REWARDVERSE: RUBRIC-GUIDED POLICY OPTIMIZATION FOR VIDEO REWARD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) is vital for optimizing video generation models, with a robust reward model (RM) serving as the cornerstone. However, existing video reward models often produce unstable scalar scores because they directly map complex, subjective video quality into a single score without explicit evaluation criteria. This leads to *scalar drift*, where the scoring scale collapses or shifts across different prompts, making the reward unreliable for RL. Drawing inspiration from professional human annotation engineering, we address this problem with **RewardVerse**, a rubric-based video reward framework that introduces a *dynamic rubric* as an intermediate representation between the evaluation query and the scorer. Instead of unconstrained direct scoring, RewardVerse first generates explicit evaluation criteria and then performs rubric-guided scoring, providing a stable semantic anchor that mitigates scalar drift. To efficiently optimize this collaborative pipeline, we propose **R**ubric-**G**uided **P**olicy **O**ptimization (**RGPO**), a two-stage training algorithm. RGPO first warms up the scorer using self-evolving seed rubrics and then jointly optimizes the rubric generator to produce query-adaptive evaluation criteria while continuously aligning the scorer with human ratings. Extensive experiments on the 16-dimensional EvalVerse benchmark and external datasets demonstrate that RewardVerse mitigates scalar drift, achieves state-of-the-art performance on both pointwise and pairwise evaluation, and provides a robust and interpretable reward signal for RL in video generation.

1 INTRODUCTION

Generative video models have achieved remarkable progress in recent years. As these models continue to improve, reinforcement learning (RL) has become an important paradigm for aligning generated videos with human preferences, with a robust reward model (RM) serving as the cornerstone. Despite recent progress, existing video RMs still struggle to produce stable pointwise rewards. Most methods directly map a generated video to a single scalar score, either through discriminative regression or generative reasoning. Since video quality is inherently subjective and multi-dimensional (Wang et al., 2026a), this direct scoring process lacks explicit evaluation criteria. Consequently, the scoring scale often becomes unstable, causing scores to collapse into a narrow range (Jin et al., 2026) or shift under different prompts and contexts. We refer to this phenomenon as *scalar drift*.

To address this challenge, we draw inspiration from professional human annotators, who rarely assign scores directly. Instead, they first decompose an evaluation task into explicit criteria before producing a final judgment, creating a stable semantic anchor that maintains consistency across different samples (Tong et al., 2025; Xu et al., 2026a). Existing video reward models largely omit this explicit criterion-setting stage. By forcing a model to map queries to scores in an unconstrained step, their internal scoring standards shift across different prompts, leading to unstable rewards.

Based on this insight, we propose **RewardVerse**, a highly data-efficient video evaluation framework that natively supports both pointwise and pairwise evaluation. As compared in Figure 1, instead of directly mapping a video to a score, we insert a *dynamic rubric* as an intermediate representation to decouple evaluation into standard generation and objective execution. Specifically, the rubric generator first decomposes the holistic text query into explicit evaluation themes, weights, and scoring tips. The scorer then evaluates the video against these rubrics independently, rather than relying

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.

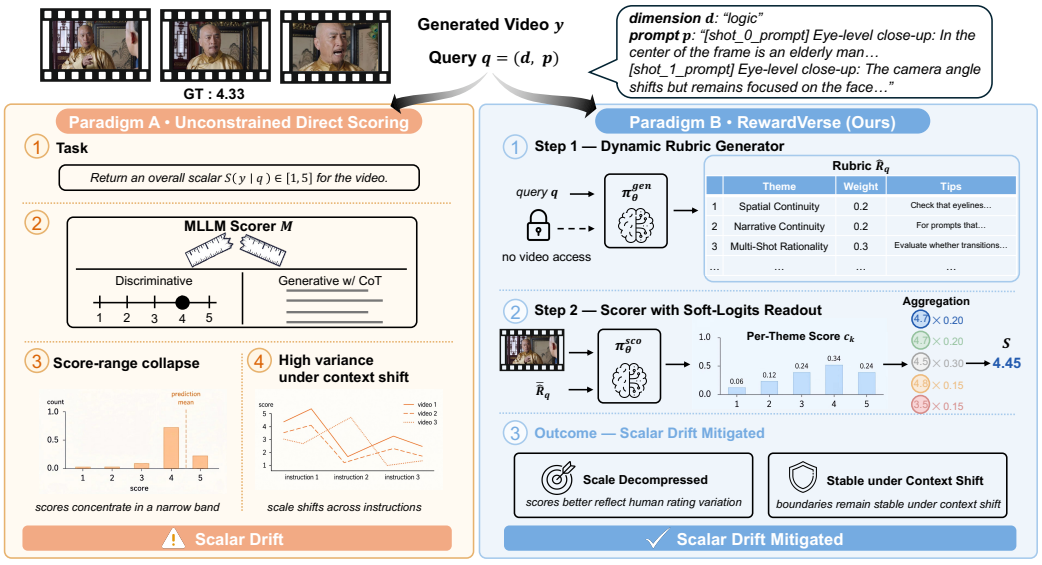


Figure 1: A comparison of different video evaluation paradigms. Compared to unconstrained discriminative and generative scoring that suffer from severe scalar drift, RewardVerse inserts a dynamic rubric as an intermediate representation to anchor the scoring process.

on an implicit internal standard. This decoupling provides a clear cognitive anchor that mitigates score-range collapse and improves robustness under context shift.

To efficiently learn dynamic rubrics and optimize the collaborative pipeline with minimal preference pairs and without prior supervised fine-tuning (SFT), we introduce a two-stage training paradigm called **Rubric-Guided Policy Optimization** (**RGPO**) under Group Relative Policy Optimization (GRPO). We first warm up the scorer using self-evolving seed rubrics synthesized offline. Then, we jointly optimize the rubric generator to produce query-adaptive evaluation criteria while continuously aligning the scorer with human ratings. Through this optimization, RewardVerse effectively mitigates scalar drift, requiring only 30 preference pairs per dimension to deliver state-of-the-art performance and provide a robust, highly interpretable reward for downstream generation models.

Our main contributions are summarized as follows:

- **Rubric-as-Reward for Video RMs:** To our knowledge, we are the first to introduce the rubric-as-reward paradigm to video reward modeling. We propose **RewardVerse**, a framework that employs a dynamic rubric as an intermediate representation to stabilize pointwise scores and mitigate scalar drift.
- **Data-Efficient Joint Policy Optimization:** We design **RGPO** (Rubric-Guided Policy Optimization), a two-stage training algorithm that learns a dynamic rubric generator from only 30 preference pairs per dimension, while maintaining a human-aligned scorer.
- **Systematic Empirical Validation:** We conduct extensive evaluations on the multi-dimensional EvalVerse-adapted dataset and external benchmarks. RewardVerse achieves state-of-the-art evaluation performance and serves as a stable, interpretable reward for downstream video generation optimization.

2 Related Work

The Rubric-as-Reward Paradigm. Rubrics decompose quality evaluation into explicit criteria corresponding to concrete aspects of the desired output, providing a natural handle for open-ended, non-verifiable tasks. In LLM and autonomous agent domains, recent works bring rubrics into RL for subjective tasks such as medicine, writing, and research (e.g., RaR (Gunjal et al., 2025), RLCF (Viswanathan et al., 2026), HealthBench (Arora et al., 2025), RLCER (Sheng et al., 2026), OpenRubrics (Liu et al., 2026b), Rubric-ARM (Xu et al., 2026b), and DR Tulu (Shao et al., 2025)). Similarly, in image generation and editing, recent works use explicit rubrics, checklists, or decom-

posed criteria to evaluate and optimize visual outputs (e.g., AlphaGRPO (Huang et al., 2026), Edit-R1 (Guo et al., 2026), EditReward-Compass (Bai et al., 2026), ARR-RPO (Tian et al., 2026), and RewardHarness (Zhang et al., 2026)). These works establish explicit criteria as useful reward signals across open-ended generation tasks. In contrast, RewardVerse makes the rubric a dynamic, learned intermediate representation: its themes, weights, and scoring tips are generated from each evaluation query without access to the candidate video, and are jointly optimized with the scorer to stabilize pointwise video rewards.

Video Reward Modeling. Existing video reward models generally fall into two categories: discriminative and generative. Discriminative models (e.g., VideoScore (He et al., 2024), VideoReward (Liu et al., 2026a)) train regression layers on top of visual encoders to output continuous scalar scores. Generative models (e.g., UnifiedReward (Wang et al., 2025), VideoScore2 (He et al., 2025)) prompt multi-modal LLMs to output scores directly. Because these models score videos directly without explicit anchors, they often suffer from scalar drift. As analyzed in Section 3, scalar drift causes scores to collapse into a narrow range or change wildly across prompts, making them unstable for online RL training. By jointly training rubric generation and scoring under our structured rubric bottleneck, RewardVerse mitigates scalar drift, achieves high data efficiency, and provides a stable reward signal for downstream video generation models.

3 The Phenomenon of Scalar Drift in Video Evaluation

To establish the empirical foundations of our work, we conduct an analytical study of video evaluation behaviors in modern Multi-modal Large Language Models (MLLMs). We focus on dissecting scalar drift under unconstrained direct scoring, analyzing how rubric-guided scoring alleviates this problem, and identifying the remaining limitations that motivate joint policy optimization.

3.1 Defining Scalar Drift

We consider an MLLM $\mathcal{M}$ tasked with returning a single overall scalar reward $S(y \mid q) \in [1, 5]$ for a generated video y given a textual query q. Because video quality assessment is multi-dimensional and largely subjective, this direct task is *unconstrained*: the model has no explicit anchor for what each point on the 1–5 scale should mean. Under this setting, the internal scoring rubric drifts in two interacting ways, which together we define as **scalar drift**:

- **Score-range collapse.** Pointwise scores concentrate in a narrow "safe" high band, shrinking the score gap between preferred and non-preferred videos and leaving downstream RL with near-flat gradients.
- **High variance under context shift.** Without explicit anchors, the scoring scale may shift across paraphrased instructions or different input orderings, producing inconsistent scores or pairwise decisions for the same content.

A workable pointwise reward must therefore (a) preserve scale resolution across the population of videos, and (b) be robust to context phrasing and ordering. See Appendix A.1 for further analysis and Appendix A.1.3 for score distributions across existing reward models.

3.2 Empirical Study: Rubric is Necessary but Not Sufficient

To isolate which design choice cures which symptom of scalar drift, we construct a 2×2 ablation design space covering rubric injection (presence vs. absence) and decoding format (natural-language generation vs. soft-logits continuous expectation). This yields four variants evaluated on a hold-out pool of 240 pointwise videos with human labels (details in Appendix A.1.1): 1) **V1 (No Rubric + Natural Float)**; 2) **V2 (No Rubric + Soft-Logits)**; 3) **V3 (With Rubric + Soft-Logits, Deployed)**; and 4) **V4 (With Rubric + Natural Float)**. Figure 2 reveals three key insights that drive our downstream framework design:

Cognitive Anchoring via Rubrics. Without rubric guidance, unconstrained direct scoring (V1 and V2) suffers from severe score saturation and narrow variance. Injecting a rubric as a cognitive anchor substantially expands these distributions. For example, holding the soft-logits format fixed, adding

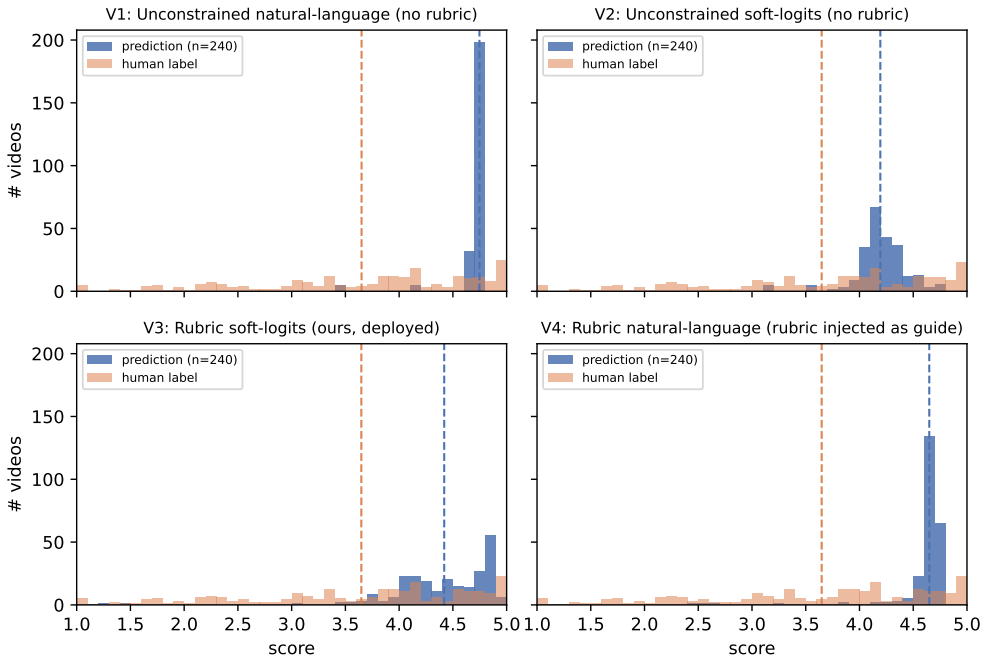


Figure 2: Pointwise score distributions of four scoring variants. Blue bars are predictions, orange bars are human labels, and dashed lines show the means. **V1** and **V2** show score-range collapse, while **V4** remains affected by natural-language bias. Only **V3** (rubrics with soft-logits) effectively spreads predictions, but it still deviates from human ratings, motivating RGPO in Section 4.

the rubric ($V2 \rightarrow V3$) expands the scoring standard deviation by $+82.6\%$ (σ: $0.258 \rightarrow 0.471$) for Qwen2.5-VL-7B. Under natural text generation, the rubric acts as a vital regularizer: it expands the collapsed standard deviation of Qwen2.5-VL-7B ($V1 \rightarrow V4$: $0.203 \rightarrow 0.275$), while reducing the excessively noisy and unguided variance of Gemini-3.1-Pro (reducing σ from 1.500 to 1.158), bringing their score variance closer to that of human judgments (see Appendix A.1.2).

Continuous Scale Decompression via Soft-Logits. Crucially, however, the rubric *alone* is insufficient if paired with natural text generation. In V4, two-thirds of the comparison pairs still tie because natural-language outputs remain anchored to the model's high-band linguistic prior. Further mitigating scalar drift requires coupling the rubric with a continuous, logit-based readout mechanism. Extracting a continuous expected score directly from the logits of rating tokens (soft-logits) successfully bypasses the model's generation bias (formulated in Section 4). Consequently, pairing a rubric with soft-logits (V3) better captures the variation in human ratings and reduces score-range collapse. For our deployment, we choose this pointwise multi-theme soft-logits protocol (V3) as our primary configuration; detailed protocol comparisons are provided in Appendix A.1.5.

The Remaining Limitations. Although rubric-guided soft-logits scoring mitigates scalar drift, two limitations remain. First, static rubrics cannot adapt their themes, tips, and weights to different queries. Second, the predicted score distribution may still deviate from human ratings, even for stronger proprietary MLLMs such as Gemini-3.1-Pro (Appendix A.1.2). These results show that rubric prompting alone cannot fully resolve scalar drift, motivating RGPO to jointly optimize rubric generation and scoring. As verified in Appendix A.1.7, the complete RGPO training further improves both score alignment and correlation with human ratings.

4 Methodology

Rather than treating the rubric as a fixed prompt scaffold, we make it a learnable intermediate representation. This allows the model to generate query-adaptive evaluation criteria while keeping the scorer aligned with human ratings. This joint optimization further mitigates the scalar drift that remains after rubric-guided scoring. Building on this formulation, we present **RewardVerse**, which parameterizes adaptive rubric generation and human-aligned scoring within a dual-role policy model

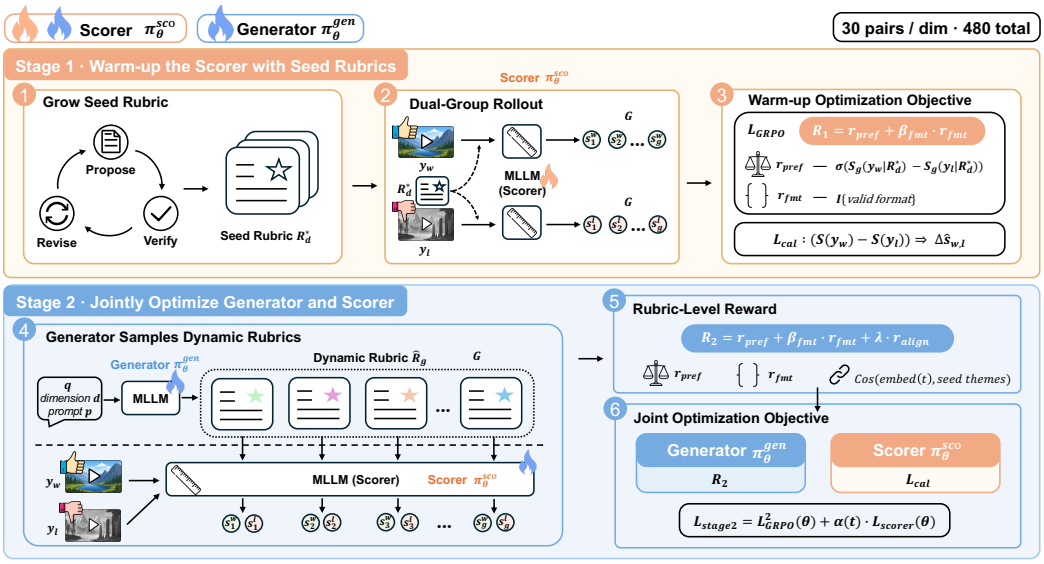


Figure 3: Overview of RewardVerse and its two-stage RGPO training. Stage 1 warms up the scorer with seed rubrics using preference and format rewards plus a calibration loss. Stage 2 learns the dynamic rubric generator while continuing scorer calibration.

π_θ. We optimize the model via **Rubric-Guided Policy Optimization (RGPO)**, a two-stage training algorithm. Stage 1 warms up the scorer using self-evolving seed rubrics (§4.2), and Stage 2 optimizes the dynamic rubric generator while continuing to calibrate the scorer (§4.3).

4.1 RewardVerse Pipeline and Formulation

Inputs. An evaluation query $q = (d, p)$ pairs a target dimension $d \in \mathcal{D}$ with the text prompt p used to synthesize the candidate video y. Each video is scored independently during both training and inference, avoiding the pairwise position bias analyzed in Appendix A.1.6.

Module 1 — Dynamic Rubric Generator π_θ^{gen}. The policy first acts as a generator that parses the query q and emits a dynamic rubric under a fixed schema:

$$\hat{R}_q \sim \pi_\theta^{\text{gen}}(\cdot \mid q), \qquad \hat{R}_q = \{(t_k, w_k, T_k)\}_{k=1}^{K}, \tag{1}$$

where t_k is a theme, $w_k \in (0, 1)$ is its weight with $\sum_k w_k = 1$, and T_k is a set of execution tips based on queries.

Candidate-Independent Rubric Generation: Notably, the generator takes only the query q as input without seeing the video y. This decoupling prevents the generator from producing biased, video-dependent rubrics (e.g., being overly lenient to low-quality videos or overly strict to high-quality ones), safeguarding evaluation objectivity while keeping the criteria query-adaptive.

Module 2 — Scorer π_θ^{sco} with Soft-Logits Readout. Given a rubric R, the policy emits one score per theme. To avoid parsing digits under JSON-constrained decoding, at each score slot we read the logits l_v of the five rating tokens $v \in \{1, \ldots, 5\}$ (corresponding to the string representations of integers "1" to "5" in the vocabulary) and take the expected value:

$$c_k(y \mid R) = \sum_{v=1}^{5} v \cdot P_\theta\big(v \mid y, t_k, T_k\big), \qquad P_\theta\big(v \mid y, t_k, T_k\big) = \frac{\exp(l_v)}{\sum_{v'=1}^{5} \exp(l_{v'})}. \tag{2}$$

The pointwise reward aggregates theme scores by their weights:

$$S(y \mid R) = \sum_{k=1}^{K} w_k \cdot c_k(y \mid R). \tag{3}$$

The generator and scorer share the same backbone and are distinguished by role-specific prompts. In Stage 2, the shared policy is trained with separate signals for the two roles. The main prompt templates are provided in Appendix A.4.1.

4.2 STAGE 1: SEED-GUIDED SCORER WARM-UP

Learning dynamic rubric generation and calibrated scoring at the same time from scratch is highly under-determined. A poor generator produces chaotic rubrics that disrupt scorer learning. Therefore, Stage 1 uses fixed seed rubrics $R_d^\star$ and updates the shared policy through the scorer role, establishing human-aligned score margins before dynamic rubric generation is introduced.

Seed Rubric Self-Evolving. For each dimension d, we build a seed rubric $R_d^\star$ through an offline loop driven by a frontier MLLM $\mathcal{M}$ (Gemini-3.1-Pro) over 30 preference pairs per dimension:

- **Propose ($\mathcal{M}^{\text{gen}}$).** Given (q, y_w, y_l), $\mathcal{M}$ drafts a candidate rubric $\tilde{R}$ that explains why $y_w \succ y_l$.
- **Verify ($\mathcal{M}^{\text{ver}}$).** $\mathcal{M}$ scores both videos with $\tilde{R}$; $\tilde{R}$ is accepted only if it recovers $y_w \succ y_l$.
- **Revise ($\mathcal{M}^{\text{ref}}$).** On failure, $\mathcal{M}$ refines $\tilde{R}$ using a critique of the misalignment.

Verified rubrics are pooled into $\mathcal{V}_d$, deduplicated by a two-level Jaccard filter, and reduced to $M{=}5$ representative entries by an MCR^2 sampler (Yu et al., 2020), giving $R_d^\star = \{r_{d,1}^\star, \dots, r_{d,M}^\star\}$.

Scorer Optimization via Policy Gradient. For each pair $(q, y_w, y_l) \sim \mathcal{D}_{\text{train}}$ under the seed rubric $R_d^\star$, the scorer produces G joint scoring samples. Each sample contains one completion $a_{w,g}$ for the preferred video and one completion $a_{l,g}$ for the non-preferred video, sampled from $\pi_{\theta_{\text{old}}}^{\text{sco}}$ for $g = 1, \dots, G$. Each completion pair g is scored by a pairwise preference-driven reward:

$$r_{\text{pref}}(a_{w,g}, a_{l,g}) = \sigma\big(S_g(y_w \mid R_d^\star) - S_g(y_l \mid R_d^\star)\big), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, and S_g represents the pointwise score (equation 3) decoded from the corresponding completion $a_{x,g}$ ($x \in \{w, l\}$). The full warm-up reward is defined as:

$$\mathcal{R}_1(a_{w,g}, a_{l,g}) = r_{\text{pref}}(a_{w,g}, a_{l,g}) + \beta_{\text{fmt}}\left(r_{\text{fmt}}(a_{w,g}) + r_{\text{fmt}}(a_{l,g})\right), \tag{5}$$

where β_{fmt} controls the contribution of the format reward, and $r_{\text{fmt}}(a_{x,g}) \in \{0, 1\}$ is a binary indicator checking whether the completion conforms to the requested scoring format.

Following GRPO (Shao et al., 2024), the joint reward of each pair is standardized across the group to yield the trajectory-level advantage $\hat{A}_g = \big(\mathcal{R}_1(a_{w,g}, a_{l,g}) - \text{mean}(\mathcal{R}_1)\big)/\text{std}(\mathcal{R}_1)$, which is broadcast to all tokens in both $a_{w,g}$ and $a_{l,g}$. Denoting the per-token importance ratio between the current and behavior scorers by $\rho_{x,g,t}(\theta) = \frac{\pi_\theta^{\text{sco}}(a_{x,g,t}|y_x, R_d^\star)}{\pi_{\theta_{\text{old}}}^{\text{sco}}(a_{x,g,t}|y_x, R_d^\star)}$, the Stage-1 GRPO objective optimizes both preferred and non-preferred completions:

$$\begin{aligned}\mathcal{L}_{\text{GRPO}}^{(1)}(\theta) = -\frac{1}{G}\sum_{g=1}^{G}\frac{1}{2}\sum_{x\in\{w,l\}}\frac{1}{|a_{x,g}|}\sum_{t=1}^{|a_{x,g}|}\Big[\min\big(\rho_{x,g,t}(\theta)\hat{A}_g,\ \text{clip}(\rho_{x,g,t}(\theta), 1{-}\epsilon, 1{+}\epsilon)\hat{A}_g\big) \\ - \beta_{\text{KL}}\,\mathbb{D}_{\text{KL}}\big[\pi_\theta^{\text{sco}} \,\big\|\, \pi_{\text{ref}}^{\text{sco}}\big]_{x,g,t}\Big].\end{aligned} \tag{6}$$

Here, ϵ is the clipping coefficient, β_{KL} controls the KL penalty, and $D_{\text{KL}}[\cdot\|\cdot]_{x,g,t}$ is the per-token KL penalty against a frozen reference policy $\pi_{\text{ref}}^{\text{sco}}$.

Human-Aligned Margin Calibration. Although the preference reward encourages the scorer to rank the preferred video above the non-preferred one, it does not control the magnitude of the predicted score difference. We therefore add a calibration loss based on the human score margin:

$$\mathcal{L}_{\text{cal}}(\theta \mid R) = \frac{1}{S_{\max} - S_{\min}}\Big|\big(S(y_w \mid R) - S(y_l \mid R)\big) - \Delta\hat{s}_{w,l}\Big|, \tag{7}$$

where $S_{\max} = 5$, $S_{\min} = 1$, and $\Delta\hat{s}_{w,l} \in [0, 4]$ is the target score margin derived from human scores. This loss calibrates the predicted score difference against the human-annotated margin, preventing arbitrary compression or inflation of reward gaps. We focus on relative rather than absolute calibration, since reward models mainly require reliable preference ordering and meaningful score differences, while direct absolute score regression can easily overfit in our low-data setting. The overall Stage-1 objective combines the GRPO loss with this calibration term:

$$\mathcal{L}_{\text{stage1}}(\theta) = \mathcal{L}_{\text{GRPO}}^{(1)}(\theta) + \lambda_{\text{cal}}\,\mathcal{L}_{\text{cal}}(\theta \mid R_d^\star). \tag{8}$$

4.3 Stage 2: Joint Policy Optimization via RGPO

Stage 2 learns the rubric generator to produce evaluation criteria tailored to each query. For each triple (q, y_w, y_l), the generator samples G dynamic rubrics $\{\hat{R}_g\}_{g=1}^{G} \sim \pi_{\theta_{\text{old}}}^{\text{gen}}(\cdot \mid q)$. Each rubric is judged by how well its induced pointwise scores separate the preferred and non-preferred videos:

$$r_{\text{pref}}(\hat{R}_g) = \sigma\big(S_g(y_w \mid \hat{R}_g) - S_g(y_l \mid \hat{R}_g)\big). \tag{9}$$

To keep the generated rubric aligned with the target dimension and enforce the required output format, we define the Stage-2 reward as:

$$\mathcal{R}_2(\hat{R}_g) = r_{\text{pref}}(\hat{R}_g) + \beta\, r_{\text{fmt}}(\hat{R}_g) + \lambda\, r_{\text{align}}(\hat{R}_g), \tag{10}$$

where $r_{\text{align}}(\hat{R}_g)$ measures the cosine similarity between the BGE-M3 embeddings (Chen et al., 2024) of the generated and seed themes. Here, λ and β are fixed coefficients for rubric alignment and format regularization, respectively. $r_{\text{fmt}}(\hat{R}_g)$ checks whether the output conforms to the structured JSON schema and whether the rubric weights sum to one.

The composite reward of each generated rubric is standardized across the group to yield the rubric-level advantage $\hat{A}_g = \big(\mathcal{R}_2(\hat{R}_g) - \text{mean}(\mathcal{R}_2)\big)/\text{std}(\mathcal{R}_2)$, which is broadcast to all tokens in the generated rubric $\hat{R}_g$. The generator is optimized using the same clipped GRPO objective as in equation 6, with $\hat{R}_g$ as the optimized sequence and $\rho_{g,t}(\theta) = \frac{\pi_{\theta}^{\text{gen}}(\hat{R}_{g,t}|q)}{\pi_{\theta_{\text{old}}}^{\text{gen}}(\hat{R}_{g,t}|q)}$ as the per-token importance ratio. We denote this generator objective by $\mathcal{L}_{\text{GRPO}}^{(2)}(\theta)$; its full form is provided in Appendix A.4.2. As the rubric distribution changes during Stage 2, we retain scorer calibration to keep the scores aligned with human margins under the sampled rubrics.

Asymmetric Optimization Signals. If the scorer were optimized directly by the relative reward $\mathcal{R}_2$, it could increase the reward simply by enlarging the score difference between the preferred and non-preferred videos, without improving score calibration. We therefore use different optimization signals for the two roles. The generator is optimized by the rubric-level GRPO objective, while the scorer receives no direct policy-gradient update from $\mathcal{R}_2$ and is instead optimized by the margin calibration loss. The generated rubric is treated as fixed text context when computing the scorer loss, so the calibration loss does not backpropagate through the rubric:

$$\mathcal{L}_{\text{scorer}}(\theta) = \frac{1}{G}\sum_{g=1}^{G} \mathcal{L}_{\text{cal}}\big(\theta \mid \hat{R}_g\big). \tag{11}$$

The overall Stage-2 objective is

$$\mathcal{L}_{\text{stage2}}(\theta) = \mathcal{L}_{\text{GRPO}}^{(2)}(\theta) + \alpha(t)\, \mathcal{L}_{\text{scorer}}(\theta). \tag{12}$$

Here, $\alpha(t)$ is a scheduled coefficient that activates the scorer calibration loss after the initial generator warm-up, improving training stability under the shared parameters. Under this joint objective, the generator learns query-adaptive rubrics beyond the static seed rubrics. The tips T_k can adapt to prompt-specific details, while the weights w_k vary with the query; meanwhile, the scorer remains guided by the human-aligned margin loss. A qualitative example is provided in Figure 7.

5 Experimental Evaluation

We evaluate RewardVerse from four aspects: (i) pointwise correlation with human ratings across the 16 EvalVerse dimensions; (ii) transfer to pairwise ranking on VGRB; (iii) the contribution of each RGPO component; and (iv) downstream use as an RL reward for video generation.

5.1 Benchmarks and Baselines

We validate our method on two complementary benchmarks. First, we curate a pointwise test set from **EvalVerse** (Yang et al., 2026), currently the most comprehensive evaluation suite for generative

Table 1: Per-dimension **PLCC** (top row) and **SRCC** (bottom row) on the 16 secondary dimensions of EvalVerse (dimension names defined in Appendix A.2). **Avg.** denotes the macro-average over the 10 dimensions supported by all compared reward models. Best in **bold**, second-best underlined. '–' means the reward model does not natively support that dimension.

Method		Scene	Const.	Action	Expr.	Comp.	Lens	Pace	Visual	Chrom.	Mater.	Light.	Ground.	Prog.	Logic	Rhythm	Char.	Avg.
External Reward Models																		
VideoScore-v1.1	PLCC	-0.086	0.102	-0.038	–	–	–	–	0.119	-0.105	-0.156	-0.031	–	–	-0.245	0.070	0.042	-0.033
	SRCC	0.002	0.255	-0.042	–	–	–	–	0.242	0.117	0.012	0.195	–	–	0.099	0.280	0.155	0.132
VideoScore2	PLCC	0.449	0.245	0.390	–	–	–	–	0.507	0.367	**0.697**	0.534	–	–	0.372	0.427	0.346	0.433
	SRCC	0.443	0.337	0.363	–	–	–	–	0.481	**0.324**	**0.698**	0.502	–	–	0.370	0.340	0.409	0.427
UnifiedReward	PLCC	0.461	0.307	0.175	–	–	–	–	0.404	-0.020	0.333	0.463	–	–	0.465	0.359	0.396	0.334
	SRCC	0.468	0.321	0.183	–	–	–	–	0.332	0.063	0.317	0.340	–	–	0.549	0.354	0.355	0.328
VideoReward	PLCC	0.313	0.381	0.326	–	–	–	–	0.462	0.237	0.573	0.483	–	–	0.326	0.265	0.474	0.384
	SRCC	0.334	0.360	0.354	–	–	–	–	0.556	0.220	0.622	0.471	–	–	0.334	0.304	0.430	0.399
VisionReward	PLCC	0.457	0.246	0.332	–	–	–	–	0.554	0.234	0.524	0.514	–	–	0.399	0.387	0.429	0.408
	SRCC	0.400	0.220	0.186	–	–	–	–	0.519	0.191	0.508	0.479	–	–	0.171	0.370	0.466	0.351
Q-Scorer	PLCC	0.496	0.356	0.182	–	–	–	–	0.517	0.234	0.614	**0.623**	–	–	0.593	0.517	0.534	0.467
	SRCC	0.340	0.277	0.139	–	–	–	–	0.370	0.230	0.505	0.451	–	–	0.337	**0.417**	0.476	0.354
RGPO Pipeline Stages																		
Raw-seed Rubric	PLCC	0.432	0.227	0.144	0.290	0.547	0.303	0.079	0.236	0.120	0.523	0.173	0.441	0.225	0.589	0.194	0.471	0.311
	SRCC	0.449	0.372	0.141	0.251	0.275	0.278	0.159	0.266	0.148	0.404	0.213	0.341	0.251	0.478	-0.043	0.326	0.275
Raw-rubric	PLCC	0.437	0.175	0.277	0.415	0.449	0.484	0.075	0.571	0.152	0.531	0.373	0.061	0.183	0.583	0.247	0.447	0.379
	SRCC	0.417	0.178	0.293	**0.340**	0.237	0.166	0.065	0.557	0.055	0.457	0.309	0.034	0.213	0.297	0.107	0.263	0.293
Joint (Ours)	PLCC	**0.520**	**0.395**	**0.566**	**0.448**	**0.575**	**0.486**	**0.471**	**0.661**	**0.498**	0.538	0.557	**0.456**	**0.510**	**0.750**	**0.523**	**0.535**	**0.554**
	SRCC	**0.493**	**0.381**	**0.403**	0.285	**0.440**	**0.333**	**0.431**	**0.610**	0.243	0.514	**0.515**	**0.391**	**0.324**	**0.656**	0.146	**0.500**	**0.446**

video models, spanning 16 fine-grained secondary dimensions (details in Appendix A.2). Second, we evaluate on **VideoGen-RewardBench (VGRB)** (Liu et al., 2026a), a large-scale external pairwise preference benchmark of over 26.5K video pairs. We report the Visual Quality split as transfer on a seen evaluation dimension and the Text Alignment split as transfer to an unseen dimension.

We train our method on Qwen2.5-VL-7B (Team, 2025) and compare against six state-of-the-art video reward models: VideoScore-v1.1 (He et al., 2024), VideoScore2 (He et al., 2025), UnifiedReward and its thinking variant (Wang et al., 2025), VideoReward (Liu et al., 2026a), VisionReward (Xu et al., 2026a), and Q-Scorer (Tang et al., 2026). To trace the effect of each stage in our pipeline, we further analyze two internal baselines: **Raw-seed Rubric** (zero-shot with the seed rubric), **Raw-rubric** (zero-shot with dynamically generated but uncalibrated rubrics).

5.2 Pointwise Evaluation on EvalVerse

Table 1 reports the per-dimension pointwise evaluation results across the 16 EvalVerse dimensions. RewardVerse (**Joint (Ours)**) achieves the strongest overall performance across the 16 dimensions, with the highest PLCC in 14 dimensions. It outperforms other models by a clear margin on challenging cognitive and temporal tasks, such as Logic (PLCC of 0.750 vs. the next-best 0.593) and Action (PLCC of 0.566 vs. 0.390). These results show that RGPO effectively aligns dynamic rubric-guided scores with human judgments. External video scorers can be poorly calibrated on these fine-grained axes (e.g., VideoScore-v1.1 has a PLCC of -0.245 on Logic). Interestingly, Q-Scorer remains competitive on some non-quality dimensions such as Lighting and Logic, suggesting that these dimensions are partly correlated with low-level visual quality. Beyond correlation, RewardVerse also shows better score alignment with human ratings after RGPO, as further analyzed in Appendix A.1.7.

5.3 Pairwise Transfer on VGRB

Table 2 reports the pairwise agreement on the large-scale external benchmark VGRB. Most notably, RewardVerse achieves the best performance among all non-oracle baselines on both splits. Despite training on only 30 pairs per dimension (480 pairs in total), our model outperforms the strongest non-oracle baseline (VisionReward) on the seen Visual Quality (VQ) split by 7.1 (Acc w/ Tie) and 7.0 (Acc w/o Tie) percentage points, closing much of the gap to the oracle VideoReward. Furthermore, our model demonstrates strong zero-shot generalization on the unseen Text Alignment (TA) split. It achieves the highest accuracy (0.471 with ties and 0.623 without ties). These results show that mitigating scalar drift in pointwise scoring can also improve pairwise preference alignment, even when transferring to out-of-distribution benchmarks and unseen evaluation dimensions.

Table 2: Pairwise agreement on VGRB. VQ is seen for RewardVerse; TA is unseen at training. VideoReward is the model whose training set defines VGRB and is shown in gray as an oracle upper bound rather than a comparable baseline.

	Text Alignment (unseen)		Visual Quality (seen)	
Method	Acc w/ Tie	Acc w/o Tie	Acc w/ Tie	Acc w/o Tie
VideoScore-v1.1	0.372	0.503	0.474	0.477
VideoScore2	0.301	0.497	0.374	0.630
UnifiedReward	0.303	0.564	0.412	0.394
UnifiedReward-Thinking	0.428	0.582	–	–
VisionReward	0.465	0.611	0.474	0.590
RewardVerse (ours)	**0.471**	**0.623**	**0.545**	**0.660**
VideoReward (oracle upper bound)	0.538	0.722	0.596	0.756

Table 3: Component-wise ablation on EvalVerse (averaged over 16 dimensions). Best in **bold**.

Configuration	PLCC ↑	SRCC ↑
Full RGPO	**0.530**	**0.416**
(a) w/o Stage-1 Scorer Warm-up	0.431	0.355
(b) w/o Stage-2 Joint Optimization	0.456	0.364
(c) w/o Hierarchical tips T_k (flat rubric)	0.497	0.398
Rubric-Free Baselines (equal 480*-pair budget)*		
(d) V2-Raw (zero-shot, soft-logits)	0.361	0.234
(e) V2-Training (direct pointwise training)	0.439	0.351

5.4 Ablation Studies and Downstream Evaluation

Table 3 reports macro-averaged PLCC and SRCC on EvalVerse, together with two rubric-free baselines under the same 480-pair budget. Three findings emerge. First, using a rubric as an intermediate representation is important for low-data efficiency. Under the same 480-pair budget, direct rubric-free pointwise training (e) improves only slightly over its zero-shot counterpart (d) and remains well below Full RGPO, showing the benefit of introducing the rubric between the query and the scorer when training data is limited. Second, Stage-1 scorer warm-up is important for score calibration. Removing Stage 1 (a) drops PLCC by 0.099 and SRCC by 0.061, showing the benefit of calibrating the scorer before dynamic rubric learning begins. Third, Stage-2 joint optimization improves over static rubrics. Skipping Stage 2 (b) yields a 0.074 PLCC gap compared with the full model, confirming the benefit of adapting rubrics to different queries. Qualitative results are provided in Appendix A.4.3.

Downstream Video Generation. We further use RewardVerse as the reward for GRPO fine-tuning of Wan-2.2-A14B (Wan et al., 2025) on Visual Quality. RewardVerse improves Imaging Quality from 0.640 to 0.653, while VBench-Quality remains stable (0.808 to 0.809) and VBench-Text improves from 0.428 to 0.446. In comparison, VideoReward improves Imaging Quality to 0.648 but reduces VBench-Quality and VBench-Text to 0.804 and 0.392. These results show that RewardVerse improves the target quality while better preserving other video capabilities. Full settings and results are provided in Appendix A.3.

6 Conclusion

In this paper, we introduced RewardVerse, a robust pointwise video reward model, and its training algorithm, Rubric-Guided Policy Optimization (RGPO). Addressing scalar drift in unconstrained MLLM scoring, we demonstrated that inserting a dynamic rubric improves score resolution, while RGPO further alleviates the remaining scalar drift through joint optimization. We also showed that both stages of RGPO—seed-guided warm-up and joint optimization—are essential to this process. With only 30 preference pairs per dimension, RewardVerse achieves high pointwise correlation on EvalVerse, strong pairwise agreement on VGRB, and improves downstream video generation while being less prone to reward hacking. These results show that mitigating scalar drift is crucial for using video reward models as reinforcement learning signals.

AI USE STATEMENT

In this work, we used generative AI tools to generate, verify, and iteratively refine the seed rubrics used for reward-model training. We did not use generative AI tools for research ideation, method or experiment design, implementation, data analysis, result interpretation, or manuscript writing. All AI-generated seed rubrics were reviewed and validated through the procedure described in the paper. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work does not involve human-subject studies or private data. The models, datasets, and benchmarks used in our experiments are publicly available research resources. As with other learned reward models, RewardVerse may inherit biases from its training data and human preference annotations. We encourage transparent evaluation criteria and human oversight when such reward models are used to optimize generative video systems.

REPRODUCIBILITY STATEMENT

We provide the data construction and evaluation details in Appendix A.2, the downstream RL setup in Appendix A.3, and the prompt templates and training configurations in Appendix A.4. We will release the code and training/evaluation splits to facilitate reproduction of our results.

REFERENCES

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Xuehai Bai, Yang Shi, Yi-Fan Zhang, Xuanyu Zhu, Yuran Wang, Yifan Dai, Xinyu Liu, Yiyan Ji, Xiaoling Gu, and Yuanxing Zhang. Edit-compass & editreward-compass: A unified benchmark for image editing and reward modeling. *arXiv preprint arXiv:2605.13062*, 2026.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Hanzhong Guo, Jie Wu, Jie Liu, Yu Gao, Zilyu Ye, Linxiao Yuan, Xionghui Wang, Yizhou Yu, and Weilin Huang. Leveraging verifier-based reinforcement learning in image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34343–34352, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. Videoscore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025.

Runhui Huang, Jie Wu, Rui Yang, Zhe Liu, and Hengshuang Zhao. Alphagrpo: Unlocking self-reflective multimodal generation in unified multimodal models via decompositional verifiable reward. In *Forty-third International Conference on Machine Learning*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Xin Jin, Huanqia Cai, Zhen Li, Zechao Zhan, Dengyang Jiang, Aiming Hao, Yuming Jiang, Xiangpeng Yang, Chunle Guo, Peng Gao, et al. Z-reward: Beyond scalar rewards by internalizing reasoning into score distributions. *arXiv preprint arXiv:2606.09076*, 2026.

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, 2026a.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17417–17437, 2026b.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Chengqi Deng, Y. K. Li, Fuli Luo, Ruoyu Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in common language models. *arXiv preprint arXiv:2402.03300*, 2024.

Leheng Sheng, Wenchang Ma, Ruixin Hong, Xiang Wang, An Zhang, and Tat-Seng Chua. Reinforcing chain-of-thought reasoning with self-evolving rubrics. *arXiv preprint arXiv:2602.10885*, 2026.

Zhenchen Tang, Songlin Yang, Bo Peng, Zichuan Wang, and Jing Dong. Revisiting mllm based image quality assessment: Errors and remedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9475–9483, 2026.

Qwen Team. Qwen2.5-vl: Making language models can actually see the world. *Alibaba Qwen Blog*, 2025. URL `https://qwen.ai/blog?id=qwen2.5-vl`.

Juanxi Tian, Fengyuan Liu, Jiaming Han, Yilei Jiang, Yongliang Wu, Yesheng Liu, Haodong Li, Furong Xu, and Wanhua Li. Auto-rubric as reward: From implicit preferences to explicit multimodal generative criteria. *arXiv preprint arXiv:2605.08354*, 2026.

Haibo Tong, Zhaoyang Wang, Zhaorun Chen, Haonian Ji, Shi Qiu, Siwei Han, Kexin Geng, Zhongkai Xue, Yiyang Zhou, Peng Xia, et al. Mj-video: Fine-grained benchmarking and rewarding video preferences in video generation. *arXiv preprint arXiv:2502.01719*, 2025.

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Sherry Wu. Checklists are better than reward models for aligning language models. *Advances in Neural Information Processing Systems*, 38:114728–114754, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xinran Wang, Songyu Xu, Shan Xiangxuan, Yuxuan Zhang, Muxi Diao, Xueyan Duan, Kongming Liang, Zhanyu Ma, et al. Cinetechbench: A benchmark for cinematographic technique understanding and generation. *Advances in Neural Information Processing Systems*, 38, 2026a.

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

Yuan Wang, Ouxiang Li, Yulong Xu, Borui Liao, Jiajun Liang, Jinghan Li, Meng Wang, Xintao Wang, Pengfei Wan, Kuien Liu, et al. Think, then score: Decoupled reasoning and scoring for video reward modeling. *arXiv preprint arXiv:2605.05922*, 2026b.

Hou Xia, Zheren Fu, Fangcan Ling, Jiajun Li, Yi Tu, Zhendong Mao, and Yongdong Zhang. Videolevelgauge: Investigating contextual positional bias in video language models. In *International Conference on Learning Representations*, volume 2026, pp. 45877–45905, 2026.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026a.

Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training. *arXiv preprint arXiv:2602.01511*, 2026b.

Songlin Yang, Haobin Zhong, Ruilin Zhang, Xiaotong Zhao, Shuai Li, Kai Zheng, Xuyi Yang, Zhe Wang, Zhenchen Tang, Yang Li, et al. Evalverse: Pipeline-aware and expert-calibrated benchmarking for professional cinematic video generation. *arXiv preprint arXiv:2605.23271*, 2026.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in neural information processing systems*, 33:9422–9434, 2020.

Yuxuan Zhang, Penghui Du, Bo Li, Cong Wei, Junwen Miao, Huaisong Zhang, Songcheng Cai, Yubo Wang, Dongfu Jiang, Yuyu Zhang, et al. Rewardharness: Self-evolving agentic post-training. *arXiv preprint arXiv:2605.08703*, 2026.

Shijie Zhao, Xuanyu Zhang, Weiqi Li, Junlin Li, Li Zhang, Tianfan Xue, and Jian Zhang. Reasoning as representation: Rethinking visual reinforcement learning in image quality assessment. *arXiv preprint arXiv:2510.11369*, 2025.

A APPENDIX

Anonymous Code Link: https://anonymous.4open.science/r/rewardverse-anonymous-FE7E/

A.1 EXTENDED EMPIRICAL DIAGNOSTICS ON SCALAR DRIFT

A.1.1 FULL 4-VARIANT ABLATION FOR QWEN2.5-VL

Table 4 expands upon the core conclusions of Section 3.2. To ensure a controlled and rigorous evaluation, all four variants are evaluated using the same multi-modal LLM backbone (Qwen2.5-VL-7B). They share an identical pointwise hold-out set ($n_{\text{pw}} = 240$) and pairwise comparison pool ($n_{\text{pair}} = 240$), spanning sixteen key dimensions from EvalVerse's global taxonomy to ensure comprehensive coverage. All inference hyperparameters are held constant across variants, utilizing a frame rate of FPS $= 2$ and greedy decoding. For evaluation metrics, we let $\sigma(\hat{s})$ denote the standard deviation of predicted scores, bias $= \overline{\hat{s}} - \overline{s_{\text{label}}}$ represent the systematic overestimation error, and $\overline{|\Delta|}$ represent the mean prediction gap on decided pairs (i.e., those whose constituent videos receive distinct scores).

Table 4: A four-variant ablation study on the white-box **Qwen2.5-VL-7B** backbone across the shared 16-dimensional hold-out set ($n_{\text{pw}} = 240$) and pair pool ($n_{\text{pair}} = 240$). The deployed configuration (V3) is highlighted in **bold**.

Variant	Rubric	Format	$\overline{\hat{s}}$	$\sigma(\hat{s})$	Bias	Decided	Pair Acc.	$\overline{\lvert\Delta\rvert}$
V1 (NL free-float)	✗	natural	4.747	0.203	+1.097	58/240	$0.931^{\dagger}$	0.338
V2 (soft-logits)	✗	soft-log.	4.194	0.258	+0.546	239/240	0.795	0.168
V3 (rubric + soft-logits)	✓	soft-log.	4.419	**0.471**	+0.771	**240/240**	0.754	**0.216**
V4 (rubric + multi-theme NL float)	✓	natural	4.649	0.275	+1.001	76/240	$0.737^{\dagger}$	0.425

† V1 and V4 produce only a small portion of distinct pair scores (the rest tie); the reported pair accuracy is evaluated on the decided subset and therefore over-states real-world separability.

A.1.2 LARGE-SCALE EVALUATION OF GEMINI-3.1-PRO ON 1K TEST SET

To evaluate the scalability and generalizability of our rubric-guided paradigm on frontier models, we extend our evaluation to **Gemini-3.1-Pro** using the complete 1K multi-dimensional pointwise test set. Given that closed-source APIs restrict access to token-level logits, we benchmark the V1 (No Rubric) and V4 (With Rubric) configurations under natural-language float output formats. Table 5 lists the PLCC, SRCC, standard deviation and mean scores on the 1K dataset, while Figure 4 shows the distribution of our predictions against human labels.

Table 5: Comprehensive pointwise calibration and correlation metrics of **Gemini-3.1-Pro** on the large-scale **1K pointwise test set** (16 evaluation dimensions) under V1 and V4 configurations. Standard deviation, biases, and discrete rating distributions are included to show scaling behavior and systematic offsets.

Variant	Rubric	$\overline{\hat{s}}$	$\sigma(\hat{s})$	Bias	PLCC	SRCC
V1	✗	3.558	1.500	-0.398	+0.490	+0.480
V4	✓	3.372	1.158	-0.584	**+0.593**	**+0.564**

The empirical results on the 1K dataset demonstrate that the rubric-as-an-intermediate-representation paradigm generalizes robustly to proprietary, state-of-the-art systems. Integrating the rubric ($V1 \rightarrow V4$) unlocks substantial alignment gains: Pearson correlation (PLCC) increases by +0.103 (a 21.0% relative improvement), while Spearman rank correlation (SRCC) increases by +0.084 (17.5% relative), demonstrating that rubric-guided sub-dimensions align much closer with human multi-dimensional preferences.

Notably, the calibration behavior of Gemini-3.1-Pro reveals a striking contrast to Qwen2.5-VL. While Qwen2.5-VL suffers from systematic overestimation (positive bias), Gemini-3.1-Pro exhibits a persistent underestimation (negative bias of -0.398 in V1 and -0.584 in V4). This divergence demonstrates that different foundational models, regardless of their scale, possess intrinsic score-scale biases. Relying solely on zero-shot static rubrics cannot eliminate these systematic biases,

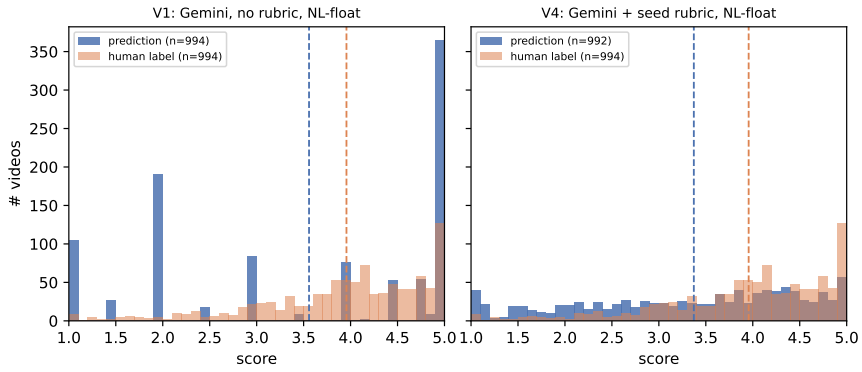


Figure 4: Pointwise prediction histograms of Gemini-3.1-Pro (V1 vs. V4) on the complete 1K pointwise test set, overlaid on human gold standard ratings (orange). Rubrics improve correlation but worsen calibration. Decomposing evaluations (V4) aligns the rating variance closer to human baselines, reducing the uncalibrated standard deviation from 1.500 (V1) to 1.158 (V4).

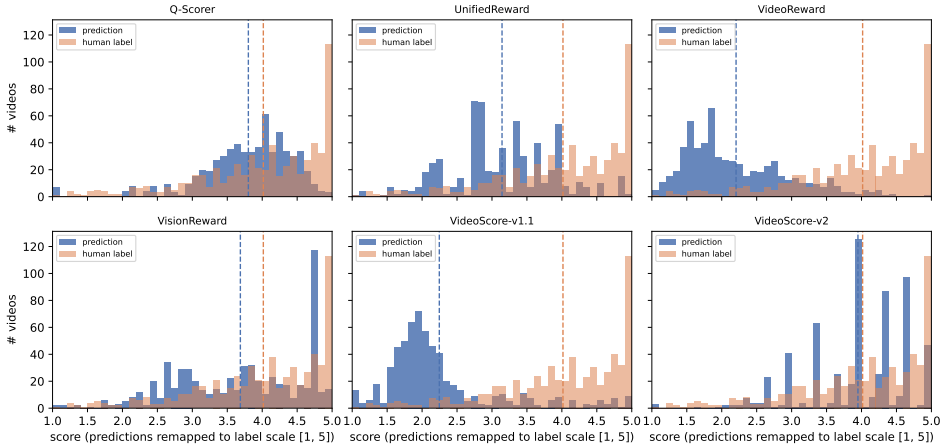


Figure 5: Score distributions of six reward models on the shared 10-dimensional test subset. Predictions are remapped to the human rating scale $[1, 5]$ for visualization. Blue bars denote model predictions and orange bars denote human labels. Dashed lines show their means. Most reward models exhibit score-range compression or distribution shift. Q-Scorer produces a relatively broader score distribution, but it still does not fully match human ratings.

which strongly justifies the necessity of our proposed RGPO training paradigm to explicitly align the reward model with human anchors.

A.1.3 Scalar Drift across Existing Reward Models

We further examine the score distributions of six existing reward models on the shared 10-dimensional test subset. Since these models use different native output ranges, we remap their predictions to $[1, 5]$ for visualization. This remapping is applied only to display the distributions.

As shown in Figure 5, most reward models produce scores within a narrow range or show a clear shift from the human rating distribution. VideoReward and VideoScore-v1.1 mainly concentrate in the low-score region, while VisionReward and VideoScore-v2 place many predictions near the upper end. UnifiedReward also produces several narrow score peaks. Q-Scorer is the only model that gives a relatively broad and continuous score distribution, although its mean and overall distribution still differ from human ratings. These results show that score compression or scale shift is common across existing reward models.

A.1.4 PER-DIM BREAKDOWN

The pooled bias of the deployed V3 configuration is $+0.771$ (bootstrap 95% CI: $[+0.64, +0.90]$, $n_{\text{boot}} = 2000$, $n = 240$). Variants V1 ($+1.10$), V2 ($+0.55$), and V4 ($+1.00$) similarly overshoot human ratings, ruling out the rubric scaffold as the primary source of overestimation. As shown in Figure 6, 15 out of 16 evaluation dimensions exhibit positive bias—with *rhythm* being the sole negative exception (-0.23). The largest calibration gaps occur in *pacing* ($+1.68$), *consistency* ($+1.64$), and *character* ($+1.48$). This systematic positive offset constitutes the limitation that our proposed RGPO training algorithm (Section 4) aims to mitigate.

A dimensional analysis reveals a clear semantic divide: higher-level temporal reasoning and consistency dimensions (e.g., *pacing*, *consistency*) suffer from severe overestimation, whereas low-level sensory dimensions (e.g., *rhythm*) show minimal bias. This pattern suggests that uncalibrated MLLMs act as overly optimistic critics: they align well with low-level sensory cues but remain blind to fine-grained temporal artifacts and physical plausibility. An explicit reinforcement learning loop (RGPO) is vital to break through this limitation.

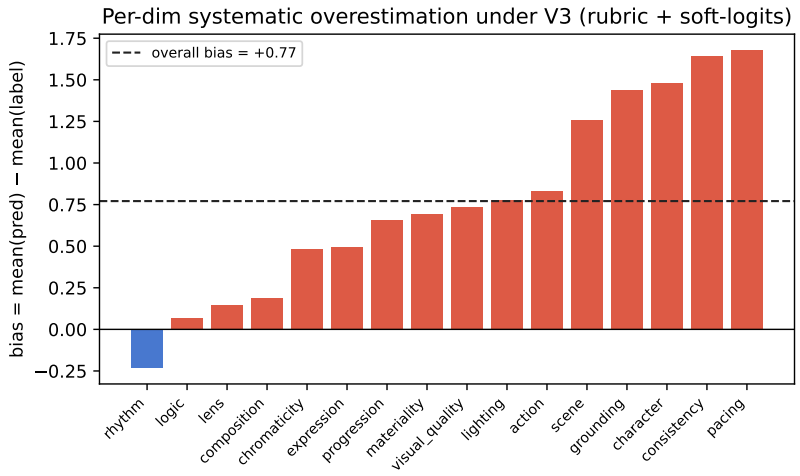


Figure 6: Per-dimension systematic overestimation bias of the V3 protocol ($n = 15$ per dimension). Bars represent the difference between predicted means and human ratings ($\overline{\hat{s}_d} - \overline{s_{\text{label},d}}$), with the dashed line marking the pooled bias ($+0.77$). The pervasive positive offset across 15 out of 16 dimensions highlights a limitation that static, zero-shot rubric prompting fails to resolve.

A.1.5 PROTOCOL SELECTION: HOW TO USE THE RUBRIC

Section 3.2 establishes that the deployed evaluator must combine a rubric with a soft-logits readout. A secondary design question remains: *within* the rubric paradigm, what is the best way to query and format the prompt? We benchmark four candidate protocols on the *visual quality* hold-out set ($n_{\text{pw}} = 30$, $n_{\text{pair}} = 30$), all utilizing the same rubric:

- **PR1 (Pairwise A/B):** The model evaluates both videos simultaneously in a single forward pass, selecting video A or B with the rubric injected as an evaluation reference.
- **PR2 (Pointwise Overall Integer):** One forward pass is conducted per video; the model is prompted to emit a single integer score in the range $[1, 5]$.
- **PR3 (Pointwise Multi-Theme Integer):** One forward pass is conducted per video; the model outputs a JSON object containing per-theme integer scores.
- **PR4 (Pointwise Multi-Theme Soft-Logits):** One forward pass is conducted per video; the continuous soft-logits expectation is computed for each rubric theme, and the overall score is aggregated via a weighted average. This is the deployed configuration (V3).

The *format-failure rate* counts forward passes that either cannot be successfully parsed as valid JSON/integers. Ties at integer granularity are reflected separately by the number of decided pairs. Since PR1 directly outputs a binary preference rather than a scalar value, its correlation metrics (PLCC, SRCC) and score margins are not defined.

Table 6: A comparative evaluation of different rubric-querying protocols on the *visual_quality* hold-out set (30 pointwise samples and 30 evaluation pairs). The proposed **PR4** is the unique protocol that simultaneously achieves a 0% format-failure rate, generates continuous scalar outputs, and preserves full pairwise separability (100% decided pairs).

Protocol	Format-Fail Rate	Decided	Pair Acc.	PLCC	$\overline{\lvert\Delta\rvert}$
PR1 — pairwise A/B	0.000	30/30	1.000	—	—
PR2 — pointwise integer (text)	0.467	9/30	1.000	0.313	1.333
PR3 — multi-theme integer (text)	0.533	6/30	0.833	0.374	1.667
PR4 — multi-theme soft-logits	**0.000**	**30/30**	0.833	0.229	0.269

Although PR1 achieves perfect pairwise accuracy, it is unsuitable as an RL reward signal due to its lack of scalar outputs. Furthermore, PR1 is highly sensitive to input ordering (see Appendix A.1.6), consistent with contextual positional bias observed in video-language models (Xia et al., 2026). While proprietary models (e.g., Gemini-3.1-Pro) handle multi-video comparison robustly, using them as iterative RL reward models is computationally and financially prohibitive.

For integer-based configurations (PR2 and PR3), formatting parser failures and scale quantization ties severely hinder evaluation, yielding failure rates of 46.7% and 53.3%, respectively. These high rates demonstrate that free-form text score generation is fragile, as minor JSON variations or tokenization shifts block automated parsing. Conversely, PR4 bypasses parsing bottlenecks by extracting soft logits directly from vocabulary tokens, guaranteeing a 0% failure rate by construction. By avoiding multi-video reasoning bottlenecks and providing a continuous optimization surface, PR4 serves as our final deployment variant.

A.1.6 Stability under Prompt Perturbation and Pair Swap

Two diagnostic studies probe the second symptom of scalar drift: *high variance under context shift*.

Diagnostic 1: Prompt perturbation (V2 vs. V3). We evaluate 48 videos (3 per dimension across 16 dimensions) three times under three semantically equivalent overall-quality instructions, reporting the mean per-video standard deviation. We compare V2 and V3 because both utilize the soft-logits readout, leaving the rubric as the sole independent variable. Variant V1 is excluded as its scores collapse onto a single peak, rendering stability metrics meaningless. To ensure commensurability between the two distinct scoring scales, each variant's per-video standard deviation is normalized by its global standard deviation $\sigma(\hat{s})$ from Table 4.

Table 7: Per-video score standard deviation across $K = 3$ paraphrased instructions ($N = 48$ videos). Statistics are computed over per-video prompt-perturbation standard deviations $\{\sigma_i\}_{i=1}^{N}$. "Normalized" reports the Raw Mean scaled by the global standard deviation $\sigma(\hat{s})$.

	Raw Mean	Median	Max	Normalized
V2 (soft-logits, no rubric)	0.046	0.040	0.120	0.179
V3 (rubric + soft-logits)	0.094	0.038	0.418	0.199

After normalization, both variants exhibit comparable relative variance under prompt paraphrase (0.18 vs. 0.20). This demonstrates a crucial structural benefit: adding the rubric (V3) decompresses the scoring scale by 82.6% ($\sigma(\hat{s})$ increases from 0.258 to 0.471) while maintaining similar relative stability under prompt paraphrases. The smaller raw variation of V2 mainly results from its compressed score range. Under the soft-logits readout, the rubric substantially expands score resolution without causing a comparable increase in relative sensitivity to prompt paraphrases.

Diagnostic 2: Pair A/B order swap (no rubric vs. seed rubric). We evaluate every video pair in both input orderings, $(y_w,\, y_l)$ and $(y_l,\, y_w)$, and report the rate at which preference verdicts disagree across the swap (flip rate). The pairwise A/B protocol is tested on 48 sampled pairs (3 per dimension) under two settings: *without* the rubric (baseline) and *with* the seed per-dimension rubric injected as a reference (PR1). Table 8 summarizes the results.

Introducing the rubric reduces the flip rate by 23.0 percentage points (from 0.792 to 0.562), confirming that a rubric serves as a grounding reference that aligns model decisions with objective criteria rather than order-sensitive heuristics. However, a flip rate of 56.2% remains very high, confirming

Table 8: Pair A/B swap flip rates with and without the rubric (48 pairs). Injecting the rubric substantially reduces the flip rate, yet the residual bias remains a significant bottleneck for the pairwise protocol.

Setting	Decided pairs	Flip rate
Pairwise A/B, no rubric	48/48	0.792
Pairwise A/B, seed rubric	48/48	0.562

Table 9: Pointwise scoring before and after RGPO under the same dynamic-rubric protocol. Bias is the difference between the mean prediction and the mean human rating. The 95% confidence interval is obtained by bootstrap sampling.

Model	N	Mean	$\sigma(\hat{s})$	Bias	95% CI	PLCC	SRCC
Pre-RGPO	240	4.198	0.457	+0.610	[+0.488, +0.736]	0.240	0.222
Post-RGPO	240	3.752	1.332	+0.164	[+0.012, +0.313]	0.505	0.458

that the binary A/B format is highly vulnerable to position bias when processing two complex video sequences simultaneously.

These findings yield two critical conclusions. First, this high residual flip rate disqualifies the pairwise A/B protocol from serving as an RL reward model, as unstable preference signals would generate highly noisy gradients. This provides an independent justification for adopting our pointwise multi-theme soft-logits protocol (PR4 / V3). Second, it highlights the division of labor in our framework: the rubric calibrates the scoring criteria, while the pointwise soft-logits format ensures greater robustness under context shifts by completely bypassing order-dependent comparison bottlenecks.

A.1.7 Scalar Drift Before and After RGPO

The four-variant study in Appendix A.1.1 analyzes the effect of rubric injection and score readout before training. We further evaluate whether the complete RGPO training further mitigates the remaining scalar drift. We compare the original Qwen2.5-VL-7B model with the jointly trained RewardVerse model. Both models generate a dynamic rubric for each query and use the same multi-theme soft-logits scoring protocol. Score calibration is evaluated on the same pointwise hold-out set ($n = 240$) in Appendix A.1.1, while prompt stability is evaluated on the 48 videos and three paraphrased instructions used in Appendix A.1.6.

As shown in Table 9, the untrained model remains systematically optimistic even when using dynamic rubrics, with a mean bias of $+0.610$. Although RGPO is optimized for relative score margins rather than absolute scores, the bias decreases to $+0.164$, while PLCC and SRCC increase from 0.240 and 0.222 to 0.505 and 0.458, respectively. The prediction standard deviation also increases from 0.457 to 1.332, showing that RGPO further expands the previously compressed score range. At the dimension level, the absolute calibration bias is reduced on 10 out of 16 dimensions, with the largest reductions appearing in pacing, materiality, character, lighting, and scene. We additionally repeat the prompt-perturbation diagnostic under the same dynamic-rubric protocol. For each video, we compute the standard deviation of its scores across three semantically equivalent instructions. Since RGPO substantially expands the overall score range, we normalize the mean per-video standard deviation by the global prediction standard deviation.

As shown in Table 10, the raw variation slightly increases from 0.094 to 0.113 because RGPO expands the overall score range. After normalization, however, the relative variation decreases from 0.206 to 0.085, corresponding to a 58.7% reduction. This shows that RGPO improves robustness to prompt paraphrases while preserving the expanded score resolution. These results complement the protocol analysis in Appendix A.1. Rubrics with soft-logits reduce score-range collapse before training, while RGPO further improves both score alignment and robustness to prompt paraphrases.

A.2 EvalVerse 16-Dimensional Taxonomy

Throughout the paper, we adopt the sixteen fine-grained secondary dimensions of the EvalVerse suite, which systematically cover the six coarse axes used in generative video evaluation: *Visual Concept Design*, *Acting*, *Cinematography*, *Aesthetics*, *Affectivity*, and *Multi-Shot Cutting*. Table 11 provides the definitions and evaluation focus of each dimension. Rubric generation and scoring in

Table 10: Prompt-perturbation stability before and after RGPO under the same dynamic-rubric scoring protocol. Raw statistics are computed over the per-video standard deviations across three paraphrased instructions. Normalized is the Raw Mean divided by the global prediction standard deviation.

Model	Raw Mean	Median	Max	Global $\sigma(\hat{s})$	Normalized
Pre-RGPO	0.094	0.038	0.418	0.457	0.206
Post-RGPO	0.113	0.042	0.755	1.332	**0.085**

RewardVerse are conditioned on the target dimension d, enabling the generator to learn dimension-adaptive themes and weights rather than a single generic scoring policy.

Table 11: The sixteen secondary evaluation dimensions adapted from the EvalVerse test suite, organized by their high-level axis and specific evaluation focus. All training and testing in Sections 3 and 5 are conducted using this taxonomy.

ID	Dimension	Axis	Evaluation Focus
1. Visual Concept Design			
D_1	Character	Visual Concept	Clarity, recognizability, and adherence of character appearance to the designated settings.
D_2	Scene	Visual Concept	Plausibility and appeal of the scene setup; establishment of a clear, distinct, and aesthetically valuable visual style.
2. Acting			
D_3	Consistency	Acting	Coherence of character identities and attributes across frames without temporal drifting, morphing, or flickering.
D_4	Action	Acting	Naturalness and physical plausibility of character movements aligned with their established identity.
D_5	Expression	Acting	Accuracy and vividness of facial expressions in conveying the intended emotions.
3. Cinematography			
D_6	Composition	Cinematography	Clarity, stability, and explicit visual intent of individual shot compositions.
D_7	Lens	Cinematography	Reasonableness, clarity, and stability of camera parameters (focal length, depth-of-field, etc.).
D_8	Pacing	Cinematography	Appropriateness of camera movements and blocking; absence of jarring or unmotivated motion.
4. Aesthetics			
D_9	Visual Quality	Aesthetics	Low-level perceptual quality, temporal consistency, and absence of digital artifacts.
D_{10}	Chromaticity	Aesthetics	Color harmony, saturation control, and emotive power.
D_{11}	Materiality	Aesthetics	Aesthetic appeal and self-consistent realism of textures, surface properties, and fine details.
D_{12}	Lighting	Aesthetics	Physical plausibility and artistic execution of illumination, shading, and cinematic lighting logic.
5. Affectivity			
D_{13}	Grounding	Affectivity	Stability and recognizability of the overall emotional direction and atmospheric setting.
D_{14}	Progression	Affectivity	Natural and controlled build-up, transition, and resolution of emotion across the temporal dimension.
6. Multi-Shot Cutting			
D_{15}	Logic	Multi-Shot	Continuity of time, space, and action between shot transitions; effectiveness of editing in aiding narrative comprehension.
D_{16}	Rhythm	Multi-Shot	Overall control of shot durations and cutting cadence; effective emotional guidance and clear temporal organization.

For each dimension, we manually curate a small seed pool of 30 preference pairs (q, y_w, y_l), totaling 480 pairs across the entire taxonomy. These pairs serve both to self-evolve the seed rubric $R_d^\star$ (Section 4.2) and to train the RGPO framework. All test-set correlations reported in Table 1 are evaluated on a disjoint pointwise hold-out set containing 1K evaluation samples ($n = 994$) across the taxonomy.

Table 12: Baseline scoring protocols and dimension alignment used in Table 1.

Method	Scoring protocol and alignment
VideoScore-v1.1	The model predicts visual quality, temporal consistency, dynamic degree, text-to-video alignment, and factual consistency. Visual quality is used for Visual Quality, Chromaticity, Materiality, and Lighting; temporal consistency for Consistency and Rhythm; dynamic degree for Action; text-to-video alignment for Scene and Character; and factual consistency for Logic.
VideoScore2	The model generates visual quality, text-to-video alignment, and physical/common-sense consistency scores. Visual quality is used for Visual Quality, Chromaticity, Materiality, and Lighting; text-to-video alignment for Scene and Character; and physical/common-sense consistency for Consistency, Action, Logic, and Rhythm.
UnifiedReward	We use its Style, Physics, and Alignment scores. Style is used for Visual Quality, Chromaticity, Materiality, Lighting, and Character; Physics for Consistency, Action, and Rhythm; and Alignment for Scene and Logic.
VideoReward	We use its visual quality (VQ), motion quality (MQ), and text alignment (TA) scores. VQ is used for Visual Quality, Chromaticity, Materiality, Lighting, and Character; MQ for Consistency, Action, and Rhythm; and TA for Scene and Logic.
VisionReward	We use the score from its fixed video checklist for all reported dimensions. The checklist is not changed across dimensions.
Q-Scorer	We sample video frames, score each frame using the official image-quality prompt, and average the frame scores. The same frame-averaged score is used for all reported dimensions.

Baseline Evaluation and Dimension Alignment. We briefly introduce the external reward models used in Table 1. VideoScore-v1.1 is a discriminative video evaluator with five regression heads for visual quality, temporal consistency, motion, text alignment, and factual consistency. VideoScore2 is a generative video evaluator that scores visual quality, text alignment, and physical/common-sense consistency. UnifiedReward is an MLLM-based reward model that evaluates generated content from Style, Physics, and Alignment. VideoReward is a discriminative video reward model trained with human preferences and predicts visual quality, motion quality, and text alignment. VisionReward is a fine-grained reward model that evaluates videos using a fixed weighted checklist. Q-Scorer is a state-of-the-art IQA model that is particularly strong at structural quality assessment.

We use the official checkpoints without additional training. Since these models do not cover all 16 EvalVerse dimensions, we report the dimensions supported in Table 1 and mark the others as "–". For models with coarse outputs, we use the closest native score following Table 12.

A.3 DOWNSTREAM GRPO FINE-TUNING

A.3.1 FINE-TUNING SETUP AND EVALUATION

We further evaluate whether RewardVerse can serve as an effective reward signal for downstream video generation optimization. We fine-tune Wan-2.2-A14B with GRPO using either VideoReward or RewardVerse as the reward model, while the original Wan-2.2-A14B is used as the base model. The optimization targets only the *Visual Quality* dimension. All training settings are kept identical across the two GRPO runs. We train with LoRA rank 64 and a rollout group size of $G = 16$. For evaluation, we use the VBench prompt suite (Huang et al., 2024) and generate 17-frame videos at 480×832 resolution with deterministic sampling. We report VBench-Quality, averaged over six official quality dimensions, and VBench-Text, averaged over three official text-related dimensions, together with several representative VBench dimensions. Since only Visual Quality is optimized, the remaining dimensions are used to examine whether reward optimization degrades other video capabilities.

A.3.2 DOWNSTREAM RESULTS

As shown in Table 13, RewardVerse improves Imaging Quality from 0.640 to 0.653, which is the dimension most directly related to the optimized Visual Quality reward. It also improves Human Action from 0.900 to 0.950, while Background Consistency, Temporal Flickering, Motion Smoothness, and Subject Consistency remain close to the base model. In contrast, optimization with VideoReward reduces both VBench-Quality from 0.808 to 0.804 and VBench-Text from 0.428 to 0.392.

These results show that RewardVerse improves the target visual quality without clear degradation of other video capabilities. In comparison, VideoReward improves Imaging Quality but reduces the overall VBench performance, suggesting that RewardVerse is less prone to reward hacking during downstream optimization.

Table 13: Downstream GRPO fine-tuning of Wan-2.2-A14B on the Visual Quality dimension. RewardVerse improves the target quality metric while largely preserving other video capabilities. Best results are shown in bold.

Metric	Base	VideoReward	RewardVerse
Imaging Quality	0.640	0.648	**0.653**
Human Action	0.900	0.800	**0.950**
Background Consistency	0.919	0.916	**0.920**
Temporal Flickering	0.943	0.938	**0.944**
Motion Smoothness	**0.966**	0.960	**0.966**
Subject Consistency	**0.898**	0.881	0.894
VBench-Quality (6D)	0.808	0.804	**0.809**
VBench-Text (3D)	0.428	0.392	**0.446**

A.4 Prompts and Training Details

A.4.1 Prompt Templates

Seed Rubric Generation. For each training pair, the teacher receives the target dimension, generation prompt, videos, and human scores. It is asked to generate observable and dimension-specific criteria, with all rubric weights summing to one. The generated rubric is then used to score the videos and is revised when its prediction differs from the human label by more than 0.5. We use at most five revision rounds and remove near-duplicate rubrics using a token-level Jaccard threshold of 0.85.

Dynamic Rubric Generation. The rubric generator uses the following prompt:

```
You are a cinematic video evaluation expert. Generate exactly five
    rubric items for scoring videos on the {dimension} dimension.

Task:
{task_description}

Caption:
{caption}

Each rubric contains:
- a clear theme;
- a weight in (0, 1), with all weights summing to 1;
- at least two observable evaluation tips.

Output strict JSON:
{
  "rubrics": [
    {
      "theme": "...",
      "weight": 0.NN,
      "tips": ["...", "..."]
    },
    ...
  ]
}
```

Rubric-Guided Scoring. Given a video and a generated rubric, the scorer uses the following prompt:

```
You are a strict video reward judge for the {dimension} dimension.

Score the single video on each rubric using an integer from 1 to 5.

Caption:
{caption}

Rubric:
{rubric_text}

Respond with exactly one line:
theme_1: _ theme_2: _ theme_3: _ theme_4: _ theme_5: _
```

At each score position, we read the logits of tokens "1" to "5" and compute their expected value. This fixed output format provides continuous theme scores without parsing free-form text.

A.4.2 Reward Model Training Settings

We train RewardVerse on Qwen2.5-VL-7B using the 480 preference pairs described in Appendix A.2. The training is divided into two separate runs corresponding to the two stages of RGPO. Both the scorer warm-up and joint optimization are trained for one epoch. Each epoch contains approximately 60 optimization steps. For each training pair, we sample a rollout group of size $G = 8$.

For both stages, we use AdamW with a learning rate of 2×10^{-6}, 30 learning-rate warm-up steps, a preference temperature of 1.5, and a maximum gradient norm of 1.0. The GRPO KL coefficient is set to 0.1. Training uses bfloat16 precision and DeepSpeed ZeRO-3 on 8 NVIDIA H20 GPUs.

Stage 1: Scorer Warm-up. The student and reference policies are initialized from Qwen2.5-VL-7B. The seed rubric is fixed for all rollouts, so this stage only updates the scorer role. The scorer update coefficient is fixed at 0.4 throughout the epoch. We set the margin calibration weight λ_{cal} to 0.2, the score-format gate to 0.55, and use an additional score-side KL coefficient of 0.005 to limit score-scale drift. The model is trained for one epoch.

Stage 2: Joint Optimization. We initialize the student from the final Stage-1 checkpoint. The same checkpoint is used as the frozen reference policy for both rubric generation and scoring. The generator produces dynamic rubrics from the query, and the scorer evaluates both videos under each sampled rubric. The rubric-alignment weight and format reward weight are both fixed at 0.5. We increase the margin calibration weight to 0.3 and use a score-side KL coefficient of 0.015. The scorer update coefficient is linearly increased from zero at step 20 to 0.4 at step 30. Joint optimization is also run for one epoch. The Stage-2 checkpoint at step 45 is selected based on the macro-averaged PLCC on the held-out validation set and used for all evaluations.

Stage-2 GRPO Objective. For completeness, the generator-side GRPO objective used in Stage 2 is:

$$\mathcal{L}^{(2)}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|\hat{R}_g|}\sum_{t=1}^{|\hat{R}_g|}\Big[\min\big(\rho_{g,t}(\theta)\hat{A}_g,\ \text{clip}(\rho_{g,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_g\big) - \beta_{\text{KL}}\,\mathbb{D}_{\text{KL}}\big[\pi_\theta^{\text{gen}} \,\big\|\, \pi_{\text{ref}}^{\text{gen}}\big]_{g,t}\Big], \tag{13}$$

where $\hat{A}_g$ and $\rho_{g,t}(\theta)$ are defined in Section 4.3, and $\pi_{\text{ref}}^{\text{gen}}$ is the frozen reference generator.

A.4.3 Qualitative Study: Inside the Learned Rubrics

Figure 7 visualizes the learned rubrics generated by RewardVerse across different prompts and evaluation dimensions. Notably, the generator automatically adjusts the level of abstraction in the tips based on the target dimension. For global, technical, or aesthetic axes such as *Visual Quality*, the generated tips remain abstract to avoid overfitting to the specific prompt content. In contrast, for

Table 14: Main hyperparameters for the two-stage RewardVerse training. Both stages are trained for one epoch.

Setting	Stage 1: Warm-up	Stage 2: Joint
Initialization	Qwen2.5-VL-7B	Stage-1 checkpoint
Reference policy	Qwen2.5-VL-7B	Stage-1 checkpoint
Training epochs	1	1
Optimization steps	∼60	∼60
Rollouts per sample G	8	8
Learning rate	2×10^{-6}	2×10^{-6}
LR warm-up steps	30	30
Preference temperature	1.5	1.5
GRPO KL coefficient	0.1	0.1
Score-side KL coefficient	0.005	0.015
Margin calibration weight	0.2	0.3
Rubric-alignment weight	–	0.5
Format reward weight	–	0.5
Maximum scorer update weight	0.4	0.4
Scorer update schedule	Fixed	Steps 20–30
Score-format gate	0.55	0.55
Maximum gradient norm	1.0	1.0



Figure 7: Qualitative view of dynamic rubrics generated by RewardVerse. Left (*Visual Quality*): themes and tips stay abstract. Right (*Consistency*): tips dynamically ground on specific entities in the prompt, illustrating scene-adaptation.

identity-bound or scene-bound axes such as *Consistency*, the tips dynamically ground on specific entities in the prompt (e.g., monitoring "the light yellow robe with gold patterns on Person_01"), as these entities are exactly what must remain stable. This example illustrates how the generator adjusts the tip granularity to different evaluation dimensions.

A.5 Inference-Time Cost

We measure the inference latency of the four pointwise scoring variants on a single H20 under the same inference settings (bf16, greedy decoding, FPS = 2, and a 64-frame budget). All variants use the same visual backbone and frame-sampling pipeline.

As shown in Table 15, rubric generation introduces only a small inference overhead. V3 takes approximately 2.0 s per video, compared with 1.7 s for the rubric-free V2 baseline, corresponding

to an increase of about 18%. V4 is slightly slower because it additionally generates multiple theme scores in natural text.

The rubric depends only on the evaluation query rather than the video. Therefore, when multiple candidate videos are evaluated under the same query, the rubric can be generated once and reused, further reducing the effective per-video overhead.

Table 15: Per-video inference latency of the four pointwise scoring variants on a single H20.

Variant	Rubric	Readout	Latency
V1 (no rubric, NL float)	✗	NL float	≈ 1.7 s
V2 (no rubric, soft-logits)	✗	soft-logits	≈ 1.7 s
V3 (rubric + soft-logits, deployed)	✓	soft-logits	≈ 2.0 s
V4 (rubric + multi-theme NL float)	✓	NL float	≈ 2.2 s

A.6 Limitations and Future Work

Our current results should be read with a few caveats in mind.

Backbone Capacity. For a controlled and fair evaluation against existing video reward models, we intentionally selected the standard-sized Qwen2.5-VL-7B as our backbone. However, the effectiveness of RGPO is bounded by the capability of the underlying MLLM. RGPO can better align the model's existing evaluation ability with human judgments, but it cannot recover visual or temporal details that the backbone fails to recognize. Therefore, stronger backbones are expected to produce more reliable rubrics and scores, leading to a higher performance ceiling for RL optimization.

Theme-Level Calibration. Our current calibration objective directly aligns only the aggregated score margin between preferred and non-preferred videos. The score of each individual rubric theme is not directly supervised. As a result, the final reward can be well aligned with human preferences while some theme scores remain biased or inconsistent. Directly calibrating theme-level scores is therefore an important direction for improving both score reliability and interpretability.

Scaling. Our current experiments are conducted with a 7B reward backbone, only 30 preference pairs per dimension, and one 14B video generation model for downstream RL. It remains necessary to scale RewardVerse to stronger reward backbones, larger training sets, and more video generation models to verify whether the observed gains remain consistent at scale.

Relation to CoT-Based Scoring. Recent CoT-based visual scoring methods treat reasoning as an intermediate representation that summarizes visual evidence before producing a quality score (Zhao et al., 2025; He et al., 2025; Wang et al., 2026b). RewardVerse operates at a different stage. Its rubric is generated from the evaluation dimension and prompt without video access, and specifies what should be checked before scoring. Therefore, CoT summarizes the evidence observed in the video, whereas the rubric defines the evaluation criteria. The two designs are complementary, and the generated rubric may further guide a CoT-based scorer.

Future Work. Future work will focus on stronger MLLM backbones, direct theme-level calibration, larger-scale training and downstream RL experiments, and rubric-guided CoT scoring.