations, and a final preference to perform cold start. These generated answers help the model acquire the structured comparison format and basic quality discrimination before encountering ambiguous cases. For the more challenging comparisons between outputs with subtle quality differences, Gemini's final preferences alone are not sufficiently reliable. We thus use rejection sampling. Gemini 3.1 (Google DeepMind, 2026a) first generates CoT answers for hard pairs, and we retain only format-valid responses whose final preferences and human-selected dimension score orderings agree with the annotations. This allows the model to learn from consistently formatted reasoning traces grounded in human judgments. Finally, since a single binary preference label is too sparse to guide fine-grained preference learning, we perform dimension-wise reinforcement learning with GRPO (Shao et al., 2024). This separately rewards score orderings on human-identified quality dimensions, including visual quality, audio quality, synchronization, and semantic faithfulness, and therefore provides denser optimization signals than the final preference alone.

In conclusion, our contributions are as follows: (1) We construct VAPref-10K and VA-Judger-Bench, comprising 10.3K human-labeled comparisons across three training models and five additional benchmark models. (2) We propose VA-Judger, a chain-of-thought omni-reward model for joint video-audio generation, trained via easy-to-hard preference learning, structured-response distillation, and dimension-wise reinforcement learning with human feedback. To our knowledge, we are the first to explore reward modeling for joint video-audio generation via human feedback. (3) We introduce VA-Judger-Bench for evaluating video-audio reward models against human preferences. VA-Judger achieves better alignment with human judgments in both in-domain and out-of-domain settings. Using VA-Judger as a reward for post-training further improves human-perceived generation quality.

2 Related work

2.1 Joint video-audio generation models

Recent generative models have moved from silent video generation (Pan et al., 2026) toward joint video-audio generation. Closed-source models (OpenAI, 2025; Google DeepMind, 2025) demonstrate this trend at product scale. Regarding open-source models, UniVerse-1 stitches pre-trained video and music experts (Wang et al., 2025), whereas Ovi uses twin diffusion backbones with blockwise cross-modal fusion (Low et al., 2025). MOVA (Team et al., 2026) and daVinci-MagiHuman (Chern et al., 2026) use large-scale video data. LTX-2 (HaCohen et al., 2026) aims to handle the capacity imbalance between modalities. Baton (Tu et al., 2026) introduces explicit semantic blueprints for joint video-audio generation. The Javis (Liu et al., 2026a;b) series further emphasizes synchronization. JoyAI-Echo (Li et al., 2026) targets long-form generation. However, existing models are still primarily optimized by model-side objectives or expert metrics, leaving substantial room for improving human-perceived quality and video-audio coherence.

2.2 MULTIMODAL REWARD MODELS

Reward modeling is a standard way to turn human preference data into optimization signals. In image generation (Pan et al., 2025; Qiu et al., 2026), Pick-a-Pic (Kirstain et al., 2023) provides image preference data, and ImageReward (Xu et al., 2023) learns an image reward model. In video generation, LiFT (Wang et al., 2024b) trains LiFT-Critic for preference alignment, and VideoReward (Liu et al., 2025b) studies multi-dimensional human preference annotations. Furthermore, IXC-2.5-Reward (Zang et al., 2025) is trained on broad image and video preference data, while UnifiedReward-Think (Wang et al., 2026) provides reward signals for both multimodal understanding and generation tasks. However, neither model is trained on joint text-video-audio preference comparisons or explicitly evaluates whether the text, video, and audio form a coherent event. VA-Judger instead targets human preference alignment for joint video-audio generation by jointly assessing visual and audio quality, semantic fidelity, and cross-modal consistency.

2.3 REINFORCEMENT LEARNING FOR GENERATIVE MODELS

Recent works adapt online reinforcement learning to diffusion models. FlowGRPO (Liu et al., 2025a) makes GRPO applicable to text-to-image flow models. DanceGRPO (Xue et al., 2025) designs a unified GRPO framework for text-to-image/video and image-to-video generation. DiffusionNFT (Zheng et al., 2026) uses positive and negative generations to define an implicit policy-improvement direction. In joint video-audio generation, OmniNFT (Zhang et al., 2026) applies online diffusion reinforcement learning, while related systems such as JavisDiT++ (Liu et al., 2026b) explore preference optimization through different formulations. OmniNFT's optimization still depends on separate evaluation metrics such as VideoAlign (Liu et al., 2025b), HPSv3 (Ma et al., 2025), Audiobox Aesthetics (Tjandra et al., 2025), CLAP (Wu et al., 2023), and DeSync (Iashin et al., 2024). This motivates our dimension-wise reinforcement learning with VA-Judger, which learns the reward signal from human preference annotations instead of optimizing only narrow expert metrics.

3 METHOD

VA-Judger is designed to provide a reward signal for post-training text-to-video-audio generation models. Concretely, we first construct VAPref-10K from difficulty-aware video-audio pairs (Fig 8 and Sec 3.2). We then train VA-Judger using the three-stage pipeline (Fig 2), which combines an easy-pair cold start, hard-pair SFT using human-aligned and format-valid responses (Sec 3.3), and dimension-wise reinforcement learning (Sec 3.4).

3.1 TASK FORMULATION

We formulate video-audio reward modeling as pairwise preference prediction with structured dimension-wise reasoning. Each example consists of a text prompt p, two generated video-audio clips $x_1 = (v_1, a_1)$ and $x_2 = (v_2, a_2)$, and a preferred index $y^\star \in \{1, 2\}$. The input to the model is a fixed system prompt that specifies the evaluation rubric, followed by a user message containing the text prompt and the two videos with audio. This input format forces our model to jointly evaluate the combined text-video-audio performance, not a separate audio or video score. The model then produces an interpretable chain-of-thought comparison (Wei et al., 2022; Wang et al., 2026). For each clip, it assigns a 1–10 score with a brief justification on five dimensions, namely (A) prompt alignment, (B) video-audio consistency, (C) audio quality, (D) video quality, and (E) completeness and coherence, then aggregates these into a total score and emits a final preference in the format `<answer>video k is better</answer>`.

3.2 HUMAN-PREFERENCE DATA CONSTRUCTION

VAPref-10K contains human-annotated hard comparisons and out-of-domain comparisons. The 4,390 easy cross-model pairs form a separate cold-start set and are not part of VAPref-10K. To make VAPref-10K comprehensive, it covers a diverse range of text prompts across various scenarios. Recent benchmarks combine taxonomy-driven synthetic prompts with curated or video-derived examples (Hua et al., 2026; Cao et al., 2025; Zhou et al., 2026). In contrast, VAPref-10K uses

10,173 real video-audio clips sourced from YouTube, Bilibili, films, and TV series as the primary source for prompt construction. This preserves naturally occurring events and cross-modal interactions, yielding more complex and realistic prompts. We trim the collected videos into 5-10-second segments and organize them into six categories.

As shown in Fig 8, source videos are collected from the web and organized into six broad categories: multi-person dialogue, voice-over narration, music, cinematic ambience, environmental events, and talks. Source video distribution is shown in Fig 10. We then caption these source videos with Qwen3.5-Omni (Team, 2026). Although these captions capture rich multimodal detail, their temporally structured format produces dense timestamps and fragmented event descriptions in our data, making them unsuitable as direct inputs to video-audio generation models. Thus, we use a Qwen (Yang et al., 2025) to remove the timestamp-heavy structure and rewrite each caption into a coherent, generation-ready *text prompt* while preserving the visual content, audio events, speech, and cross-modal relations. This produces 10,083 rewritten prompt candidates. We then remove candidates that do not satisfy the generation and data-quality requirements. The remaining approximately 9K prompts are used to generate candidate clips and construct preference pairs.

Pair construction is deliberately difficulty-aware. If all training pairs are near ties, the base omni-model may struggle to learn the rubric format and produce unstable judgments. If all pairs are easy, it may overfit to superficial model-specific cues and fail on realistic near-quality comparisons. Thus, we build two complementary splits. The easy pool contains 4,390 pairs with clear quality gaps. Across all six categories, we compare Ovi (Low et al., 2025) with LTX-2 (HaCohen et al., 2026). In our preliminary evaluation, LTX-2 is substantially better than Ovi, so the preferred output is usually apparent from overall quality. For multi-person dialogue and voice-over, we also compare Ovi with daVinci-MagiHuman, which is strong in human-centric speech generation (Chern et al., 2026). The raw hard pool contains 9,146 pairs with much smaller quality gaps. It compares two samples generated from the same prompt by the same model using different random seeds. Distinguishing these pairs requires close inspection of subtle audio-visual differences. More details are in Sec 3.3.

3.3 Curriculum Supervised Fine-Tuning

VA-Judger produces structured chain-of-thought judgments rather than opaque scalar scores. A base omni-model lacks this capability, as it is not trained to compare clips under a multi-dimensional rubric, and binary preference supervision alone cannot induce such structured reasoning. Thus, we use supervised fine-tuning to establish the output format. We initialize VA-Judger from Qwen3-Omni (Xu et al., 2025) and train it on complete rubric responses that include dimension-wise scores, explanations, total scores, and a final `<answer>` tag.

3.3.1 Stage 1: Easy Cold-start

Stage 1 aims to teach VA-Judger the required response format and lets the model learn preference discrimination from unambiguous examples. We begin with easy cross-model pairs, where large quality gaps make preferences clear and allow the model to learn comparison without relying on subtle perceptual differences. This easy-to-hard schedule follows curriculum learning, which improves optimization by presenting simpler examples before harder ones (Bengio et al., 2009). We use Gemini 3.1 Pro (Google DeepMind, 2026a) to generate structured reasoning responses, as it agrees with human annotators on 80% of final preferences in a 200-pair pilot study, avoiding costly large-scale annotation at this stage. Training on these responses forces the model to follow the five-dimensional rubric, aggregate scores, and produce valid `<answer>` outputs, allowing the subsequent hard stage to focus on fine-grained preference alignment rather than output formatting.

3.3.2 Stage 2: Hard Preference Alignment

Hard pairs contain clips of similar overall quality, so their preference depends on subtle differences that must be identified through close inspection of the audio and video. This ambiguity makes the supervision generated by Gemini in Stage 1 unreliable. In a separate pilot study of 200 hard pairs, Gemini 3.1 Pro agrees with human judgments on only 60% of the pairs. As the reward model is intended to represent subjective human preference (Christiano et al., 2017; Ouyang et al., 2022), we replace Gemini's final choices with human-provided preference labels. Annotators label all 9,146

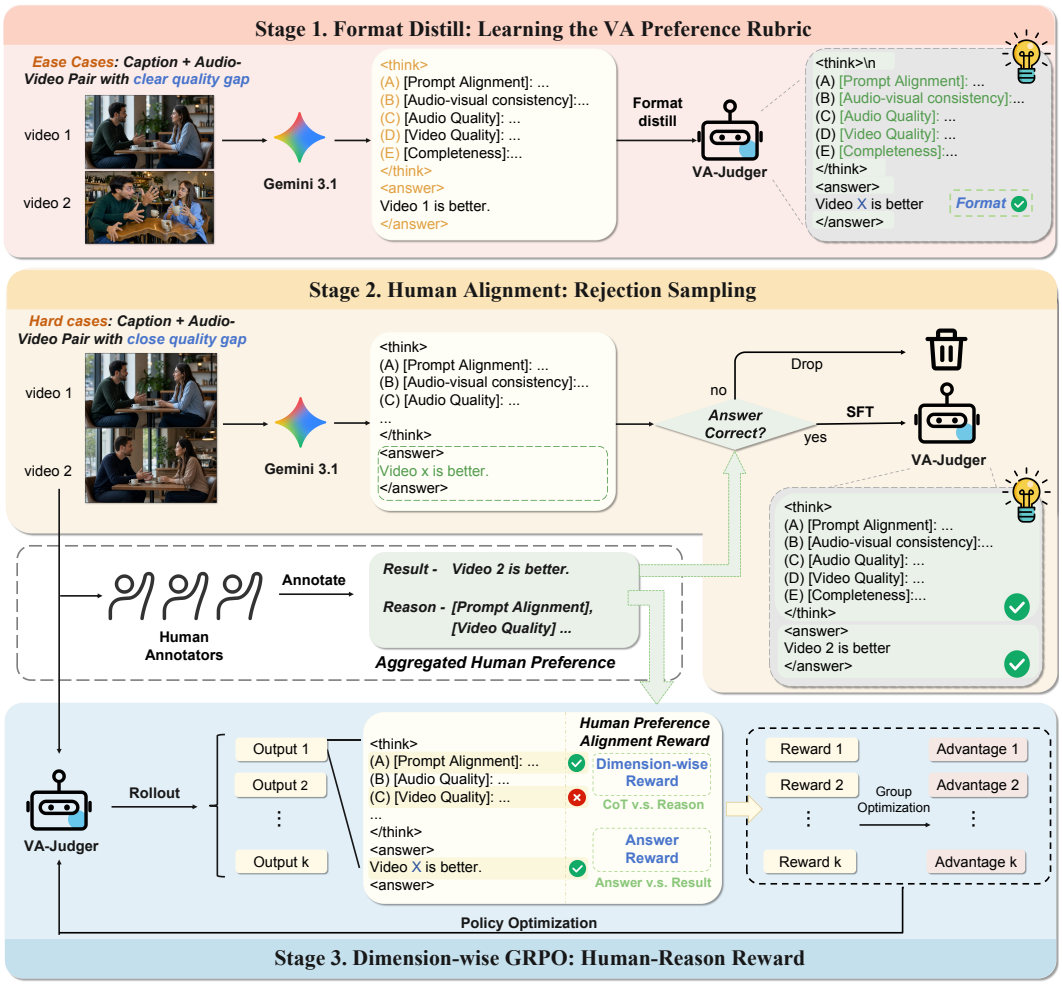


Figure 2: Overview of the VA-Judger training pipeline. Stage 1 cold-starts the model on easy preference pairs. Stage 2 aligns the model on harder pairs using responses filtered by final-choice agreement, dimension-score agreement, and format validation. Stage 3 performs GRPO using answer-level and dimension-level rewards grounded in human reasons.

raw hard pairs by selecting the preferred clip and identifying the rubric dimensions that support their decision. Annotator training and quality control are described in Appendix C.4.

Furthermore, human labels alone do not provide the complete structured responses required for SFT. We therefore use Gemini to generate rubric responses and retain only those that match the human preference, pass the regular-expression format parser, and assign the human-preferred clip a strictly higher score on every human-selected supporting dimension, that is, $s_{y^\star,d} > s_{3-y^\star,d}$ for all $d \in \mathcal{D}_{\mathrm{h}}$. This yields 4,436 hard-pair responses whose final choices and dimension-level score orderings agree with the human annotations. Training on these structured responses exposes VA-Judger to hard pairs with similar overall quality while preserving the response format learned in Stage 1.

3.4 Dimension-wise Reinforcement Learning

The two SFT stages teach the model to imitate structured rubric responses, but imitation alone has a fundamental limitation. The likelihood objective treats every token in the target response equally and does not distinguish whether the model's own generated judgments actually agree with human preferences. A model can achieve low SFT loss while still producing dimension scores that contradict its final answer, or choosing the wrong clip with confident but internally consistent reasoning. We therefore use an objective that rewards sampled judgments according to their agreement with human labels.

Thus, we propose *Dimension-wise GRPO*, which extends GRPO (Shao et al., 2024) with a human-grounded verifiable reward. A key observation is that a single binary reward (correct or incorrect final answer) is too sparse for structured judgments. The model may guess the right winner while assigning dimension scores that do not support its decision, learning a shortcut rather than dimension-level agreement. Our reward addresses this by jointly verifying two aspects of each response. Given input $x = (p, x_1, x_2)$ consisting of a prompt and two clips, the rollout policy samples G responses $\{o_i\}_{i=1}^{G}$. Each human annotation provides the preferred clip $y^\star$ and a set of supporting dimensions $\mathcal{D}_\mathrm{h} \subseteq \mathcal{D}$ that justify the preference (the full rubric $\mathcal{D}$ is shown in Appendix Figure 7). For response o_i, we parse the predicted preference $\hat{y}_i$ and dimension scores $s_{i,k,d}$ for clip k on dimension d. The reward combines an answer component that verifies the final preference with a dimension component that checks whether the score ordering on each human-specified dimension supports the human decision. For clarity, we define both components for a generic response o, with predicted preference $\hat{y}$ and score $s_{k,d}$ for clip k on dimension d:

$$r_{\mathrm{ans}}(o) = \mathbf{1}[\hat{y} = y^\star], \qquad r_{\mathrm{dim}}(o) = \frac{1}{|\mathcal{D}_\mathrm{h}|} \sum_{d \in \mathcal{D}_\mathrm{h}} \mathbf{1}[s_{y^\star,d} > s_{3-y^\star,d}]\,. \tag{1}$$

For each sampled response, we define the combined reward as $R_i = \frac{1}{2}\big(r_{\mathrm{ans}}(o_i) + r_{\mathrm{dim}}(o_i)\big)$. This design rewards both the final choice and score orderings that agree with the human-identified quality dimensions. Responses whose dimension scores disagree with these dimensions receive only partial reward. Annotations without any valid dimension in $\mathcal{D}$, including Other-only annotations, are excluded from Dimension-wise GRPO, ensuring that $|\mathcal{D}_\mathrm{h}| > 0$.

We optimize the policy with a group-normalized on-policy variant of GRPO. The G rewards within each group are normalized to advantages A_i. We then maximize the advantage-weighted mean log likelihood of the generated tokens:

$$A_i = \frac{R_i - \bar{R}}{\sigma_R + \epsilon}, \qquad \mathcal{J}_{\mathrm{GRPO}}(\theta) \quad = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{A_i}{|o_i|}\sum_{t=1}^{|o_i|}\log \pi_\theta(o_{i,t} \mid x, o_{i,<t})\right]. \tag{2}$$

Here, $\bar{R}$ and σ_R are the group mean and standard deviation of rewards, and ϵ stabilizes normalization. The submitted implementation uses neither PPO ratio clipping nor a reference-model KL penalty. This choice keeps the optimization objective consistent with the sampled group-relative advantages and avoids an additional reference-model copy.

4 Experiments

4.1 Experiment Setup

Reward model evaluation. VA-Judger-Bench contains three splits: 400 easy pairs, 250 in-domain pairs, and 500 out-of-domain pairs. The easy split contains clear quality gaps and tests coarse preference discrimination. The in-domain split follows the hard training distribution and tests subtle judgments on familiar generation models. The out-of-domain split uses clips from unseen closed-source generation models to test the model's generalization capability. **None of the 1,150 benchmark pairs overlaps with the training set.**

Generation model post-training. We use LTX-2 (HaCohen et al., 2026) as the policy and retain OmniNFT's reinforcement-learning framework (Zhang et al., 2026), but replace its independent expert rewards with VA-Judger. VA-Judger evaluates pairwise candidates and provides audio, video, and shared cross-modal rewards derived from its five rubric dimensions. We freeze the LTX-2 backbone and optimize only LoRA parameters (Hu et al., 2021). Full reward construction and training details are provided in Appendix F.2.

4.2 Reward Model Evaluation

We evaluate seven representative automatic metrics spanning unimodal quality, text consistency, audio-video consistency, synchronization, and aggregate quality (Liu et al., 2025b; Tjandra et al., 2025; Hessel et al., 2021; Girdhar et al., 2023; Iashin et al., 2024; Liu et al., 2026a). Each metric selects the clip with the better scalar score, with an exact tie receiving half credit. Table 1 shows

Table 1: Accuracy (%) against human pairwise preferences on 400 easy, 250 in-domain, and 500 out-of-domain pairs. Reward models report Total Acc / Parsed Acc. Overall is the micro accuracy over all 1,150 pairs.

Model	Easy	In-domain	Out-of-domain	Overall (micro)
Automatic evaluation metrics				
Video quality: VideoAlign	62.25	53.20	48.40	54.26
Audio quality: Audiobox Aesthetics	58.50	50.00	44.00	50.35
Text-video: CLIP Score	55.75	54.80	53.80	54.70
Text-audio: ImageBind T-A	58.50	52.80	51.60	54.26
Audio-video: ImageBind	61.25	55.60	51.80	55.91
Synchronization: SynchFormer	61.50	46.80	48.20	52.52
Overall: Javis Score	60.00	55.60	**55.40**	57.04
Metric ensemble: Hard voting	64.50	52.00	50.00	55.48
Metric ensemble: Z-score soft voting	**67.00**	**58.00**	52.40	**58.70**
Reward models				
Qwen3-Omni Captioner (No CoT)	57.00 / 57.00	54.00 / 54.00	56.20 / 56.20	56.00 / 56.00
Qwen3-Omni Captioner (CoT)	52.50 / 57.69	49.60 / 54.87	46.60 / 56.97	49.30 / 56.76
Qwen3-Omni Instruct NoCoT	60.75 / 60.75	54.00 / 54.00	54.60 / 54.60	56.61 / 56.61
Qwen3-Omni Instruct CoT	63.25 / 64.71	54.80 / 55.92	55.00 / 55.33	57.83 / 58.69
VA-Judger (Ours)	**76.25** / 76.25	**66.00** / 66.00	**63.40** / 63.40	**68.43** / 68.43

Table 2: Results on the randomly sampled 200-prompt JavisBench subset. LTX-2 + Qwen3-Omni-RL uses the untuned Qwen3-Omni as the reward model under the same post-training configuration as LTX-2 + VA-Judger (Appendix F.2). Best results are highlighted in green and second-best results are underlined. Higher is better unless marked with ↓.

	Video Quality		Audiobox Aesthetics					Cross-Modal Alignment		
Model	VQ ↑	MQ ↑	AQ ↑	AB-CE ↑	AB-CU ↑	AB-PC ↑	AB-PQ ↑	TV-Align ↑	TA-Align ↑	ViCLIP ↑
LTX-2	2.248	0.697	4.767	3.957	5.904	3.041	6.166	0.303	0.105	0.232
LTX-2 + Qwen3-Omni-RL	2.908	0.635	4.900	4.214	6.200	2.923	6.398	0.272	0.143	0.224
LTX-2 + OmniNFT	3.727	0.947	5.399	4.745	6.797	3.150	6.905	0.279	0.133	0.219
LTX-2 + VA-Judger	3.942	1.183	5.610	5.136	6.766	3.606	6.932	0.310	0.180	0.242

	Text Consistency				AV Consistency and Synchrony					
Model	TV-IB ↑	TA-IB ↑	CLIP ↑	CLAP ↑	AV-IB ↑	CAVP ↑	AV-Align ↑	AVHScore ↑	DeSync ↓	JavisScore ↑
LTX-2	0.355	0.102	0.303	0.304	0.091	0.786	0.192	0.091	0.430	0.074
LTX-2 + Qwen3-Omni-RL	0.370	0.140	0.317	0.354	0.142	0.799	0.200	0.142	0.335	0.122
LTX-2 + OmniNFT	0.336	0.134	0.298	0.394	0.154	0.800	0.208	0.146	0.226	0.122
LTX-2 + VA-Judger	0.376	0.179	0.326	0.425	0.265	0.799	0.228	0.261	0.592	0.230

limited agreement with human preference: overall micro accuracy ranges from 50.35% to 57.04%, while VideoAlign and Audiobox Aesthetics drop to 48.40% and 44.00% out of domain.

Combining all seven metrics does not close the gap. Label-free Z-score soft voting improves overall accuracy to 58.70%, but reaches only 52.40% out of domain, while hard voting reaches 50.00%. Separate metrics observe only partial aspects of the joint experience, and embedding similarity or temporal offset can overlook local defects and semantic mismatch. Appendix G.1 gives the ensemble definitions, and Appendix Tables 9–14 define the metrics and report the complete results.

VA-Judger reaches 68.43% overall Total Acc, improving over Qwen3-Omni Instruct with CoT by 10.60 percentage points. Total Acc counts invalid final choices as incorrect, while Parsed Acc uses only valid responses. Appendix Tables 4–7 report curriculum and reward ablations, input-order robustness, parsing coverage, and micro and macro accuracy.

4.3 Post-training LTX-2 with VA-Judger

We evaluate all models on 200 uniformly sampled JavisBench prompts (Liu et al., 2026a). Appendix Table 3 reports the subset distribution. Higher is better except for DeSync.

Quantitative results. VA-Judger ranks first on 17 of the 20 reported metrics and receives the highest human preference. Relative to LTX-2, visual, motion, and audio quality rise from 2.248, 0.697, and 4.767 to 3.942, 1.183, and 5.610, while JavisScore and AVHScore rise from 0.074 and 0.091 to 0.230 and 0.261. OmniNFT obtains the lowest DeSync value, but DeSync is explicitly included in its post-training reward. This advantage likely reflects direct optimization of that metric and is

consistent with metric-specific reward hacking rather than uniformly better video-audio quality. The broad improvement of VA-Judger is consistent with learning a unified reward from holistic human comparisons rather than optimizing separate expert metrics (Zhang et al., 2026).

Qualitative results. Figure 3 shows that the VA-Judger post-trained model better follows the requested temporal actions. In the rabbit example, the baselines omit either the requested freeze-and-dart motion or the intended temporal sequence, whereas our result preserves the brief pause followed by rapid movement. In the spacecraft example, our result maintains the requested three-leg structure and depicts the synchronized compression, while the baselines exhibit incorrect support geometry or fail to show the specified motion. Appendix A provides additional qualitative comparisons.

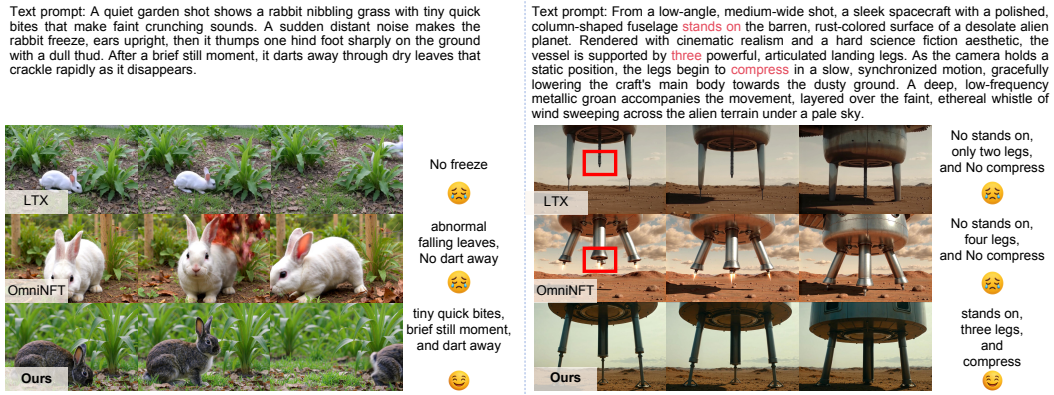


Figure 3: Qualitative comparisons of LTX-2, OmniNFT, and our VA-Judger post-trained model. VA-Judger better follows the requested temporal actions in both examples.

Human evaluation. Across 4,000 selections from 200 three-way comparisons judged by 20 adults, VA-Judger receives 62.30% of preferences, compared with 27.63% for OmniNFT and 10.08% for LTX-2, as shown in Figure 4. Appendix Figure 11 shows the study interface, and Appendix H provides the full protocol.

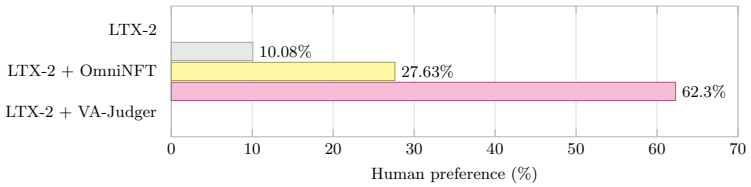


Figure 4: Human preference rates for 200 prompts, aggregated over 4,000 three-way selections by 20 participants, among LTX-2, LTX-2 post-trained with OmniNFT, and LTX-2 post-trained with VA-Judger as the learned reward model.

5 Conclusion

We introduced VA-Judger, a human-aligned omni-reward model for joint video-audio generation, together with VAPref-10K and VA-Judger-Bench. VA-Judger combines easy-to-hard supervised fine-tuning with human-matched hard examples and dimension-wise reinforcement learning to learn final preferences and structured dimension-level comparisons. Unlike separate expert metrics that optimize isolated properties, its unified reward jointly evaluates visual quality, audio quality, semantic fidelity, and cross-modal coherence. On VA-Judger-Bench, it improves preference prediction over existing omni-models and automatic metrics across both in-domain and out-of-domain evaluations. As a post-training reward for LTX-2, it leads on 17 of 20 automatic metrics and receives 62.30% human preference. Together, these results establish holistic human preference supervision as an effective foundation for improving joint video-audio generation.

ETHICS STATEMENT

The authors have read and adhere to the ICLR Code of Ethics. This work involves human preference annotation for VAPref-10K and a user study with 20 adult participants. Only aggregate preference results are reported, and the paper does not disclose participant-identifying information. VA-Judger is developed to evaluate and improve generated video-audio content. We encourage responsible use consistent with applicable laws, licenses, and platform policies.

REPRODUCIBILITY STATEMENT

We provide the task formulation and training objectives in Sections 3–3.4. Section 3.2 and Appendix C describe data construction, source permissions, pair formation, source distributions, and human annotation. Section 4.1 defines the benchmark splits and evaluation protocol. Appendix E specifies the controlled ablations, parsing rules, tie handling, and micro and macro aggregation, while Appendix F.1 and Appendix F.2 report the reward-model and generation model post-training configurations. Appendix G provides the complete metric results. The anonymous supplementary archive contains reward-model SFT, dimension-wise GRPO, and evaluation code, together with generation model post-training and inference code, prompts, configuration files, dependency specifications, access instructions, and generated video examples.

AI USE STATEMENT

In this work, we used generative AI tools to edit manuscript text for clarity, grammar, and readability, to help structure sections and tables, and to retrieve potentially relevant literature. These tools also provided feedback on experimental and ablation design, assisted with LaTeX formatting and aggregate-metric calculations, and suggested preliminary interpretations of reported trends. The authors independently checked the cited papers and bibliographic information, verified the calculations against the underlying results, and made all final methodological and interpretive decisions. Generative AI tools were also used directly in data construction. As detailed in Sections 3.2 and 3.3, Qwen3.5-Omni captioned source video-audio clips, Qwen rewrote the captions into generation-ready prompts, and Gemini 3.1 Pro generated structured rubric responses for the easy-pair cold start and candidate hard-pair responses. The hard-pair responses were retained only when their final choices and human-selected dimension score orderings matched the annotations and their output formats passed regex-based validation. Video-audio generation models produced the candidate clips used to form preference comparisons. The authors reviewed all AI-assisted text and research artifacts, filtered or verified the data-construction outputs according to the procedures described in the paper, and take responsibility for the final manuscript, claims, methods, results, and released artifacts.

REFERENCES

Alibaba Cloud. Wan 2.6 models. https://www.alibabacloud.com/help/en/model-studio/video-generate-edit-model, 2026a. Accessed: 2026-09-08.

Alibaba Cloud. Wan 2.7. https://www.alibabacloud.com/help/en/model-studio/video-generate-edit-model, 2026b. Accessed: 2026-09-08.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, volume 2026, pp. 118456–118470, 2026.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jiahao Wang, Jialu Chen, Miao Deng, Yubin Guo, Chenxi Liao, et al. T2av-compass: Towards unified evaluation for text-to-audio-video generation. *arXiv preprint arXiv:2512.21094*, 2025.

Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Google DeepMind. Introducing Veo 3.1 and advanced capabilities in Flow. `https://blog.google/innovation-and-ai/products/veo-updates-flow/`, 2025. Accessed: 2026-06-25.

Google DeepMind. Gemini 3.1 pro. `https://deepmind.google/models/gemini/pro/`, 2026a. Accessed: 2026-06-23.

Google DeepMind. Gemini-omni. `https://deepmind.google/models/gemini-omni/`, 2026b. Accessed: 2026-09-08.

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.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Kuaishou Technology. Kling ai launches video 2.6 model with "simultaneous audio-visual generation" capability, redefining ai video creation workflow. Official release, December 2025.

Haoran Li, Fredreic Li, Shichen Ma, Jie Huang, Yijun Liu, Jiaqi Shi, and Yanwen Ma. Joyai-echo: Pushing the frontier of long audio-visual generation. 2026.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal LLMs to follow zero-shot demonstrative instructions. In *International Conference on Learning Representations*, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2025b.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Scofield Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, volume 2026, pp. 139160–139194, 2026a.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Scofield Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. In *International Conference on Learning Representations*, volume 2026, pp. 150592–150618, 2026b.

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

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*, 2021.

OpenAI. Sora 2 is here. `https://openai.com/index/sora-2/`, 2025. Accessed: 2026-06-25.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Kaihang Pan, Wang Lin, Zhongqi Yue, Tenglong Ao, Liyu Jia, Wei Zhao, Juncheng Li, Siliang Tang, and Hanwang Zhang. Generative multimodal pretraining with discrete diffusion timestep tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26136–26146, 2025.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, Yue Wu, Liefeng Bo, Siliang Tang, and Zhao Zhong. Omniweaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Haiyi Qiu, Kaihang Pan, Jiacheng Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. Spatialfusion: Endowing unified image generation with intrinsic 3d geometric awareness. *arXiv preprint arXiv:2604.26341*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 886–890. IEEE, 2022.

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.

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.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

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.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024a.

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024b.

Yibin Wang, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, Jiaqi Wang, et al. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *Advances in neural information processing systems*, 38:159130–159157, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

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.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 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*, 2025.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6463–6474. IEEE, 2024.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6547–6563, 2025.

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*, 2026.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8018–8027. IEEE, 2024.

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

Ziwei Zhou, Zeyuan Lai, Rui Wang, Yifan Yang, 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. In *Forty-third International Conference on Machine Learning*, 2026.

APPENDIX

A ADDITIONAL POST-TRAINING QUALITATIVE RESULTS

Figures 5 and 6 provide additional comparisons of LTX-2, OmniNFT, Qwen3-Omni-RL, and our VA-Judger post-trained model on diverse prompts. Qwen3-Omni-RL corresponds to the LTX-2 + Qwen3-Omni-RL row in Table 2, and its training configuration is described in Appendix F.2.

Text prompt: A brown and white cow, with large, gentle eyes and a slightly bewildered expression, stands on a sun-drenched pasture. It is awkwardly holding an acoustic guitar, its hooves fumbling with the strings. The cow attempts to strum a chord, but instead of music, a series of loud, resonant "Moo!" sounds emanate from its throat, clearly not aligning with the intended melody. The camera employs a medium shot to capture the cow's comical struggle, with a slight zoom in to emphasize its earnest but unsuccessful efforts. The lighting is bright and natural, characteristic of a clear day, with a slightly desaturated color palette to give a touch of whimsicality. The overall style leans towards gentle, humorous illustration, with a focus on the visual absurdity of the scene. The dominant sound is the repeated, drawn-out "Moo!", punctuated by the faint, comical rustle of hay。

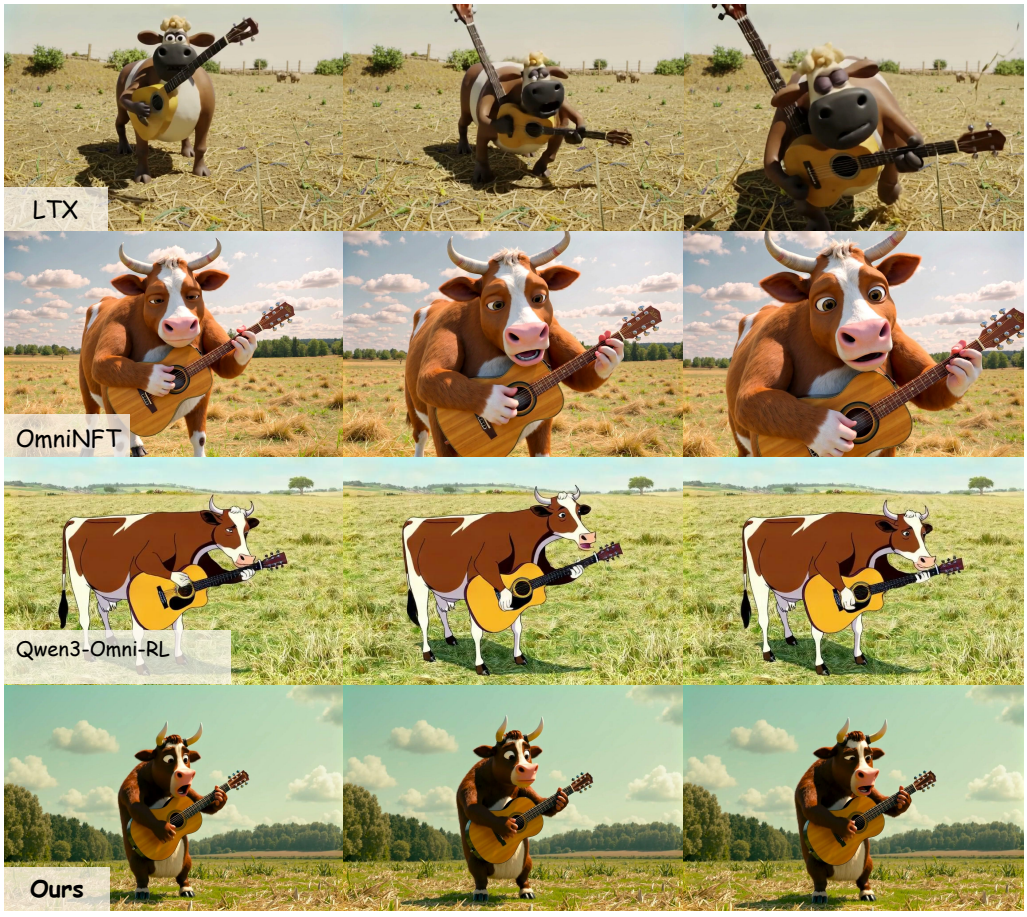


Figure 5: Additional qualitative comparison for a prompt describing a cow playing an acoustic guitar. Each row shows three frames generated by the corresponding method.

Text prompt: From a slight high-angle, an extreme close-up captures a pair of steady hands guiding a large chef's knife as it slices a ripe tomato on a wooden cutting board. The scene is shot in a high-saturation photographic style, with bright, even light enhancing the dominant reds of the vegetables and the warm tones of the wood grain. Scattered across the board are diced red bell pepper pieces, and whole cherry tomatoes sit softly blurred in the background. In the top-left corner, a circular logo displays the text "Men With The Pot" with three stylized pine trees above it. The hands continue their work, gliding the knife back and forth in a steady, rhythmic motion. With each pass of the blade, small, neatly cut pieces of tomato fall away to join the other diced vegetables on the board. The shallow focus keeps the sharp edge of the knife and the central tomato in high detail, while the background tomatoes and the wider surface of the board remain out of focus, emphasizing the texture and motion of the chopping. The audio shows soft birdsong filtering through the air, creating a gentle and natural ambient rhythm for the scene.

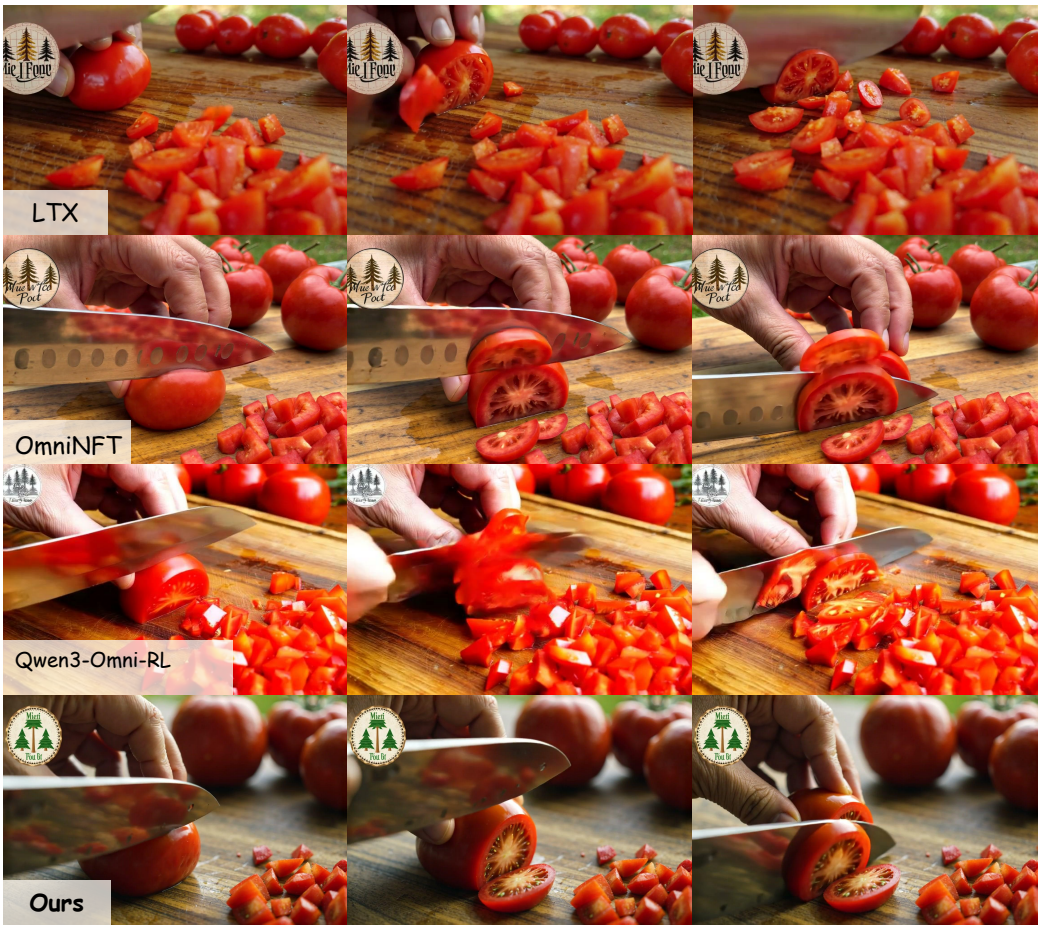


Figure 6: Additional qualitative comparison for a prompt describing a tomato-cutting scene. Each row shows three frames generated by the corresponding method.

B QUALITATIVE EXAMPLE OF A VA-JUDGER JUDGMENT

System Prompt:

Given a caption and two generated audio/video files based on this caption, evaluate which video is better.\n\ncompare the two videos on the following fixed dimensions:\n\n
(A) Better match to the prompt (characters, content, or speech)\n
(B) Better audio-visual consistency (e.g., lip sync, sound matches the image)\n
(C) Higher audio quality\n
(D) Higher video quality\n
(E) More complete and coherent content\n\n
For each dimension, provide:\n- A score between 1-10 for both videos (e.g., video 1: 8/10, video 2: 6/10)\n- A concise rationale explaining the score\n\nThen compute the total score for each video by summing all dimension scores.\n\n

Use the following output format exactly:\n\n
<think>\n
(A) Better match to the prompt: video 1 (.../10) - ...; video 2 (.../10) - ...\n
(B) Audio-visual consistency: video 1 (.../10) - ...; video 2 (.../10) - ...\n
(C) Audio quality: video 1 (.../10) - ...; video 2 (.../10) - ...\n
(D) Video quality: video 1 (.../10) - ...; video 2 (.../10) - ...\n
(E) Completeness and coherence: video 1 (.../10) - ...; video 2 (.../10) - ...\n\n
Total score:\nvideo 1: ... = ...\nvideo 2: ... = ...\n
</think>\n\n
Finally, in the <answer> tag, output exactly one of the following strings based on the total scores:\nvideo 1 is better\nor\nvideo 2 is better\n\nNo additional text or quotation marks are allowed in the <answer> section. Write the answer tag on one line, for example:\n<answer>video 1 is better</answer>"

Video Text Prompt:

Cinematic medium shot, locked-off and eye-level, opens inside a warmly lit café or tea-room bathed in soft, diffuse daylight that blends with cozy interior tones. The color palette is dominated by earthy browns, muted beiges, and deep blacks, punctuated by the vivid yellow of neckerchiefs and green foliage. Two teenage girls in matching black sailor-style school uniforms sit side-by-side at a dark wooden table, their navy collars edged in white piping and bright yellow neckerchiefs tied in loose bows, small red name badges pinned to their left chests. In front of them stand two tall, curved glasses filled with a cloudy pale drink, each garnished with a lemon slice on the rim, a single cherry at the bottom, and a white straw, while a smaller empty tumbler rests near the girl on the left. As a faint café room tone hums in the air, an off-screen adult woman's calm, confident Japanese voice begins to speak, her standard Tokyo accent crisp and mid-to-high pitched: \"お二人は転校早々、十人もの不良どもをやっつけたと聞きました。\" The girl on the right turns her head slightly toward the unseen speaker, listening attentively with her eyes lowered and lips closed, while the girl on the left keeps her gaze forward, brows faintly knit as she absorbs the words, her posture steady and respectful.\n\nThe woman's voice continues, shifting to an encouraging tone with a slight smile audible in her diction as she delivers her next line in Japanese: \"あんたらなら、やつらを倒せると思って。\" The girl on the left lifts her chin, leans forward a fraction, and parts her lips, signaling her intent to respond, her expression tightening with quiet determination. Beside her, the girl on the right remains composed, watching her companion with steady eyes. The camera holds perfectly still, capturing the subtle micro-expressions and restrained body language that convey a shared sense of resolve against the softly shadowed backdrop. The scene lingers on this poised anticipation, the ambient café sounds fading into the weight of the unspoken challenge hanging in the warm, layered space.



Video 1



Video 2

VA-Judger Response:

<think>\n
(A) Better match to the prompt:
video 1 (7/10) - Captures the cinematic style, setting, and specific character actions (left girl leaning forward, right girl watching) well, though it uses a real-name actress for the voiceover instead of an adult woman and misses the exact lemon placement;
video 2 (4/10) - Misses the lemon slices entirely, features a tie instead of a bow, and fails to depict the requested character actions, acting more as a static background.\n
(B) Audio-visual consistency:
video 1 (8/10) - The off-screen voiceover fits the ambient cafe setting perfectly, and the characters' subtle reactions align with the narrative pacing;
video 2 (5/10) - The voiceover feels disconnected from the environment, and the lack of character reaction makes the audio feel like a detached narration rather than dialogue directed at them.\n
(C) Audio quality:
video 1 (9/10) - Clear, professional-sounding voiceover with natural ambient noise;
video 2 (3/10) - The voiceover is heavily distorted, garbled, and difficult to understand.\n
(D) Video quality:
video 1 (9/10) - High-quality, realistic visuals with natural lighting and textures;
video 2 (6/10) - Basic 2D animation with flat colors and static shots that lack visual interest.\n
(E) Completeness and coherence:
video 1 (8/10) - Delivers a cohesive and engaging scene that follows the prompt's progression, despite the voiceover mismatch;
video 2 (4/10) - Disjointed and incomplete, hindered significantly by the poor audio and lack of dynamic visual storytelling.\n\n
Total score:\n
video 1: 7 + 8 + 9 + 9 + 8 = 41\n
video 2: 4 + 5 + 3 + 6 + 4 = 22\n\n
</think>\n\n
<answer>video 1 is better</answer>

Figure 7: A complete VA-Judger input and response. The system prompt defines dimensions A to E, and the response compares both clips on every dimension before producing the final preference.

C VAPREF-10K CONSTRUCTION AND FINAL PAIR DISTRIBUTION

C.1 CONSTRUCTION PIPELINE

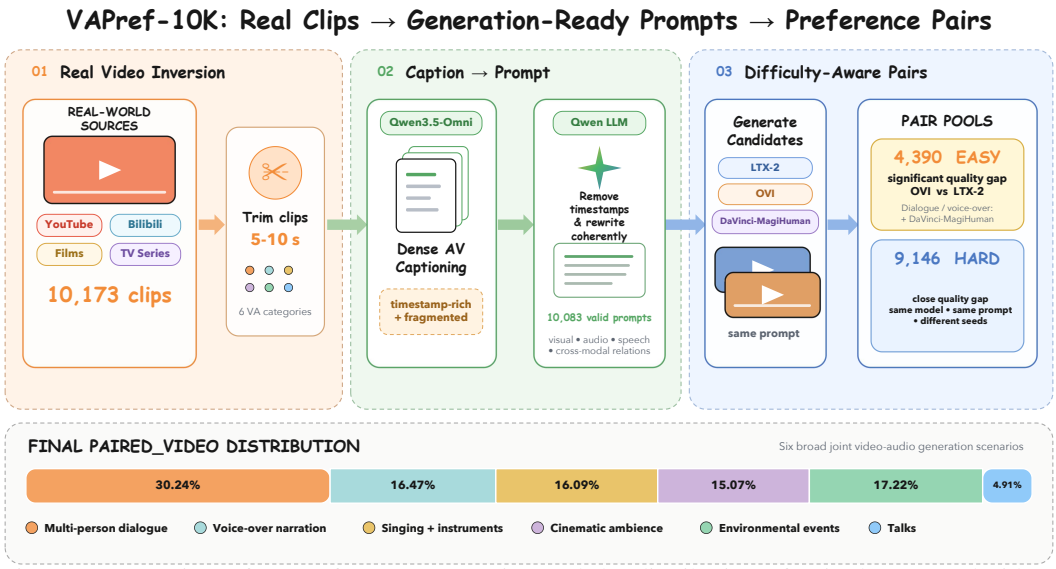


Figure 8: Overview of VAPref-10K construction and the distribution of paired generated videos retained after filtering and pair construction. Real video-audio clips are captioned by Qwen3.5-Omni, rewritten into text prompts by an LLM, and passed to LTX-2, Ovi, and daVinci-MagiHuman to generate candidate clips. A prompt may yield multiple generated clips and therefore contribute more than one within-model comparison. The diagram reports 4,390 easy cold-start pairs and 9,146 human-annotated raw hard pairs. The separately constructed 1,150-pair benchmark is not included in the diagram.

C.2 SOURCE VIDEO PERMISSIONS

The source videos used for prompt construction were collected primarily from Bilibili and YouTube. Before including a collected video in the dataset construction and annotation pipeline, we contacted the corresponding uploader through a private message and requested permission to use the video for non-commercial academic research. We retained only videos for which the uploader provided affirmative consent.

C.3 HUMAN ANNOTATION INTERFACE

For each of the 9,146 raw hard pairs, the annotation platform presents two video-audio clips generated from the same prompt. Annotators select the preferred clip and then identify one or more dimensions that support their decision. These dimensions correspond to prompt alignment, video-audio consistency, audio quality, video quality, and completeness and coherence. An additional "Other" option allows annotators to report a reason outside the predefined rubric. When constructing the Dimension-wise GRPO training set, we exclude every annotation that selects only "Other" and contains no valid rubric dimension. Consequently, every retained example satisfies $|\mathcal{D}_{\mathrm{h}}| > 0$, so the dimension reward in Equation 1 is always well-defined.

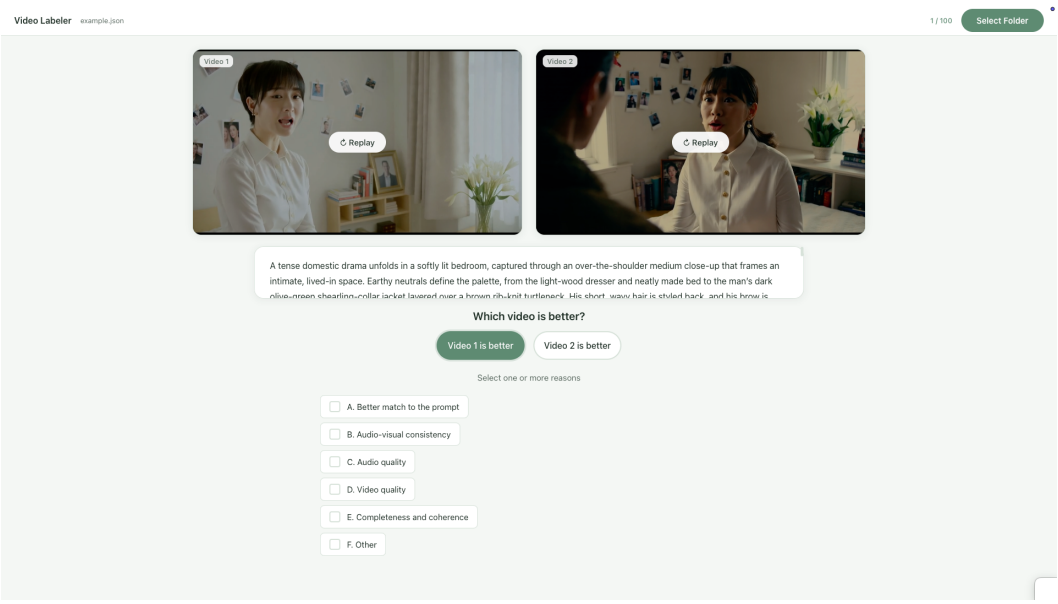


Figure 9: Interface used to collect pairwise preferences and supporting quality dimensions for VAPref-10K. Annotators can replay both clips, inspect the shared prompt, select the better clip, and mark the reasons for their choice.

C.4 Annotator Training and Quality Control

Annotator training. All human preference labels in VAPref-10K and VA-Judger-Bench are provided by 23 annotators. Before annotation, all annotators attend a joint training session of approximately 3 hours, in which we explain the meaning of each of the five rubric dimensions. They then annotate example pairs during the session, and we review their annotations in real time and provide immediate feedback. Annotators are instructed to play both clips with audio in full, to judge the complete text-video-audio experience rather than a single modality, and to select up to three rubric dimensions that support their choice. Because many clips contain speech or singing, each clip is assigned only to annotators proficient in its language.

Qualification test. Reference labels and spot-check judgments are provided by an expert panel of three members. After training, each annotator labels a qualification set of 100 easy pairs and 100 hard pairs whose reference labels are determined by the expert panel. We measure preference agreement separately for the two difficulty levels. Only annotators who reach at least 90% agreement on easy pairs and at least 80% agreement on hard pairs proceed to formal annotation. The lower threshold for hard pairs reflects that near-quality comparisons are inherently more ambiguous.

We additionally measure dimension overlap with the expert annotations. Let $\mathcal{D}_i^{\text{ann}}$ and $\mathcal{D}_i^{\text{expert}}$ denote the dimension sets selected by the annotator and expert panel. We define

$$\text{OverlapAcc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\mathcal{D}_i^{\text{ann}} \cap \mathcal{D}_i^{\text{expert}} \neq \emptyset\right]. \tag{3}$$

A dimension annotation is consistent when the two sets share at least one dimension. Only annotators whose dimension-overlap rate exceeds 90% proceed to formal annotation.

Spot checks during annotation. During formal annotation, we randomly sample 20 pairs labeled by each annotator and the expert panel independently re-judges these pairs. We compute preference agreement on the sampled easy and hard pairs using the same thresholds as in the qualification test, and require the dimension-overlap rate to exceed 90%. If an annotator fails any criterion, all pairs labeled by that annotator are discarded and re-annotated. Thus, every retained label comes from an annotator who passes both preference and dimension-level quality control in the qualification test and the spot check.

Table 3: Category distributions of the full JavisBench pool and the random evaluation subset. Each cell reports the number and percentage of examples.

Dimension	Category	Full pool (10,140)	Random subset (200)
Event Scenario	Living Scenario	4,029 (39.7%)	73 (36.5%)
	Natural Scenario	2,452 (24.2%)	37 (18.5%)
	Urban Scenario	2,140 (21.1%)	66 (33.0%)
	Industrial Scenario	1,090 (10.7%)	20 (10.0%)
	Virtual Scenario	429 (4.2%)	4 (2.0%)
Visual Style	Camera Shooting	8,586 (84.7%)	174 (87.0%)
	2D Animation	915 (9.0%)	17 (8.5%)
	3D Animation	639 (6.3%)	9 (4.5%)
Sound Type	Musical Sounds	5,065 (50.0%)	73 (36.5%)
	Ambient Sounds	4,183 (41.3%)	126 (63.0%)
	Biological Sounds	2,215 (21.8%)	59 (29.5%)
	Mechanical Sounds	1,940 (19.1%)	49 (24.5%)
	Speech Sounds	1,090 (10.7%)	29 (14.5%)
Spatial Composition	Multiple Subjects	7,588 (74.8%)	167 (83.5%)
	Single Subject	2,502 (24.7%)	33 (16.5%)
	Off-screen Sound	2,165 (21.4%)	59 (29.5%)
Temporal Composition	Sequential Events	3,698 (36.5%)	46 (23.0%)
	Simultaneous Events	3,306 (32.6%)	99 (49.5%)
	Single Event	3,136 (30.9%)	55 (27.5%)

C.5 Source Video Distribution

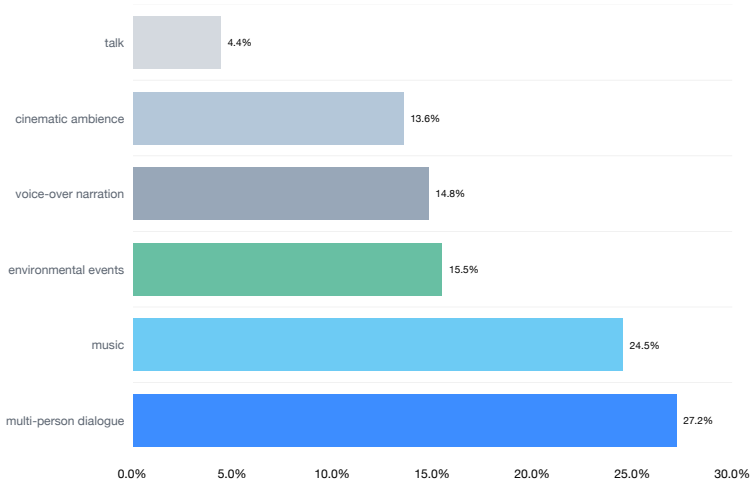


Figure 10: Distribution of the source videos across the six categories in VAPref-10K.

D JavisBench Evaluation Subset Distribution

Table 3 compares the category distribution of the full 10,140 prompt JavisBench pool with the 200 prompt subset used for generation model evaluation. The subset is sampled uniformly at random without category-based filtering or stratification. Sound Type and Spatial Composition allow multiple labels, so their percentages do not sum to 100%.

E Reward Model Ablations and Robustness

This section isolates the effects of the SFT curriculum and the two reinforcement-learning reward components. Unless stated otherwise, all variants use Qwen3-Omni-30B-A3B-Instruct, freeze the visual and audio encoders, update the language backbone, and use the same decoding and parsing procedure as the main model.

Table 4: Curriculum SFT ablations. Every entry reports Total Acc / Parsed Acc in percent. The hard-only, joint, and sequential variants use the compute-matched protocol described above.

Schedule	Easy	In-domain	Out-of-domain	Overall (micro)
Easy-only	72.00 / 72.00	59.20 / 59.20	56.20 / 56.20	62.35 / 62.35
Hard-only, compute-matched	68.00 / 68.00	63.00 / 63.00	55.80 / 55.80	61.61 / 61.61
Easy+Hard joint, compute-matched	72.50 / 72.50	63.20 / 63.20	58.80 / 58.80	64.52 / 64.52
Easy→Hard sequential	74.50 / 74.50	63.60 / 63.60	60.20 / 60.20	65.91 / 65.91

Table 5: Dimension-wise GRPO reward ablations. Every entry reports Total Acc / Parsed Acc in percent.

Reward	$(\lambda_{\text{ans}}, \lambda_{\text{dim}})$	Easy	In-domain	Out-of-domain	Overall (micro)
Answer only	$(1, 0)$	73.50 / 73.50	64.40 / 64.40	56.40 / 56.40	64.09 / 64.09
Dimension only	$(0, 1)$	72.75 / 72.75	64.00 / 64.00	60.40 / 60.40	65.48 / 65.48
Answer + dimension	$(0.75, 0.25)$	76.00 / 76.00	64.60 / 64.60	63.00 / 63.00	67.87 / 67.87
Answer + dimension	$(0.50, 0.50)$	76.25 / 76.25	66.00 / 66.00	63.40 / 63.40	68.43 / 68.43
Answer + dimension	$(0.25, 0.75)$	76.75 / 76.75	65.00 / 65.00	62.80 / 62.80	68.13 / 68.13

E.1 Curriculum Training Ablations

We adopt a compute-matched comparison. Let U_{e} and U_{h} denote the numbers of updates assigned to easy and hard data in the default sequential schedule, and let $U = U_{\text{e}} + U_{\text{h}}$. Hard-only SFT trains on hard examples for all U updates. Joint SFT trains for U updates on the union of both sets and samples easy and hard examples in the ratio $U_{\text{e}} : U_{\text{h}}$. Sequential SFT first performs U_{e} easy updates and then U_{h} hard updates. Joint and sequential SFT therefore use identical data sources, sampling proportions, and total compute, differing only in ordering. Hard-only SFT uses the same compute but intentionally excludes easy data. Easy-only SFT is retained as a diagnostic intermediate checkpoint and is not used as the compute-matched curriculum comparison.

Hard-only SFT reaches 61.61% overall accuracy and underperforms the 62.35% easy-only checkpoint in this controlled run. Joint training improves the overall result to 64.52%, while the sequential easy-to-hard schedule achieves 65.91%. Because the joint and sequential variants use the same data proportions and update budget, the 1.39-point gain is consistent with a benefit from curriculum ordering.

The default sequential schedule uses 400 optimizer updates for easy SFT followed by 1,000 updates for hard SFT. The corresponding training budgets and checkpoints are summarized in Appendix F.1.

E.2 Dimension-Wise Reward Ablations

We write the GRPO reward as

$$R(o) = \lambda_{\text{ans}} r_{\text{ans}}(o) + \lambda_{\text{dim}} r_{\text{dim}}(o), \qquad \lambda_{\text{ans}} + \lambda_{\text{dim}} = 1. \tag{4}$$

All variants start from the same easy-to-hard SFT checkpoint and keep the data, sampled responses, optimizer updates, and decoding settings fixed. Answer-only and dimension-only rewards isolate the two supervision sources. The remaining variants test the sensitivity to their relative weights.

Dimension-only training improves overall accuracy from 64.09% to 65.48% relative to answer-only training, indicating that the structured dimension signal is useful even without the final-answer reward. Combining both components yields higher overall accuracy than either component alone. Equal weighting obtains the best overall accuracy of 68.43%, while the nearby 3:1 and 1:3 settings reach 67.87% and 68.13%, respectively. The three mixed-reward settings remain within 0.56 percentage points of one another.

E.3 Input-Order Robustness

To measure position bias, we evaluate every comparison in its original order (x_1, x_2) and its swapped order (x_2, x_1), mapping the swapped prediction back to the original indexing. We report both directional accuracies and their arithmetic mean (order-averaged accuracy). This aggregate comparison measures sensitivity of accuracy to presentation order; computing pairwise order consistency would additionally require retaining the two mapped predictions for every example.

Table 6: A/B order robustness of the final VA-Judger. Accuracy columns report Total Acc / Parsed Acc in percent. Order-averaged accuracy is the arithmetic mean of the original- and swapped-order accuracies.

Evaluation order	Easy	In-domain	Out-of-domain	Overall (micro)
Original	76.25 / 76.25	66.00 / 66.00	63.40 / 63.40	68.43 / 68.43
Swapped	75.75 / 75.75	65.80 / 65.80	63.80 / 63.80	68.39 / 68.39
Order-averaged	76.00 / 76.00	65.90 / 65.90	63.60 / 63.60	68.41 / 68.41

Table 7: Micro and macro accuracy in percent. Reward models report Total Acc / Parsed Acc. Automatic ensembles report a single pairwise accuracy.

Method	Micro	Macro
Javis Score	57.04	57.00
Hard voting	55.48	55.50
Z-score soft voting	58.70	59.13
Qwen3-Omni Captioner (No CoT)	56.00 / 56.00	55.73 / 55.73
Qwen3-Omni Captioner (CoT)	49.30 / 56.76	49.57 / 56.51
Qwen3-Omni Instruct NoCoT	56.61 / 56.61	56.45 / 56.45
Qwen3-Omni Instruct CoT	57.83 / 58.69	57.68 / 58.65
Easy-only SFT	62.35 / 62.35	62.47 / 62.47
Hard-only SFT	61.61 / 61.61	62.27 / 62.27
Easy+Hard joint SFT	64.52 / 64.52	64.83 / 64.83
Easy→Hard SFT	65.91 / 65.91	66.10 / 66.10
GRPO, answer only (1:0)	64.09 / 64.09	64.77 / 64.77
GRPO, dimension only (0:1)	65.48 / 65.48	65.72 / 65.72
GRPO, answer+dimension (3:1)	67.87 / 67.87	67.87 / 67.87
GRPO, answer+dimension (1:3)	68.13 / 68.13	68.18 / 68.18
VA-Judger	68.43 / 68.43	68.55 / 68.55

Swapping the two candidates changes overall accuracy from 68.43% to 68.39%, and their order-averaged accuracy is 68.41%. The 0.04-point directional difference indicates that the aggregate result is stable to A/B presentation order.

E.4 Tie Handling, Parsing, and Macro Accuracy

All automatic metrics are evaluated on the same 400 easy, 250 in-domain, and 500 out-of-domain pairs. For a scalar metric, an exact score tie receives half credit. In hard voting, a tied metric abstains, and a tie in the final ensemble vote receives half credit. Soft voting likewise assigns half credit to an exact tie in the fused score. Reward-model outputs must contain a valid choice in the `<answer>` tag. Total Acc uses all examples as the denominator and counts an unparseable output as incorrect. Parsed Acc uses only valid outputs. Consequently, the parse rate can be recovered as Total Acc divided by Parsed Acc when both are expressed as proportions.

Table 7 distinguishes micro accuracy from the macro average that gives equal weight to the three splits. For automatic metrics and Total Acc, micro accuracy is weighted by the benchmark split sizes. For Parsed Acc, it is pooled over valid outputs. Let c_s and p_s denote the numbers of correct and parsed outputs in split s:

$$\begin{aligned} \mathrm{Acc}_{\mathrm{micro}}^{\mathrm{total}} &= \frac{400\,\mathrm{Acc}_{\mathrm{easy}} + 250\,\mathrm{Acc}_{\mathrm{in}} + 500\,\mathrm{Acc}_{\mathrm{out}}}{1{,}150}, \\ \mathrm{Acc}_{\mathrm{micro}}^{\mathrm{parsed}} &= \frac{\sum_s c_s}{\sum_s p_s}, \\ \mathrm{Acc}_{\mathrm{macro}} &= \frac{1}{3}\left(\mathrm{Acc}_{\mathrm{easy}} + \mathrm{Acc}_{\mathrm{in}} + \mathrm{Acc}_{\mathrm{out}}\right). \end{aligned} \tag{5}$$

The captioner with CoT parses 91.00%, 90.40%, and 81.80% of the easy, in-domain, and out-of-domain outputs, respectively, for an overall parse rate of 86.87%. Qwen3-Omni Instruct with CoT parses 97.75%, 98.00%, and 99.40% of the three splits, respectively, for an overall parse rate of 98.52%. All reported SFT and GRPO checkpoints currently have 100% parse coverage.

Table 8: Reward-model training settings for the submitted checkpoints.

Setting	Easy SFT	Hard SFT	Dimension-wise GRPO
Initialization	Qwen3-Omni-30B-A3B	Easy SFT checkpoint	Hard SFT checkpoint
Training examples	4,390	4,436	9,136
Optimizer	AdamW (PyTorch)	AdamW (PyTorch)	Adafactor
Adam (β_1, β_2)	$(0.9, 0.95)$	$(0.9, 0.95)$	N/A
Adam ϵ	10^{-8}	10^{-8}	N/A
Learning rate	5×10^{-6}	5×10^{-6}	1×10^{-6}
Scheduler	Cosine	Cosine	Constant
Warmup ratio	0.05	0.05	0
Weight decay	0.1	0.1	0
Epochs	N/A (step-based)	N/A (step-based)	N/A (step-based)
Optimizer updates	400	1,000	600
Effective batch size	128	128	6 prompts / 48 responses
Group size G	N/A	N/A	8
Sampling temperature	N/A	N/A	1.0
KL coefficient β	N/A	N/A	0 (not used)
PPO clip ratio δ	N/A	N/A	N/A (not used)
Normalization ϵ	N/A	N/A	10^{-6}
Maximum sequence / response tokens	24,576	24,576	1,024 response
Checkpoint interval (updates)	100	100	50
Checkpoint selection	Step 400	Step 1,000	Step 600

F Implementation Details

F.1 Reward Model Training and Evaluation

Supervised fine-tuning. Stage 1 initializes from Qwen3-Omni-30B-A3B-Instruct and trains on easy pairs for 400 optimizer updates. Stage 2 continues from the Stage 1 checkpoint and trains on hard pairs for another 1,000 updates. In both stages, we fully update the language backbone while freezing the visual encoder and aligner. Each stage runs on eight GPUs using BF16 and DeepSpeed ZeRO 3 with optimizer offload. We use PyTorch AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.1, a cosine learning-rate schedule, a peak learning rate of 5×10^{-6}, and a warmup ratio of 0.05. These values are explicitly passed by the released MS-Swift 4.4.2 launch script. The maximum sequence length is 24,576, and gradient checkpointing is enabled. The batch size is one per GPU with 16 gradient accumulation steps, giving an effective batch size of 128. Checkpoints are saved every 100 updates. For each video, we sample up to 12 frames, with each frame represented by up to 128 visual tokens. Table 8 summarizes the optimization settings.

Dimension-wise GRPO. This stage starts from the hard SFT checkpoint and trains for 600 optimizer updates on the human-annotated hard-pair pool. Of the 9,146 raw comparisons, we exclude 10 annotations that select only "Other" and contain no valid rubric dimension, resulting in 9,136 training examples. The 4,390 easy cold-start pairs are not used in this stage. For every retained example, the dimension reward is computed exclusively from the supporting dimensions selected by human annotators, so all dimension-level GRPO supervision is grounded in human annotations. We freeze the visual and audio encoders and update the language thinker using six training GPUs. Two additional GPUs host independent vLLM rollout servers. Each optimizer update processes six prompts and eight sampled responses per prompt, giving 48 responses globally. Sampling uses temperature 1.0, top-$p = 1.0$, a repetition penalty of 1.05, and at most 1,024 generated tokens. We use Adafactor with a constant learning rate of 1×10^{-6}, zero weight decay, no warmup, BF16 precision, gradient checkpointing, gradient-norm clipping at 1.0, and seed 42. The answer and dimension rewards receive equal weight, and a response that omits any required dimension score receives zero reward. Advantages are standardized within each eight-response group using $\epsilon = 10^{-6}$. The implementation applies neither PPO ratio clipping nor a reference-model KL penalty. Checkpoints and evaluation results are written every 50 optimizer updates.

Evaluation. We use greedy decoding with vLLM. Each process receives the training comparison rubric. We extract the final choice from the `<answer>` tag; a response without a valid choice is incorrect for Total Acc and excluded from Parsed Acc.

F.2 LTX-2 POST-TRAINING

We initialize the policy from the 19B LTX-2 checkpoint and keep the backbone frozen. LoRA adapters with rank 32 and scaling factor 64 are inserted into the video and audio attention layers, feed-forward layers, and bidirectional cross-modal attention layers. The current and old policies use separate adapters, while disabling the adapters recovers the fixed reference policy.

For each prompt, the old policy generates eight video-audio candidates, which form 28 unique pairs. VA-Judger evaluates every pair using the five rubric dimensions shown in Appendix Figure 7. We average the scores obtained by each candidate across its seven comparisons. Dimension C provides the audio reward, dimension D provides the video reward, and the average of dimensions A, B, and E provides the shared cross-modal reward. Each reward is normalized within the eight-candidate group. The audio and video rewards are routed to their corresponding generation branches, while the shared reward updates both branches. Gemini 3.1 Pro (Google DeepMind, 2026a) is used only for offline data construction, while online post-training uses the locally deployed VA-Judger for efficient batched reward computation.

Following OmniNFT (Zhang et al., 2026), we use its 20K-prompt VGGSound training split and train for 400 optimizer updates with random seed 42. Each candidate is generated at 512×768 resolution with 121 frames at 24 FPS. Training rollouts use 20 denoising steps with video and audio classifier-free guidance scales of 1.5 and 3.0. We optimize with AdamW at a learning rate of 3×10^{-5}, $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 10^{-4}, and gradient clipping at 1.0. Training uses BF16 precision. For each update, we optimize over the first 8 scheduler timesteps in the 20-step rollout schedule, corresponding to a timestep fraction of 0.4. Advantages are clipped to $[-5, 5]$, the positive-negative prediction mixing coefficient is 1.0, and the reference regularization weight is 10^{-4}.

The video loss is reweighted using video-to-audio attention with a maximum weight of 1.5 and 400 warmup steps. Training runs with FSDP on seven GPUs, while one additional GPU hosts a VA-Judger inference server for batched reward computation. All generation model baselines use the same prompts, generation settings, and evaluation subset.

Qwen3-Omni-RL baseline. Qwen3-Omni-RL, reported as LTX-2 + Qwen3-Omni-RL in Table 2 and as Qwen3-Omni-RL in Figures 5 and 6, isolates the contribution of our reward-model training. It replaces VA-Judger with the untuned Qwen3-Omni-30B-A3B-Instruct, which is also the initialization of VA-Judger. The reward model receives the same rubric system prompt shown in Appendix Figure 7, evaluates the same 28 pairs per prompt, and its dimension scores are aggregated and routed to the audio, video, and shared rewards in the same way as described above. All other settings are identical to LTX-2 + VA-Judger, including the LoRA configuration, the 20K-prompt VGGSound training split, 400 optimizer updates, random seed 42, rollout resolution and denoising schedule, optimizer and loss hyperparameters, and hardware allocation. Consequently, the two post-trained models differ only in the reward model.

G FULL METRIC ALIGNMENT RESULTS

We convert each scalar metric into a pairwise prediction using its preferred direction and compare the prediction with the human label. Every metric is evaluated on all 1,150 benchmark pairs, comprising 400 easy, 250 in-domain, and 500 out-of-domain comparisons. Exact score ties receive half credit. Metrics selected as representatives in Table 1 are marked with $*$.

G.1 LABEL-FREE METRIC FUSION

We examine whether combining the seven representative automatic metrics provides a stronger proxy for human preference. The ensemble contains VideoAlign, Audiobox Aesthetics, CLIP Score, ImageBind text-audio similarity, ImageBind audio-video similarity, SynchFormer offset, and Javis Score. Both fusion methods assign equal importance to all metrics and do not use human preference labels to learn or tune weights. They are evaluated on the same 1,150 benchmark pairs as the individual metrics.

Table 9: Definitions and implementation details for less standard automatic metrics. All implementations use the JavisBench evaluation pipeline (Liu et al., 2026a); ↑ and ↓ indicate the preferred raw-score direction before conversion to pairwise accuracy.

Metric	Definition	Input	Dir.	Preprocessing and source
VideoPhy2	VideoPhy-2-AutoEval for semantic adherence and physical commonsense (Bansal et al., 2026)	Text, video	↑	Decode and sample the generated clip following the official evaluator through the JavisBench wrapper.
NISQA	Non-Intrusive Speech Quality Assessment, predicting perceptual speech quality (Mittag et al., 2021)	Audio	↑	Extract audio and apply the model-native waveform preprocessing and sampling rate.
DNSMOS	Deep Noise Suppression Mean Opinion Score, using the overall P.835 quality prediction (Reddy et al., 2022)	Audio	↑	Extract audio, convert it to the model-required mono waveform, and apply model-native resampling.
ViCLIP	Video-text contrastive similarity from InternVid (Wang et al., 2024a)	Text, video	↑	Sample video frames with the JavisBench wrapper and compute cosine similarity between video and text embeddings.
SyncNet	Audio-mouth synchronization network (Chung & Zisserman, 2016)	Audio, face video	Mixed	Use the JavisBench face-track and audio preprocessing. Confidence is higher-better; minimum embedding distance and absolute offset are lower-better.

Table 10: Human preference accuracy (%) of video-quality metrics. All raw metric scores are higher-better before pairwise conversion.

Metric	Easy	In-domain	Out-of-domain	Overall (micro)
VideoAlign Overall*	62.25	53.20	48.40	54.26
VideoPhy2	52.75	54.00	50.20	51.91
Motion Quality	60.00	51.20	50.00	53.74
Visual Quality	62.50	52.40	42.80	51.74
HPSv3	56.50	52.00	50.20	52.78
Video Aesthetic	57.75	50.00	47.40	51.57
Video Technical/Aesthetic	57.50	51.20	42.80	49.74
Video Technical/Overall	59.25	52.80	42.20	50.43
Video Technical/Technical	54.25	54.00	47.20	51.13

Equal-weight hard voting. For each pair, every metric votes for the clip with the better score. We orient all votes so that $v_m = +1$ selects the first clip and $v_m = -1$ selects the second clip. Higher values are preferred for all metrics except SynchFormer offset, for which a lower value is preferred. An exact score tie contributes no vote. The ensemble prediction is

$$\hat{y}_{\text{hard}} = \operatorname{sign}\left(\sum_{m=1}^{7} v_m\right). \tag{6}$$

Hard voting obtains 64.50%, 52.00%, and 50.00% accuracy on the easy, in-domain, and out-of-domain splits, respectively. The resulting micro accuracy over all 1,150 benchmark pairs is 55.48%.

Equal-weight Z-score soft voting. Hard voting ignores the magnitude of each score difference. We therefore also standardize the scores of each metric within each evaluation split,

$$z_m(x) = \frac{s_m(x) - \mu_m}{\sigma_m}, \qquad S(x) = \frac{1}{7}\sum_{m=1}^{7} \tilde{z}_m(x), \tag{7}$$

where μ_m and σ_m are computed from all clip scores for metric m in the corresponding split. We set $\tilde{z}_m(x) = -z_m(x)$ for SynchFormer offset and $\tilde{z}_m(x) = z_m(x)$ for the other six metrics so that a higher value is always preferred. The clip with the larger $S(x)$ is selected. This method obtains 67.00%, 58.00%, and 52.40% accuracy on the three splits, with an overall micro accuracy of 58.70%. Compared with Javis Score, the strongest of the seven representative individual metrics, Z-score soft voting improves easy, in-domain, and overall accuracy by 7.00, 2.40, and 1.66 percentage points, respectively, but decreases out-of-domain accuracy by 3.00 points. This result shows that the metrics contain complementary information, while their label-free combination remains insufficiently robust under generation model distribution shift.

Table 11: Human preference accuracy (%) of audio-quality metrics. All raw metric scores are higher-better before pairwise conversion.

Metric	Easy	In-domain	Out-of-domain	Overall (micro)
Audiobox Aesthetics*	58.50	50.00	44.00	50.35
Audiobox Aesthetics CE	61.50	50.00	42.20	50.61
Audiobox Aesthetics CU	62.25	48.00	38.60	48.87
Audiobox Aesthetics PC	45.25	54.40	53.60	50.87
Audiobox Aesthetics PQ	64.25	45.60	36.20	48.00
NISQA	53.75	51.60	43.80	48.96
DNSMOS	47.25	52.00	40.40	45.30

Table 12: Human preference accuracy (%) of text-video and text-audio consistency metrics. All raw similarity scores are higher-better before pairwise conversion.

Metric	Easy	In-domain	Out-of-domain	Overall (micro)
Text-video consistency				
CLIP Score*	55.75	54.80	53.80	54.70
ImageBind T-V	47.25	53.20	51.00	50.17
ViCLIP	53.75	43.60	49.60	49.74
Text-audio consistency				
ImageBind T-A*	58.50	52.80	51.60	54.26
CLAP Score	51.25	52.00	66.20	57.91

Table 13: Human preference accuracy (%) of audio-video semantic-consistency metrics. All raw scores are higher-better. Audio-Video Alignment, AVHScore, and ImageBind A-V are ImageBind-based metrics but differ in preprocessing and feature aggregation.

Metric	Easy	In-domain	Out-of-domain	Overall (micro)
Audio-Video Alignment*	61.25	55.60	51.80	55.91
AVH Score	61.75	58.80	54.40	57.91
ImageBind A-V	61.25	58.80	55.20	58.09

Table 14: Human preference accuracy (%) of synchronization and aggregate audio-video metrics. Preferred raw-score directions are repeated beside each metric.

Metric	Easy	In-domain	Out-of-domain	Overall (micro)
SyncNet Confidence ↑	64.50	50.80	40.20	50.96
SyncNet Min Distance ↓	54.50	59.20	47.00	52.26
SyncNet $\lvert\mathrm{Offset}\rvert$ (frames) ↓	65.25	48.00	42.00	51.39
SynchFormer $\lvert\mathrm{Argmax\ Offset}\rvert$ (s) ↓	63.75	47.20	50.60	54.43
SynchFormer $\lvert\mathrm{Offset}\rvert$ (s)* ↓	61.50	46.80	48.20	52.52
Desync ↓	66.25	49.20	52.00	56.35
Lip-sync Confidence ↑	63.25	53.60	42.80	52.26
Javis Score* ↑	60.00	55.60	55.40	57.04

H Human Evaluation Protocol

We recruit 20 adult participants who have normal or corrected-to-normal vision and report no hearing impairment. All participants provide informed consent before the study. Participation is voluntary and uncompensated. For each of the same 200 prompts used in the automatic evaluation, LTX-2, OmniNFT, and our VA-Judger post-trained model each generate one video-audio output, yielding 200 triplets and 600 generated clips. Every participant evaluates all 200 prompt-aligned triplets, so each triplet receives exactly 20 votes and the study produces $20 \times 200 = 4{,}000$ votes. The display order of the three candidate outputs is randomized for each comparison. Participants may replay the clips before selecting the best overall output in each three-way forced-choice comparison, and candidate identities are concealed behind generic video indices. Only complete response sheets containing all 200 selections are included. We compute the final preference percentages directly over the 4,000 votes, without first aggregating each prompt to a majority winner. VA-Judger receives 2,492 votes (62.30%), OmniNFT receives 1,105 votes (27.63%), and LTX-2 receives 403 votes (10.08%).

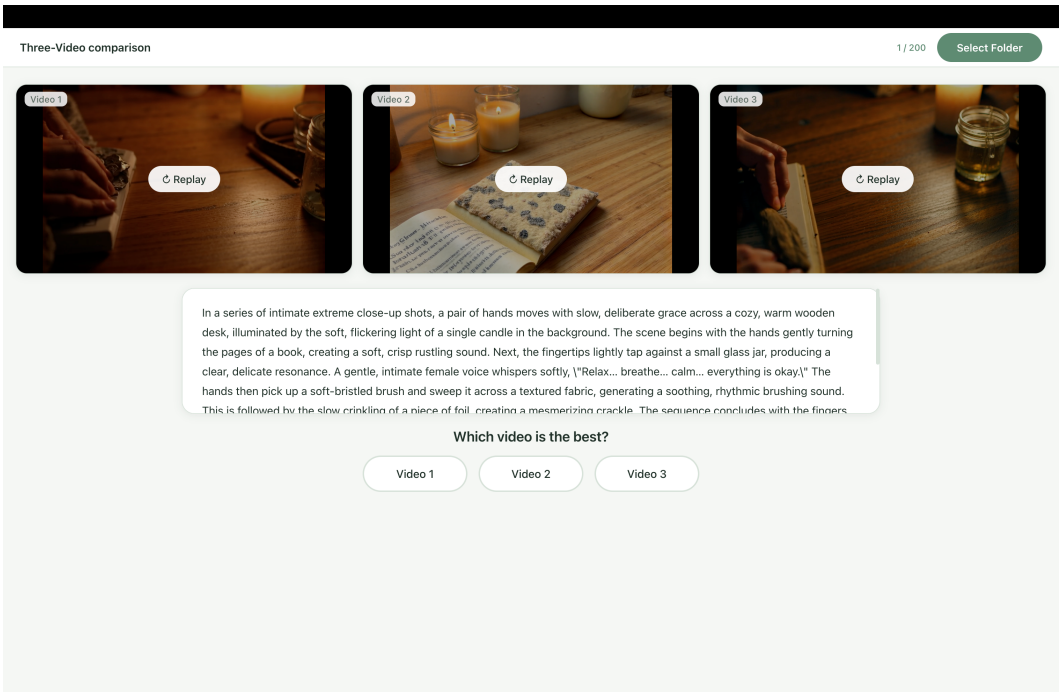


Figure 11: Interface for the three-way human preference study. For each prompt, participants replay the three video-audio outputs and select the best result among LTX-2, LTX-2 post-trained with OmniNFT, and LTX-2 post-trained with VA-Judger. The interface identifies candidates only by generic video indices.

I LIMITATIONS

The current experiments focus primarily on short video-audio clips and a fixed set of recent generation models. Future work can extend the evaluation to longer-form content and newly released generation models. In addition, expanding the benchmark with more prompt styles and application scenarios would provide an even broader test of reward-model generalization.

Failure case. Figure 12 presents a representative table-tennis example with large-amplitude player motion and a small ball moving at high speed. These demanding dynamics can produce implausible frames and motion artifacts. This primarily reflects the limited capability of the LTX-2 base generation models to model rapid, fine-grained motion, which is difficult to fully overcome through reward-based post-training alone. The complete videos are provided in the supplementary material.

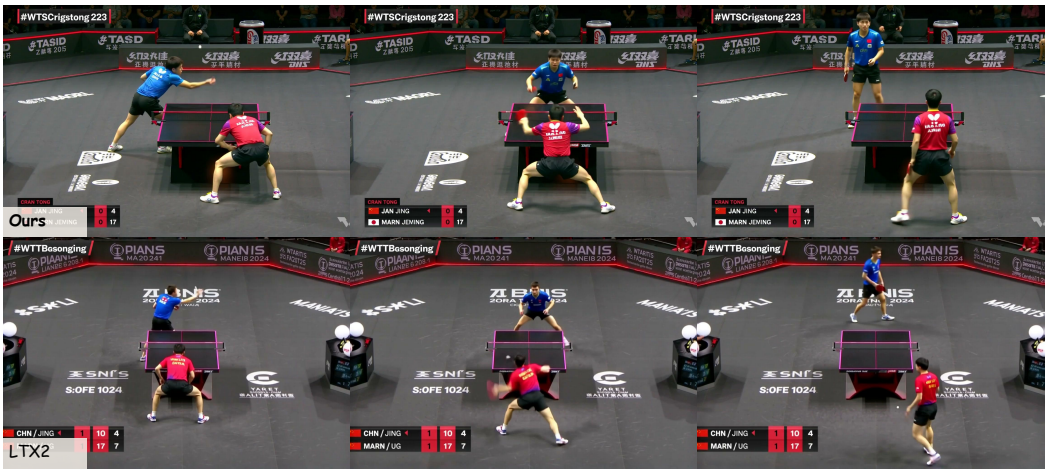


Figure 12: Representative failure case on a table-tennis scene with large-amplitude player motion and a fast-moving ball. Rapid, fine-grained dynamics remain challenging for the LTX-2 base generation models and can produce implausible frames even after post-training.