JACE: JUDGE WITH ATTRIBUTION, CALIBRATION, AND EVALUATION FOR MULTIMODAL AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation increasingly combines multimodal conditions to produce visual and auditory content, placing greater demands on evaluation. Existing evaluators struggle to jointly support diverse generation tasks and provide interpretable judgments aligned with human preferences. Accurate defect prediction requires understanding physical plausibility, temporal coherence, and cross-modal consistency, which is challenging. Even accurate diagnosis, however, does not determine preference: shared defects may offer little basis for comparison, while different defects may carry different importance. We introduce JACE (Judge with Attribution, Calibration, and Evaluation), a generalist evaluator that connects defect diagnosis to human preference through explicit scoring contributions. JACE first learns defect predictions from non-exhaustive annotations, using complementary representations for category-specific refinement. It then freezes the diagnostic module and learns defect-category weights from separately collected pairwise preferences, without requiring both annotation types on the same videos. Combining weighted defect contributions with a base quality score yields interpretable win, tie, and loss judgments. JACE supports text, image, audio, and video conditioning across text-to-video, image-to-video, reference-to-video, and video editing tasks, jointly assessing generated visuals and sound. Its shared scoring mechanism connects individual-video diagnosis and scoring, candidate comparison, and generator ranking, supporting evaluation across generators. Across four generation tasks, JACE significantly improves defect Recall@3 (+24.6%) and non-tie accuracy by (+8.09%) over the strongest baseline (Gemini 3.7 Flash).

1 INTRODUCTION

Video generation models (KlingAI, 2026; Team Wan et al., 2025; ByteDance Seed Team, 2026; MiniMax, 2026; Google Deepmind, 2026; SpaceXAI, 2026) are evolving rapidly, moving beyond text-conditioned visual synthesis toward increasingly diverse combinations of inputs and outputs. Recent systems support text, images, audio, and video as conditioning inputs, enabling tasks ranging from text-to-video generation to reference-guided synthesis and video editing (Team Wan et al., 2025; ByteDance Seed Team, 2026; Jiang et al., 2025; OpenAI, 2025; Google Deepmind, 2026). Generated outputs increasingly combine visuals and sound, further expanding the scope of generation. These developments place broader demands on evaluation: an evaluator must interpret task-specific combinations of conditions, assess visual and auditory content, and make judgments aligned with human preferences. To support comparison across generators and successive versions within a model family, they must also capture subtle quality differences and explain how specific defects influence their judgments, providing actionable feedback for model improvement.

Existing methods address complementary aspects of these requirements. Dimension-specific metrics assess attributes such as subject consistency and motion smoothness (Huang et al., 2024; 2026b), while human-supervised evaluators learn quality scores and pairwise preferences (He et al., 2024; Liu et al., 2025). Recent approaches further extend evaluation to joint audio-video content (Jin et al., 2026; Huang et al., 2026a) and connect quality judgments with structured attributes or defect explanations (Xu et al., 2026; Wu et al., 2025b). As tasks, conditions, and outputs diversify, simple

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.

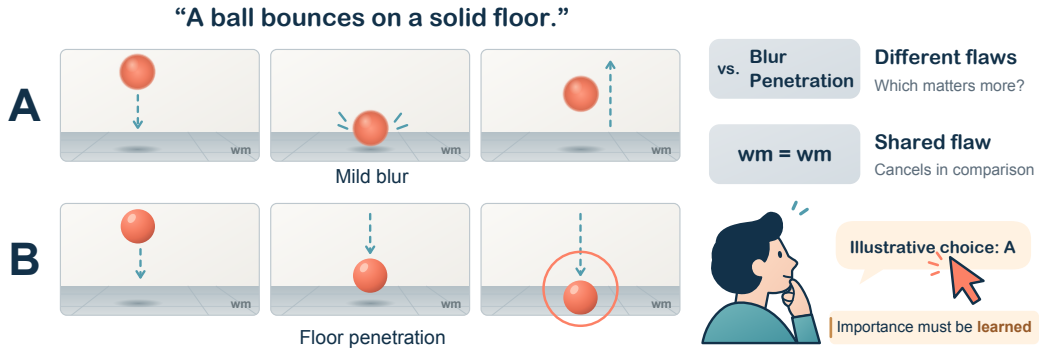


Figure 1: **From diagnosis to preference.** This illustrative pair contrasts mild blur in A with floor penetration in B. Detecting their shared watermark improves diagnosis, but equal contributions cancel in comparison. Human preference learning weights the remaining defects.

quality scores and basic evaluation dimensions remain insufficient to cover multimodal audio-video generation comprehensively while reflecting human judgments.

Connecting evaluation to human judgments requires distinguishing two complementary questions: what defects are present in an individual video, and how differences between candidates affect overall preference. These questions serve distinct needs—diagnosis and comparison—and support separate annotation workflows that can scale independently. Defect annotations remain non-exhaustive, as annotators prioritize salient problems and may overlook subtler ones, while pairwise preferences capture overall quality trade-offs without specifying each defect's influence. Crucially, identifying defects alone does not determine preference. As illustrated in Figure 1, detecting a shared watermark improves diagnosis but offers little basis for choosing between candidates; comparing mild blur with floor penetration instead requires learning their relative importance from human preferences. The challenge is therefore to connect incomplete diagnostic evidence to preference judgments without requiring both annotations on the same videos.

We introduce JACE (**J**udge with **A**ttribution, **C**alibration, and **E**valuation), connecting weakly supervised attribution learning with human-preference-driven evaluation calibration. In the diagnostic stage, the omni-modal understanding model Qwen3-Omni (Xu et al., 2025) is employed to learn defect representations from sparse annotations, using recorded defects and overall-good samples while retaining uncertainty about unannotated categories. Auxiliary diagnostic representations provide category-specific residual corrections. In the preference stage, the diagnostic module is fixed, and independent human preferences determine how its outputs contribute to evaluation. Learned influence weights combine candidate differences in attribution with a base quality term that captures factors beyond the defect taxonomy. This separates learning to identify defects from learning how they affect overall choice, while connecting both through explicit scoring contributions.

JACE brings this mechanism to text-to-video (T2V), image-to-video (I2V), reference-to-video (R2V), and video editing (V2V), jointly assessing visuals and sound under their multimodal conditions. Its scope spans three connected levels. For individual outputs, it provides diagnosis, relative quality, and category-level scoring contributions. For candidate pairs, differences in these scoring terms determine wins, ties, and losses and explain each defect's influence. Under a common protocol, aggregated comparisons yield generator rankings for cross-model comparison and within-family iteration analysis. Sharing the scoring terms across these levels makes explanations correspond directly to the quantities that reflect human preference. Our contributions are threefold:

- We develop weakly supervised attribution from sparse defect annotations, combining omni-modal representations and category-specific residual correction to establish structured diagnostic evidence.
- We introduce attribution-driven preference calibration under separate supervision. Independent preferences determine defect influence weights, enabling candidate comparison through base quality and weighted attribution differences with explicit contribution.
- We build a generalist evaluator spanning multimodal generation tasks and evaluation levels, connecting video-level diagnosis and scoring, interpretable pairwise judgements, and generator ranking.

2 Related Work

Video quality evaluation and human preference learning. Video evaluation measures appearance, motion (Unterthiner et al., 2018), and content adherence. Hierarchical benchmarks cover subject consistency, motion smoothness, image-conditioned generation, and intrinsic faithfulness (Huang et al., 2024; 2026b; Zheng et al., 2025). Specialized protocols assess composition, temporal relations, and physical commonsense (Sun et al., 2025; Feng et al., 2025; Bansal et al., 2025), while vision-language methods evaluate alignment with conditioning descriptions (Wu et al., 2024; Lin et al., 2024; Guan et al., 2025). These protocols relate quality differences to specific capabilities (Liu et al., 2024; Han et al., 2025). Human-judgment datasets support a complementary learning approach (Zhang et al., 2025; Shekhar et al., 2026): multimodal evaluators learn perceptual scores (He et al., 2024; Wang et al., 2025b), and pairwise preferences train rewards for candidate selection and generation optimization (Liu et al., 2025; Tong et al., 2025; Wu et al., 2025a; Lian et al., 2026). Such rewards can retain quality dimensions and fine-grained criteria, connecting aspect-level judgments to overall choices.

Interpretable evaluation and fine-grained attribution. Evaluation rationales ground judgments in video content (Liu & Zhang, 2026; Chinchure et al., 2026). Explanation supervision and reinforcement fine-tuning train evaluators to produce reasoning alongside scores or choices (Wang et al., 2024b; He et al., 2025; Wang et al., 2025c; 2026b; Zhang et al., 2026). Q-Save jointly learns scoring and explanation from quality and attribution annotations on the same videos (Wu et al., 2025b). REACT produces scores and attribution labels for structural distortions through frame-level reasoning and uses pairwise rewards for alignment (Wang et al., 2026c). Attribute aggregation and concept bottlenecks offer further ways to organize interpretable evidence (Xu et al., 2026; Laguna et al., 2025; Shang et al., 2024). These approaches connect quality judgments to content-level explanations. Our setting instead requires learning from non-exhaustive defect annotations and transferring their diagnostic information to preferences on a separate video collection. The distinction concerns both the reliability of defect representations and their influence on candidate choice.

Multitask and joint audio-video judging. Multitask reward models share judgment capabilities across visual understanding and generation evaluation, with extensions to audio and other modalities (Wang et al., 2025d; Jin et al., 2026). Task-dependent rubrics translate text, reference images, or source videos into evaluation criteria (Zeng et al., 2026; Jia et al., 2026). Joint audio-video benchmarks assess visual and auditory quality, synchronization, and cross-modal consistency (Hua et al., 2026; Yang et al., 2026). Omni-modal judges jointly interpret text, visuals, and sound, while preference-trained audio-video rewards incorporate cross-modal relationships into candidate choices (Liang et al., 2026; Huang et al., 2026a). Together, these directions broaden evaluation across conditioning inputs and output modalities. JACE connects this scope to sparse attribution and independent preference supervision, linking diagnosis, candidate comparison, and generator ranking through decomposable quality scores.

3 Problem Setup

We evaluate T2V, I2V, R2V, and V2V outputs under multimodal conditions. Write an input as $x = (t, c, v, u, m)$, with task t, conditioning material c, generated visuals v, audio track u, and modality-availability record m. Conditions may contain text, images, audio, reference videos, or source videos for editing; silent outputs have $u = \varnothing$. Evaluation considers output quality and adherence to these conditions. Let $\mathcal{K}$ be the shared defect taxonomy, $K = |\mathcal{K}|$, and $\mathcal{K}(x) \subseteq \mathcal{K}$ its applicable subset under the task and protocol (e.g., reference-related categories do not apply to T2V). Applicability defines evaluation scope, independently of whether labels are available.

The attribution data are $\mathcal{D}_{\text{attr}} = \{(x_i, S_i, g_i)\}_{i=1}^{N_{\text{attr}}}$, where N_{attr} is the number of videos, $S_i \subseteq \mathcal{K}(x_i)$ records salient defects and $g_i \in \{0, 1\}$ indicates overall-good status. Human perceptual limits prevent exhaustive annotation, particularly for subtle defects, so unrecorded applicable categories remain unknown. Sparsity refers to this incomplete category coverage. An overall-good label $g_i = 1$ means no salient defects were observed and supplies a sample-level negative reference rather than exhaustive category verification.

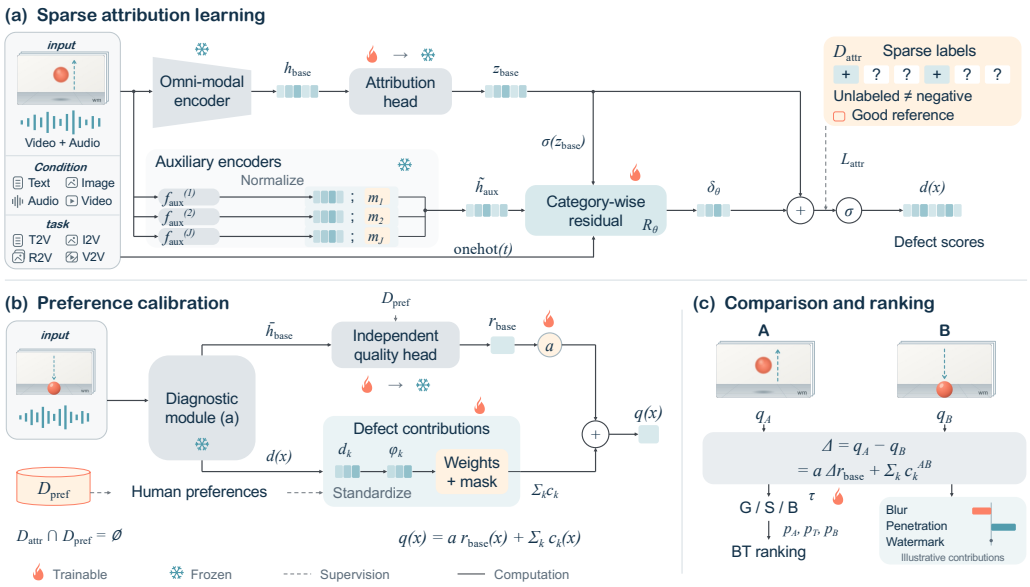


Figure 2: **Overview of JACE.** (a) Sparse supervision trains base diagnosis, then category-wise residual correction with auxiliary and task information. (b) On separate videos, preferences fit an independent quality head, then attribution weights and scoring parameters, with diagnosis frozen. (c) Shared scoring supports decomposable, tie-aware comparisons and generator ranking. Flames/snowflakes indicate trainable/frozen modules; arrows between them indicate training then freezing. Thumbnails and bars are illustrative.

The preference data are $\mathcal{D}_{\text{pref}} = \{(x_j^A, x_j^B, y_j)\}_{j=1}^{N_{\text{pref}}}$, where N_{pref} is the number of pairs and the two candidates in each pair share a task and generation conditions. They may come from different generators or different versions or configurations within a model family. Labels $y_j \in \{A, T, B\}$ denote an A win, a tie, or a B win, corresponding to Good/Same/Bad (GSB) relative to A. GSB is distinct from single-video overall-good status. Annotators judge overall perceived quality, reflecting combined quality trade-offs without supplying defect labels or importance weights. The attribution and preference collections are disjoint at the generated-video level.

We study the transfer of sparsely supervised defect information to preferences on the separate collection. For each video, the evaluator returns $d(x) \in [0,1]^K$, whose components indicate defect presence strength, and $q(x) \in \mathbb{R}$, with higher values denoting better relative quality. Explanations quantify category contributions. For a pair, it returns nonnegative probabilities $\mathbf{p} = (p_A, p_T, p_B)$ summing to one to determine GSB outcomes. Differences in category contributions explain how defects affect the candidate choice. For generators $\mathcal{G} = \{G_1, \ldots, G_L\}$, comparisons under a common protocol yield Bradley–Terry strengths β_i (Bradley & Terry, 1952; Jiang et al., 2024); larger strengths indicate higher relative quality and determine ranks within connected comparison components.

4 Method

4.1 Overview

JACE learns diagnosis on $\mathcal{D}_{\text{attr}}$, then freezes the diagnostic module and learns quality scoring on $\mathcal{D}_{\text{pref}}$. A frozen omni-modal encoder extracts $h_{\text{base}}(x) = f_{\text{base}}(x)$, while J frozen auxiliary encoders provide $h_{\text{aux}}^{(j)}(x) = f_{\text{aux}}^{(j)}(x)$. Their concatenation is $h_{\text{aux}}(x) = [h_{\text{aux}}^{(1)}(x); \ldots; h_{\text{aux}}^{(J)}(x)]$, allowing one or multiple auxiliary encoders. The base head produces defect logits $z_{\text{base}}(x) = \text{Head}_{\text{attr}}(h_{\text{base}}(x)) \in \mathbb{R}^K$. Auxiliary features support category-specific residual corrections, yielding the final diagnosis $d(x)$. Preference learning then determines how this diagnosis contributes to $q(x)$ and candidate choice. Figure 2 summarizes the training and inference paths.

4.2 Weakly Supervised Learning from Sparse Attribution

Sparse attribution distinguishes a defect's presence from its annotation (Durand et al., 2019; Cole et al., 2021). Recorded defects supply positive labels, overall-good samples supply weaker negative references, and other unannotated categories remain unknown. The diagnostic head uses fixed attribution-training statistics to normalize features and jointly predicts defect logits and an overall-good logit $z_{\mathrm{base},G}(x)$. Preprocessing and head details appear in Appendix B.1. Write $d_{\mathrm{base}}(x) = \sigma(z_{\mathrm{base}}(x))$, where σ is the sigmoid function.

For sample i, let $y_{ik} = \mathbf{1}[k \in S_i]$. We weight category supervision by annotation status:

$$\omega_{ik} = \begin{cases} 1, & k \in S_i, \\ \rho, & g_i = 1, \quad k \in \mathcal{K}(x_i) \setminus S_i, \\ 0, & \text{otherwise}, \end{cases} \qquad 0 < \rho < 1. \tag{1}$$

The smaller weight ρ reflects the coarser evidence provided by an overall-good label. Inapplicable and other unannotated categories do not enter the observed-label loss. Writing $z_{ik} = z_{\mathrm{base},k}(x_i)$ and $z_{iG} = z_{\mathrm{base},G}(x_i)$, we minimize

$$\mathcal{L}_{\mathrm{attr}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\sum_{k=1}^{K} \omega_{ik} \ell_{\mathrm{BCE}}(z_{ik}, y_{ik})}{\max\left(1, \sum_{k=1}^{K} \mathbf{1}[\omega_{ik} > 0]\right)} + \lambda_G \eta_i \ell_{\mathrm{BCE}}(z_{iG}, g_i) + \lambda_U \mathcal{U}_i(z) \right]. \tag{2}$$

Here, n is the training sample count, ℓ_{BCE} is binary cross-entropy on logits, and $\eta_i \in \{0, 1\}$ equals one when an explicit overall-good or not-good label is available. The coefficients λ_G and λ_U control sample-level supervision and unknown-category regularization. For samples with recorded defects and $g_i = 0$, $\mathcal{U}_i$ averages $\mathrm{softplus}(z_{ik}) = \log(1 + \exp(z_{ik}))$ over applicable unannotated categories; it is zero otherwise. This weak shrinkage prior limits unsupported score growth while retaining the distinction between observed labels and unknown categories.

4.3 Complementary Diagnostic Correction

The base diagnosis learns defect predictions from sparse annotations, while auxiliary encoders provide complementary content representations. The correction module combines them to learn which category predictions to strengthen or weaken. Let $\widetilde{h}_{\mathrm{aux}}(x)$ concatenate normalized auxiliary features and their availability information, using training statistics fixed at inference. A residual head R_θ combines these features with the current diagnosis and the task's one-hot encoding:

$$\begin{aligned} \delta_\theta(x) &= R_\theta\Big([d_{\mathrm{base}}(x); \widetilde{h}_{\mathrm{aux}}(x); \mathrm{onehot}(t)]\Big), \\ z_J(x) &= z_{\mathrm{base}}(x) + \delta_\theta(x), \qquad d(x) = \sigma(z_J(x)). \end{aligned} \tag{3}$$

Brackets denote concatenation. Positive and negative components of $\delta_\theta(x) \in \mathbb{R}^K$ respectively strengthen and weaken the corresponding defect judgments, preserving the base diagnosis as the reference.

After base diagnostic training, we freeze its head and all encoders and train R_θ with the attribution objective applied to corrected logits, plus a penalty on residual magnitude. The overall-good output remains unchanged. Appendix B.2 specifies the initialization and regularized objective. The resulting head and correction module are then frozen together for preference learning.

4.4 Attribution-Driven Preference Calibration

Preference calibration learns how fixed diagnostic evidence contributes to overall quality, separating defect identification from the weighting of defects in human choice. On $\mathcal{D}_{\mathrm{pref}}$, we first fit an independent quality head $r_{\mathrm{base}}(x) = \theta_q^\top \bar{h}_{\mathrm{base}}(x)$ to pairwise human preferences. The frozen base features $\bar{h}_{\mathrm{base}}$ use preference-training statistics for standardization. This head can capture quality factors beyond the defect taxonomy and remains fixed when learning attribution influence weights.

To make diagnostic scores comparable across categories, standardize each applicable score as $\phi_k(x) = (d_k(x) - \mu_k)/\max(s_k, \varepsilon)$. Here, μ_k and s_k are its mean and standard deviation over

applicable preference-training videos, and $\varepsilon > 0$ bounds the scale away from zero. We combine base quality and attribution contributions as

$$q(x) = a\, r_{\text{base}}(x) + \sum_{k=1}^{K} c_k(x), \qquad c_k(x) = -\mathbf{1}[k \in \mathcal{K}(x)]\, w_k \phi_k(x). \tag{4}$$

Human preferences determine the base scale $a \geq 0$ and defect influence weights $w_k \geq 0$. Increasing an applicable category's standardized defect score by one lowers quality by w_k; zero weights exclude categories from adjustment. The applicability indicator removes out-of-scope categories. Statistics and scoring parameters are fixed at inference.

The diagnosis $d_k(x)$ describes defect presence, whereas $c_k(x)$ quantifies its learned quality contribution. Centering makes this contribution relative to the preference-training mean: for a positive weight, above-average defect scores incur a penalty and below-average scores yield a relative bonus. Thus, preference learning determines the importance of diagnostic evidence, and explanations follow directly from the terms used in scoring.

4.5 Difference-Based Pairwise Comparison and Generator Ranking

For candidates under the same task and conditions, define the contribution difference $c_k^{AB} = c_k(x^A) - c_k(x^B)$. Their quality difference decomposes as

$$\Delta = q(x^A) - q(x^B) = a\big[r_{\text{base}}(x^A) - r_{\text{base}}(x^B)\big] + \sum_{k=1}^{K} c_k^{AB}. \tag{5}$$

A positive c_k^{AB} supports A, a negative value supports B, and its magnitude measures the effect on the comparison score. Equal contributions cancel. Diagnostic improvements therefore affect preference only through the weighted candidate differences they change.

We model an A win, a B win, and a tie by $p_A = \sigma(\Delta - \tau)$, $p_B = \sigma(-\Delta - \tau)$, and $p_T = 1 - p_A - p_B$, respectively. Increasing $\tau \geq 0$ raises tie probability at a fixed Δ. These probabilities determine GSB outcomes through a prespecified decision rule. Swapping candidates negates Δ and c_k^{AB}, exchanges win/loss probabilities, and preserves tie probability.

With the quality head and diagnostic module fixed, we jointly learn a, $w = (w_1, \ldots, w_K)$, and τ by minimizing

$$\mathcal{L}_{\text{pref}} = -\sum_{j} \gamma_j \log p_{y_j}(x_j^A, x_j^B) + \lambda_w \|w\|_2^2. \tag{6}$$

The pair weights γ_j balance prompt and media-association groups, and λ_w controls weight regularization; Appendix B.3 describes the pair-weighting principle. Overall choices supervise the scoring combination, while c_k^{AB} explains each category's contribution.

Under a common task and condition distribution, we aggregate pairwise evidence including ties to fit Bradley–Terry strengths, with $P(G_i \succ G_j) = \sigma(\beta_i - \beta_j)$ (Bradley & Terry, 1952). Strengths determine ranks within connected comparison components. At inference, diagnosis, quality scores, and attribution contributions can be computed independently for each video and reused across comparisons, connecting video-level evaluation to pairwise decisions and generator ranking.

5 Experiments and Analysis

5.1 Experimental Setup

Data and supervision. We evaluate JACE at three levels: defect diagnosis, overall quality judgment, and generator comparison, examining how sparse attribution informs human preference modeling. Experiments cover T2V, I2V, R2V, and V2V, using task-specific conditions together with generated outputs. Attribution training uses salient-defect annotations and overall-good labels for individual videos; preference training uses win/tie/loss judgments on a separate video collection. The shared taxonomy contains 46 defect categories, with only applicable categories enabled for each task, including 41 for T2V. The two supervision sources respectively learn diagnostic representations and their contributions to evaluation.

Table 1: **Overall evaluation results.** Recall@3 and accuracies are percentages; τ_b measures ranking agreement. The best results are marked in **bold**. $\times$: no defect-category output; –: attribution mapping and top-3 extraction protocol are unavailable.

Method	R@3 ↑	Non-tie acc. ↑	GSB ↑	τ_b ↑
VideoScore2 (He et al., 2025)	–	56.39	47.67	0.7454
VideoReward (Liu et al., 2025)	$\times$	58.35	48.45	0.6727
VisionReward (Xu et al., 2026)	–	57.75	46.32	0.7455
VA-Judger (Huang et al., 2026a)	–	56.04	45.87	0.6364
Qwen3-Omni (Xu et al., 2025)	18.37	53.55	37.59	0.1692
Qwen3.5-Omni-Plus (Qwen Team, 2026)	21.72	55.66	49.01	0.6000
Gemini 3.7 Flash (Google, 2026)	31.92	61.43	47.64	0.7454
JACE (Ours)	**56.52**	**69.52**	**58.31**	**0.8182**

Evaluation protocol and metrics. Evaluation data provide defect labels, pairwise preferences, and generator comparisons. Training, validation, and test data are split by prompt and media-association group. Parameter selection, normalization statistics, and decision rules use only training or validation data. We assess relative quality through the resulting candidate comparisons against independent human preference labels. Generator rankings use a common prompt distribution and are compared with human rankings on the same comparison graph. Attribution is measured by Recall@3 over observed defects, with category and task averages. Pairwise evaluation reports GSB accuracy, non-tie accuracy, and Tie-F1. At the generator level, we fit Bradley–Terry strengths and measure agreement with human rankings using Kendall's τ_b (Bradley & Terry, 1952). Ties count as half-wins, and rankings are compared within connected components of the comparison graph.

Baselines. We compare representation models trained under matched supervision, specialized evaluators, and general multimodal models that directly judge quality. Representation comparisons cover image–text, video, quality-adapted, and geometric features (Radford et al., 2021; Cherti et al., 2023; Tschannen et al., 2025; Wu et al., 2023; Wang et al., 2024a; Assran et al., 2025; Vyas et al., 2026; Wang et al., 2025a; 2026a; Xu et al., 2025), using matched training data and readout protocols to assess their support for attribution and preference learning. Direct evaluators include video reward models (Liu et al., 2025), fine-grained evaluators (He et al., 2025; Xu et al., 2026; Wu et al., 2025b), and joint audio-video judges (Huang et al., 2026a), alongside API judgments from Qwen3.5-Omni-Plus (Qwen Team, 2026) and Gemini 3.7 Flash (Google, 2026).

5.2 Overall Evaluation Results

Video-level defect attribution. Table 1 summarizes video-level diagnosis, pairwise comparison, and generator ranking. JACE achieves 56.52% Recall@3, exceeding the baselines with reported attribution results. Gemini 3.7 Flash achieves 31.92%, a gap of 24.60 percentage points. Figure 3(b) further compares attribution across backbones. JACE obtains the highest Recall@3 on all four tasks, with a task average 13.72 points above Qwen3-Omni Head's 42.80%. These results show that JACE retrieves observed defects among its top predictions, providing structured evidence for locating problems in generated content. The role of this diagnosis in relative quality scoring is assessed through candidate comparisons.

Pairwise comparison. Non-tie accuracy assesses choices with a clear human preference, whereas GSB accuracy assesses three-way decisions including ties. JACE achieves 69.52% non-tie accuracy, exceeding the strongest baseline, Gemini 3.7 Flash at 61.43%, by 8.09 percentage points. Its GSB accuracy of 58.31% exceeds Qwen3.5-Omni-Plus at 49.01% by 9.30 points. JACE leads both metrics, indicating close agreement with human candidate choices. Together, these results demonstrate JACE's ability to identify specific defects and predict overall human preferences within a unified evaluation framework. The strongest baseline differs between the two metrics, while JACE remains ahead in both, suggesting that its advantage persists when ties are included in the decision space.

Figure 3(a) relates attribution to preference performance across backbones. JACE achieves the highest mean Recall@3 and GSB accuracy. Its GSB accuracy exceeds the strongest backbone baseline,

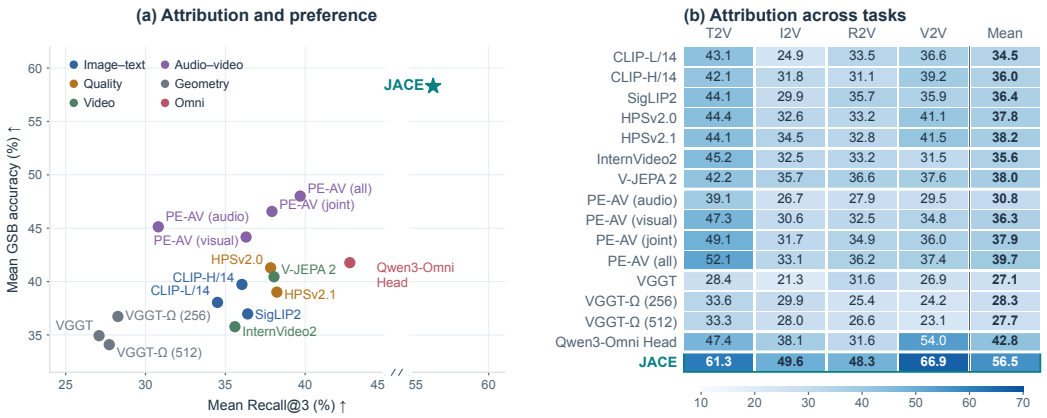


Figure 3: **Attribution and preference across backbones.** (a) Task-mean Recall@3 versus GSB (%); colors denote representation groups, the star JACE. The horizontal 45–55 interval is compressed; scales on either side are equal. (b) Per-task Recall@3 (%) and its equally weighted mean.

PE-AV (all) at 48.01%, by 10.30 percentage points. The two metrics do not order representations identically: Qwen3-Omni Head has higher mean Recall@3 than PE-AV (all), but lower mean GSB accuracy. This difference distinguishes defect recognition from overall preference judgment and motivates learning how diagnostic information affects candidate choice. For example, recovering a defect shared by both candidates can improve recall while contributing little to their score difference.

Generator ranking. We aggregate candidate comparisons into Bradley–Terry rankings and measure their agreement with human rankings. JACE achieves Kendall's $\tau_b = 0.8182$, exceeding the strongest baseline value of 0.7455 by 0.0727. Aggregating candidate-level quality judgments thus yields a generator ordering closer to the human ordering. This result supports using JACE for overall generator comparison, connecting video-level diagnosis, candidate selection, and model-level evaluation within one framework. Since τ_b depends on relative ordering rather than score magnitudes, this improvement reflects better preservation of human judgments about which generators perform better overall.

5.3 Ablation and Mechanism Analysis

Sparse attribution and complementary diagnosis. In Table 2, the full weakly supervised base diagnosis achieves a task-mean Recall@3 of 42.8%. Treating unlabeled categories as negatives, removing the overall-good reference, and removing unknown-category regularization reduce this score by 3.9, 2.1, and 1.2 percentage points, respectively. These results support distinguishing recorded defects, overall-good samples, and unknown categories. Category-specific residual correction increases task-mean and category-mean Recall@3 from 42.8% and 24.8% to 56.5% and 38.3%, respectively, outperforming both prediction-score fusion and feature concatenation. This supports using auxiliary information to correct predictions relative to the base diagnosis. The category-mean gain of 13.5 percentage points further shows that the improvement persists when defect categories receive equal weight, complementing the task-level comparison.

Joint calibration of base quality and attribution. In Table 3, quality-only and attribution-only scoring achieve GSB accuracies of 54.0% and 51.5%, respectively, compared with 58.3% when combined, indicating that the two scoring components are complementary. With the same corrected diagnosis, learning category-specific defect weights instead of equal weights further improves non-tie accuracy and GSB accuracy by 2.0 percentage points each, and Tie-F1 by 2.4 points. Distinguishing the preference contributions of defect categories improves both candidate selection and tie recognition. With diagnostic outputs held fixed, these gains reflect more effective use of the same defect evidence through weights learned from human preferences.

Transfer from attribution to preference. With the same corrected diagnosis and independent quality head, recalibration on the same preference-training set achieves 58.3% GSB accuracy and

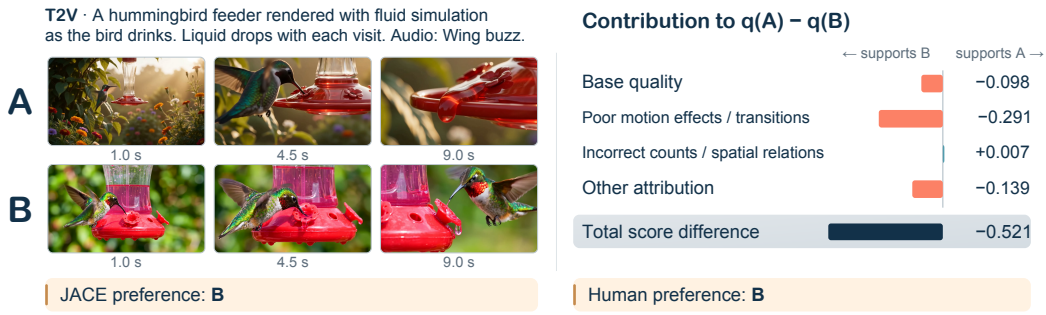


Figure 4: Hummingbird-feeding comparison with matched timestamps. The right panel decomposes the score difference, showing the largest attribution contribution favoring each candidate and aggregating the rest. Positive values favor A; negative values favor B.

Table 2: **Attribution ablations.** R@3 (%) and task-mean gain Δ (pp).

Configuration	Task ↑	Class ↑	Δ
Base diagnosis	42.8	24.8	0.0
Unlabeled as negatives	38.9	21.2	−3.9
w/o GOOD reference	40.7	24.1	−2.1
w/o unknown reg.	41.6	23.8	−1.2
PE-AV (all) only	39.7	25.8	−3.1
Prediction-score fusion	46.8	28.9	+4.0
Feature concatenation	50.1	34.0	+7.3
JACE (Ours)	**56.5**	**38.3**	**+13.7**

Table 3: **Preference ablations.** Non-tie accuracy, GSB accuracy, and Tie-F1 (%) assess pairwise decisions and tie recognition. Quality-based rows share the independent quality head. Higher is better.

Configuration	Non-tie	GSB	Tie-F1
Quality only	65.8	54.0	32.8
Attribution only	63.4	51.5	33.6
Quality + equal weights	67.5	56.3	35.6
Quality + base diagnosis	68.1	56.9	36.1
Corrected, fixed params.	67.9	56.6	35.8
JACE (Ours)	**69.5**	**58.3**	**38.0**

69.5% non-tie accuracy, improving over fixed preference parameters by 1.7 and 1.6 percentage points, respectively. These results show that learning the evaluation contributions of corrected diagnostic information through preference supervision improves agreement with human choices.

5.4 Attribution Explanations and Case Analysis

Figure 4 illustrates how JACE translates defect diagnosis into an interpretable preference judgment, concretizing the distinction highlighted in Figure 1. For the hummingbird-feeding pair, the base quality term weakly favors B (-0.098). Defect contributions reinforce this preference: poor motion effects/transitions provide the strongest category-level contribution toward B (-0.291), outweighing the small contribution toward A from incorrect counts/spatial relations ($+0.007$). Together with the remaining categories (-0.139), these terms yield a total score difference of $q(A) - q(B) = -0.521$, and JACE selects B in agreement with the human judgment. Beyond identifying defects, this decomposition quantifies how their differences contribute to the comparison, including evidence favoring either candidate. Because these contributions are explicit terms in the scoring function, the explanation is directly tied to the model's decision computation. This provides an inspectable basis for candidate selection and highlights defect categories for further investigation during refinement.

6 Conclusion

We presented JACE, connecting weakly supervised defect attribution to human-preference-driven evaluation under separate supervision. Omni-modal and auxiliary diagnostic representations identify defects, while independent preferences determine how candidate differences affect quality. Combining these contributions with base quality ties explanations directly to scoring and comparison. For multimodal video and audio-video generation, the same mechanism supports single-video diagnosis and scoring, tie-aware pairwise comparison, and generator ranking. This combination of task coverage and evaluation levels provides interpretable evidence for candidate selection, cross-generator comparison, within-family iteration analysis, and problem identification.

AI USE STATEMENT

In this work, we used generative AI tools to assist with brainstorming research ideas and retrieving and summarizing related literature. During the experimental stage, we used these tools to help write and modify portions of the experimental code. During manuscript preparation, we used them to assist with drafting, translation, and language polishing. We take full responsibility for the final content of this work, including text, claims, and code produced with the assistance of generative AI.

REFERENCES

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

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: evaluating physical commonsense for video generation. In *Int. Conf. Learn. Represent.*, pp. 102075–102121, 2025.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, pp. 324–345, 1952.

ByteDance Seed Team. Seedance 2.0 official launch. `https://seed.bytedance.com/en/blog/seedance-2-0-official-launch`, 2026.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2818–2829, 2023.

Aditya Chinchure, Sahithya Ravi, Pushkar Shukla, Vered Shwartz, and Leonid Sigal. Spotlight: identifying and localizing video generation errors using VLMs. In *Eur. Conf. Comput. Vis.*, pp. 1–17, 2026.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 933–942, 2021.

Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep ConvNet for multi-label classification with partial labels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 647–657, 2019.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. TC-Bench: benchmarking temporal compositionality in conditional video generation. In *Findings of the Assoc. Comput. Linguist.*, pp. 4638–4662, 2025.

Google. Gemini 3.7 Flash. `https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash`, 2026.

Google Deepmind. Gemini Omni system card. `https://deepmind.google/models/gemini-omni/`, 2026.

Kaisi Guan, Zhengfeng Lai, Yuchong Sun, Peng Zhang, Wei Liu, Kieran Liu, Meng Cao, and Ruihua Song. ETVA: evaluation of text-to-video alignment via fine-grained question generation and answering. In *Int. Conf. Comput. Vis.*, pp. 21299–21309, 2025.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-Bench: human-aligned video generation benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18858–18868, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: building automatic metrics to simulate fine-grained human feedback for video generation. In *Conf. Empirical Methods NLP*, pp. 2105–2123, 2024.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. VideoScore2: think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 23345–23355, 2026.

Yinming Huang, Shuyuan Tu, Xi Yan, Zihan Yang, Jianhua Han, Hang Xu, Kaihang Pan, Yu-Gang Jiang, and Zuxuan Wu. VA-Judger: reward modeling from human preference feedback for joint video-audio generation. *arXiv preprint arXiv:2608.18607*, 2026a.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: comprehensive benchmark suite for video generative models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 21807–21818, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: comprehensive and versatile benchmark suite for video generative models. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 3268–3285, 2026b.

Ziheng Jia, Jiaying Qian, Zicheng Zhang, Xiaorong Zhu, Lancheng Gao, and Xiongkuo Min. RAVEN-Eval: rubric-guided automatic evaluation for AI video generation models based on LMM preference judgement. *arXiv preprint arXiv:2608.09111*, 2026.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. GenAI Arena: an open evaluation platform for generative models. In *Adv. Neural Inform. Process. Syst.*, pp. 79889–79908, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: all-in-one video creation and editing. In *Int. Conf. Comput. Vis.*, pp. 17191–17202, 2025.

Zhuoran Jin, Hongbang Yuan, Kejian Zhu, Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Omni-Reward: towards generalist omni-modal reward modeling with free-form preferences. In *Int. Conf. Learn. Represent.*, pp. 108481–108530, 2026.

M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, pp. 239–251, 1945.

KlingAI. Kling Video 3.0 system card. `http://kling.ai/feature/kling-video-3/`, 2026.

Sonia Laguna, Katarzyna Kobalczyk, Julia E. Vogt, and Mihaela van der Schaar. Interpretable reward modeling with active concept bottlenecks. *arXiv preprint arXiv:2507.04695*, 2025.

Jiesong Lian, Ruizhe Zhong, Zixiang Zhou, Xiaoyue Mi, Long Hu, Yuan Zhou, Qinglin Lu, Yixue Hao, and Junchi Yan. SoliReward: mitigating susceptibility to reward hacking and annotation noise in video generation reward models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 12798–12807, 2026.

Susan Liang, Chao Huang, Filippos Bellos, Yolo Yunlong Tang, Qianxiang Shen, Jing Bi, Luchuan Song, Zeliang Zhang, Jason Corso, and Chenliang Xu. Omni-Judge: can Omni-LLMs serve as human-aligned judges for text-conditioned audio-video generation? *arXiv preprint arXiv:2602.01623*, 2026.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Eur. Conf. Comput. Vis.*, pp. 366–384, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Adv. Neural Inform. Process. Syst.*, pp. 82155–82192, 2025.

Xiao Liu and Jiawei Zhang. AIGVE-MACS: unified multi-aspect commenting and scoring model for AI-generated video evaluation. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pp. 5179–5188, 2026.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: benchmarking and evaluating large video generation models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 22139–22149, 2024.

MiniMax. MiniMax H3: an open model breaking the boundaries between tasks and modalities. `https://www.minimax.io/blog/minimax-h3`, 2026.

OpenAI. Sora 2 system card, https://openai.com/index/sora-2-system-card/, 2025.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pp. 8748–8763, 2021.

Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 11030–11040, 2024.

Shivanshu Shekhar, Uttaran Bhattacharya, Raghavendra Addanki, Mehrab Tanjim, Somdeb Sarkhel, and Tong Zhang. GT-SVJ: generative-Transformer-based self-supervised video judge for efficient video reward modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9847–9858, 2026.

SpaceXAI. Grok Image Video 1.5 system card. `https://x.ai/news/grok-imagine-video-1-5/`, 2026.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: a comprehensive benchmark for compositional text-to-video generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8406–8416, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haibo Tong, Zhaoyang Wang, Zhaorun Chen, Haonian Ji, Shi Qiu, Siwei Han, Kexin Geng, Zhongkai Xue, Yiyang Zhou, Peng Xia, Mingyu Ding, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. MJ-Video: benchmarking and rewarding video generation with fine-grained video preference. In *Adv. Neural Inform. Process. Syst.*, pp. 79142–79173, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: a new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matthew Le, Piotr Dollár, Christoph Feichtenhofer, Ann Lee, and Wei-Ning Hsu. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 30172–30182, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded Transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 21486–21499, 2026a.

Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. AIGV-Assessor: benchmarking and evaluating the perceptual quality of text-to-video generation with LMM. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18869–18880, 2025b.

Qunzhong Wang, Jie Liu, Jiajun Liang, Yuanxing Zhang, Yilei Jiang, Yaozhi Zheng, Xintao Wang, Pengfei Wan, Xiangyu Yue, and Jiaheng Liu. VR-Thinker: boosting multimodal reward models through Think with Image reasoning. In *Int. Conf. Mach. Learn.*, 2026b.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *Eur. Conf. Comput. Vis.*, pp. 396–416, 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, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. In *Adv. Neural Inform. Process. Syst.*, pp. 159130–159157, 2025c.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025d.

Yuan Wang, Borui Liao, Huijuan Huang, Jinda Lu, Ouxiang Li, Kuien Liu, Meng Wang, and Xiang Wang. Thinking with frames: generative video distortion evaluation via frame reward model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4526–4536, 2026c.

Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, Weisi Lin, Wynne Hsu, Ying Shan, and Mike Zheng Shou. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024.

Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, Yan Zeng, and Weilin Huang. RewardDance: reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025a.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Xiele Wu, Zicheng Zhang, Mingtao Chen, Yixian Liu, Yiming Liu, Shushi Wang, Zhichao Hu, Yuhong Liu, Guangtao Zhai, and Xiaohong Liu. Q-Save: towards scoring and attribution for generated video evaluation. *arXiv preprint arXiv:2511.18825*, 2025b.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Shiyu Huang, Xiaotao Gu, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: fine-grained multi-dimensional human preference learning for image and video generation. In *Assoc. Adv. Artif. Intell.*, pp. 11269–11277, 2026.

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

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*, 2026.

Ziyun Zeng, Zixuan Wang, Yongsheng Yu, Hang Hua, and Jiebo Luo. VideoArgus: agentic rubric-grounded unified evaluation for video generation and editing. *arXiv preprint arXiv:2608.05485*, 2026.

Xuanyu Zhang, Weiqi Li, Shijie Zhao, Junlin Li, Li Zhang, and Jian Zhang. VQ-Insight: teaching VLMs for AI-generated video quality understanding via progressive visual reinforcement learning. In *Assoc. Adv. Artif. Intell.*, pp. 12870–12878, 2026.

Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, Xiaohong Liu, and Guangtao Zhai. Q-Eval-100K: evaluating visual quality and alignment level for text-to-vision content. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10621–10631, 2025.

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

A ATTRIBUTION TAXONOMY

Figure 5 groups the 46 defect categories by instruction following, motion and temporal coherence, visual quality, audio quality, and output anomalies. Each task uses the applicable subset, with reference- and editing-related categories assessed under the corresponding conditions; T2V excludes categories that require a reference or editing condition.

Annotations record salient problems within this subset, and a video can exhibit multiple defects. Unrecorded applicable categories remain unknown under sparse supervision. GOOD indicates that no salient defect was observed and provides a sample-level reference for defect learning; it is separate from the 46 defect categories.



Figure 5: **Defect attribution taxonomy.** The 46 categories form five diagnostic groups. Asterisks mark reference- or editing-dependent categories, which are inapplicable to T2V. GOOD is an additional sample-level label for no observed salient defects, separate from the 46 defects.

B METHOD DETAILS

B.1 FEATURE PREPROCESSING AND DIAGNOSTIC HEAD

Within $\mathrm{Head}_{\mathrm{attr}}$, L_2-normalized base features are centered and rescaled before a two-layer perceptron predicts defect logits:

$$\widetilde{h}_{\mathrm{base}}(x) = \frac{h_{\mathrm{base}}(x) - \mu_h}{s_h}. \tag{7}$$

The mean μ_h and centered root mean square s_h are computed over attribution-training samples and feature dimensions:

$$\mu_h = \frac{1}{n}\sum_{i=1}^{n} h_{\mathrm{base}}(x_i), \qquad s_h = \sqrt{\frac{1}{nD_h}\sum_{i=1}^{n} \|h_{\mathrm{base}}(x_i) - \mu_h\|_2^2}. \tag{8}$$

Here, n is the number of attribution-training samples and D_h is the base feature dimension. Centering removes the shared feature offset, and rescaling normalizes the diagnostic head's input scale. Both statistics remain fixed at inference. The head predicts category logits $z_{\mathrm{base}}(x)$ and the overall-good logit $z_{\mathrm{base},G}(x)$ from the same hidden representation.

Each auxiliary branch uses the same normalization procedure with its own training statistics. Its availability information records whether the corresponding content and features were obtained. The processed branches and availability information form $\widetilde{h}_{\mathrm{aux}}(x)$. The task encoding $\mathrm{onehot}(t)$ identifies the generation task and supplies context to the residual head. For the attribution objective, we use $\rho = 0.5$, $\lambda_G = 0.2$, and $\lambda_U = 0.05$.

B.2 RESIDUAL CORRECTION

The residual head is a two-layer perceptron with a zero-initialized output layer, so its initial predictions equal the base diagnosis. Base-head parameters and all encoders are frozen during correction training. Only category-level defect logits are corrected; the overall-good output is unchanged. The regularized objective is

$$\mathcal{L}_{\mathrm{res}} = \mathcal{L}_{\mathrm{attr}}(z_J, z_{\mathrm{base},G}) + \frac{\lambda_R}{nK}\sum_{i=1}^{n} \|\delta_\theta(x_i)\|_2^2. \tag{9}$$

Here, n is the number of correction-training samples and λ_R controls residual regularization. Recorded defects and overall-good references determine correction directions through the attribution objective, while the penalty constrains departures from the base logits. The complete diagnostic module is frozen after correction training.

B.3 PREFERENCE PAIR WEIGHTING

Preference training first fits the independent quality head from base-quality differences. It then freezes that head and the diagnostic module and learns a, w, and τ with Eq. 6. Pair weights γ_j are allocated by prompt and media-association group so that each group has equal total weight. This reduces the dominance of groups containing more repeated candidates. Feature and diagnostic-score standardization uses preference-training statistics, which remain fixed for evaluation.

C Implementation Details

This appendix describes the data partitions, training configuration, and evaluation metrics. JACE uses a frozen Qwen3-Omni omni-modal understanding encoder for the base representation; the primary auxiliary branch is PE-AV. The base diagnostic head predicts 46 defect categories and one overall-good (GOOD) logit. Residual correction modifies only defect outputs, while the base head and encoders remain frozen during correction.

C.1 Data Scale and Task Distribution

The attribution data cover text-to-video (T2V), image-to-video (I2V), reference-to-video (R2V), and video editing (V2V). The attribution development and confirmation sets support model selection and final checking, while preference data are held for quality-head and defect-weight learning. Table 4 reports the currently registered task counts.

Table 4: **Data scale and task distribution.** Counts are the registered records for each stage.

Data use	T2V	I2V	R2V	V2V	Total
Attribution training	26,522	10,000	9,997	9,771	56,290
Attribution development	∼256	∼256	∼256	∼256	∼1,024
Attribution confirmation	∼256	∼256	∼256	∼256	∼1,024
Preference training	12,896	4,211	5,006	4,025	26,138
Preference calibration	∼512	∼512	∼512	∼512	∼2,048
Preference evaluation	∼1,000	∼1,000	∼1,000	∼1,000	∼4,000
Model-pair battle count	∼4,480	∼4,480	∼4,480	∼4,480	∼17,920

C.2 Backbone Selection and Training Configuration

Base backbone. JACE uses Qwen3-Omni, an open-source omni-modal understanding model, as its base backbone (Xu et al., 2025). Its joint understanding of text, images, audio, and video allows generation conditions and candidate content to be considered together for audio-video evaluation. Its language and multimodal semantic understanding also provides a basis for assessing whether generated content satisfies the instructions, supporting attribution of instruction-following defects.

Auxiliary representation. We use PE-AV as the primary auxiliary encoder (Vyas et al., 2026). Among the candidate auxiliary encoders evaluated in Table 6, PE-AV achieves the highest task-mean attribution recall and GSB accuracy. We therefore use its representations to complement the backbone's audiovisual perception and support residual diagnostic correction.

Training setup. The base backbone and auxiliary encoders remain frozen, while the diagnostic head and residual module are optimized with AdamW. The residual output layer is initialized to zero, and the base diagnostic head is fixed during correction. Preference learning first trains the independent base-quality head, then fixes both that head and the diagnostic module to learn defect weights. The tie threshold is selected by three-class macro-F1 on the preference calibration set. Normalization statistics are estimated from the corresponding training sets and remain fixed during evaluation.

Other settings. Task conditions, generated videos, and available audio tracks follow the task-specific preprocessing protocol. VideoReward and VideoScore2 use 2-fps frame sampling; the other encoders use their respective preprocessing configurations.

C.3 Evaluation Protocol and Metrics

Evaluation covers video-level attribution, pairwise comparison, and generator ranking. Recall@3 measures retrieval of observed defects; non-tie and GSB accuracy measure candidate choices; Tie-F1 measures tie recognition. At the generator level, the Bradley–Terry (BT) model aggregates battle

Table 5: **Implementation and training configuration.**

Component	Configuration
Encoders and features	Frozen Qwen3-Omni; 2,048-dimensional base feature; PE-AV as the primary auxiliary representation.
Diagnostic head	$2048 \to 256$ (GELU) $\to 47$: 46 defect logits and one GOOD logit; centering and scale statistics estimated on attribution training data.
Sparse-attribution objective	$\rho = 0.5$, $\lambda_G = 0.2$, and $\lambda_U = 0.05$; unannotated categories are not uniformly treated as negatives.
Residual correction	Two-layer residual head with zero-initialized output; $\lambda_R = 0.01$; base head and encoders frozen.
Attribution optimization	AdamW, learning rate 10^{-4}, weight decay 0.01.
Quality and preference calibration	Independent linear quality head; quality head fixed before defect-weight fitting; defect weights learned on independent preference data, with the tie threshold selected by three-class macro-F1 on calibration data.

evidence into strengths, and Kendall's τ_b measures agreement with human rankings. Recall, accuracy, and F1 are reported as percentages. Higher values are better for all reported metrics.

Attribution Recall@3. For video x_i, let S_i denote its observed defect set and $P_i^{(3)}$ the three highest-scoring categories within the applicable set $\mathcal{K}(x_i)$. Let $\mathcal{I}_t$ index evaluation videos of task t with $|S_i| > 0$. Task-specific and task-mean recall are

$$R_t@3 = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \frac{|S_i \cap P_i^{(3)}|}{|S_i|}, \qquad R_{\text{task}}@3 = \frac{1}{4} \sum_{t \in \{\text{T2V,I2V,R2V,V2V}\}} R_t@3. \tag{10}$$

Recall is first computed per video, averaged within each task, and then averaged equally across the four tasks. This prevents larger tasks from dominating the aggregate. GOOD examples and videos without observed defects are excluded from the recall denominator. Category-mean recall pools each category's positive examples across tasks before averaging equally over supported categories. With $\mathcal{I}_k = \{i : k \in S_i\}$ and $\mathcal{K}_+ = \{k : |\mathcal{I}_k| > 0\}$, it is

$$R_{\text{class}}@3 = \frac{1}{|\mathcal{K}_+|} \sum_{k \in \mathcal{K}_+} \frac{\sum_{i \in \mathcal{I}_k} \mathbf{1}[k \in P_i^{(3)}]}{|\mathcal{I}_k|}. \tag{11}$$

Here $\mathbf{1}[\cdot]$ is the indicator function. Task averaging measures performance across tasks; category averaging gives rare and frequent defects equal weight. Both evaluate retrieval of recorded defects.

Non-tie and GSB accuracy. Human labels $y_j \in \{A, T, B\}$ indicate an A win, a tie, and a B win, corresponding to Good/Same/Bad relative to A. Let $\Delta_j = q(x_j^A) - q(x_j^B)$ and $\mathcal{J}_{\text{nt}} = \{j : y_j \neq T\}$. The binary decision $\widetilde{y}_j$ is A when $\Delta_j > 0$ and B otherwise; exactly equal scores are assigned to B. The three-way decision $\widehat{y}_j$ follows the prespecified decoding rule. For JACE, $|\Delta_j| \leq \kappa$ yields a tie and the sign determines the winner outside this interval, where κ is chosen on preference calibration data and fixed for evaluation. The two accuracies are

$$\text{Acc}_{\text{nt}} = \frac{1}{|\mathcal{J}_{\text{nt}}|} \sum_{j \in \mathcal{J}_{\text{nt}}} \mathbf{1}[\widetilde{y}_j = y_j], \qquad \text{Acc}_{\text{GSB}} = \frac{1}{N_{\text{pair}}} \sum_{j=1}^{N_{\text{pair}}} \mathbf{1}[\widehat{y}_j = y_j], \tag{12}$$

where N_{pair} is the number of evaluation pairs. Non-tie accuracy tests the preference direction on human win/loss examples without applying the tie interval. GSB accuracy evaluates all three outcomes under their empirical frequencies in the evaluation set.

Tie-F1. Treating a tie as the positive class, TP_T counts correctly predicted human ties, FP_T counts predicted ties with human win/loss labels, and FN_T counts human ties predicted as wins or losses. Then

$$F_{1,T} = \frac{2TP_T}{2TP_T + FP_T + FN_T}. \tag{13}$$

This metric balances tie precision and recall, distinguishing successful tie recognition from excessive tie prediction. It complements overall GSB accuracy and is reported only on sets containing human ties.

BT generator ranking. For generators $\mathcal{G} = \{G_1, \ldots, G_L\}$ with strengths β_i, the BT model represents pairwise win probabilities as (Bradley & Terry, 1952)

$$\pi_{ab} = P(G_a \succ G_b) = \frac{\exp(\beta_a)}{\exp(\beta_a) + \exp(\beta_b)} = \sigma(\beta_a - \beta_b). \tag{14}$$

Let a_j, b_j denote the generators in battle j. Human win credit is $u_j^H = \mathbf{1}[y_j = A] + \frac{1}{2}\mathbf{1}[y_j = T]$, assigning $1, \frac{1}{2}, 0$ to a win, tie, and loss. For probabilistic evaluators, expected win credit is $u_j^J = p_{A,j} + \frac{1}{2}p_{T,j}$; methods returning only discrete decisions use the same outcome-to-credit mapping. Human and evaluator strengths are fitted separately by

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{\,s} &= \arg\min_{\sum_{i=1}^{L}\beta_i = 0} \mathcal{L}_{\mathrm{BT}}(\boldsymbol{\beta}; u^s), \quad s \in \{H, J\}, \\ \mathcal{L}_{\mathrm{BT}}(\boldsymbol{\beta}; u) &= -\frac{1}{M}\sum_{j=1}^{M}\left[u_j \log \pi_{a_j b_j} + (1 - u_j)\log(1 - \pi_{a_j b_j})\right] + \lambda_{\mathrm{BT}}\|\boldsymbol{\beta}\|_2^2. \end{aligned} \tag{15}$$

Here M is the number of ranking battles and λ_{BT} is the regularization coefficient. The zero-mean constraint fixes the common offset, while regularization stabilizes estimation from finite battles. Generators are ordered by decreasing $\widehat{\beta}_i$, using the same generator set and comparison graph for human and evaluator judgments. BT aggregates local battles into generator-level strengths for relative ranking under a common protocol.

Kendall's τ_b. We compare human and evaluator BT strengths over the same generators. Let C and D count concordant and discordant generator pairs, and let T_J and T_H count pairs tied only in the evaluator and only in the human ordering, respectively. Ranking agreement is (Kendall, 1945)

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_J)(C + D + T_H)}}. \tag{16}$$

Pairs tied in both orderings contribute to none of these four counts. These are comparisons between generators, not individual video battles; ties are determined from unrounded BT strengths. The coefficient lies in $[-1, 1]$: values near 1 indicate agreement, -1 indicates strictly reversed orderings, and values near zero indicate that concordant and discordant relations approximately balance. For disconnected comparison graphs, coefficients are computed within connected components and aggregated by their numbers of comparable generator pairs. Components with a zero denominator have no reported correlation.

D DETAILED RESULTS

Table 6 gives the numerical results shown in Figure 3. It reports each task separately, along with task-mean attribution and preference performance. PE-AV (audio), PE-AV (visual), and PE-AV (joint) use the respective representations; PE-AV (all) concatenates all three. Qwen3-Omni Head uses a trained diagnostic head, distinct from the original model in Table 1.

Table 6: **Backbone-level attribution and preference results.** Task columns report Recall@3; Mean is the equally weighted task average. Mean GSB reports task-mean three-way accuracy. All values are percentages. PE-AV (all) concatenates audio, visual, and joint audio-video representations. Bold denotes the best reported result.

Configuration	T2V	I2V	R2V	V2V	Mean	Mean GSB
CLIP-L/14 (Radford et al., 2021)	43.10	24.87	33.47	36.59	34.51	38.05
CLIP-H/14 (Cherti et al., 2023)	42.06	31.82	31.07	39.19	36.04	39.73
SigLIP2 (Tschannen et al., 2025)	44.08	29.86	35.73	35.94	36.40	36.97
HPSv2.0 (Wu et al., 2023)	44.40	32.62	33.19	41.15	37.84	41.29
HPSv2.1 (Wu et al., 2023)	44.14	34.49	32.77	41.54	38.23	39.01
InternVideo2 (Wang et al., 2024a)	45.18	32.53	33.19	31.51	35.60	35.77
V-JEPA 2 (Assran et al., 2025)	42.25	35.74	36.58	37.63	38.05	40.45
PE-AV (audio) (Vyas et al., 2026)	39.10	26.70	27.90	29.50	30.80	45.13
PE-AV (visual) (Vyas et al., 2026)	47.30	30.60	32.50	34.80	36.30	44.17
PE-AV (joint) (Vyas et al., 2026)	49.10	31.70	34.90	36.00	37.92	46.57
PE-AV (all) (Vyas et al., 2026)	52.15	33.07	36.16	37.37	39.69	48.01
VGGT (Wang et al., 2025a)	28.45	21.30	31.64	26.95	27.09	34.93
VGGT-Ω (256) (Wang et al., 2026a)	33.59	29.86	25.42	24.22	28.27	36.73
VGGT-Ω (512) (Wang et al., 2026a)	33.33	27.99	26.55	23.05	27.73	34.09
Qwen3-Omni Head (Xu et al., 2025)	47.40	38.15	31.64	54.04	42.80	41.77
JACE (Ours)	**61.33**	**49.59**	**48.26**	**66.90**	**56.52**	**58.31**