Aligning Human Sense: Calibrated Distributional Reward Learning for Video Generation

Naixin Zhai[1,11†], Weihua Cheng[2,11], Dexu Yu[3], Yikai Gu[4], Hanwen Du[5,11], Junchen Fu[6,11], Chenxi Huang[7], Yingwei Song[8], Liyuan Lillian Ma[9], Yang Ran[3], Youhua Li[1*], and Yongxin Ni[10*]

[1] City University of Hong Kong [2] ShanghaiTech University [3] Fenz.AI
[4] University of California, Berkeley [5] The Ohio State University
[6] University of Glasgow [7] Columbia University [8] University of Arizona
[9] GMI Cloud [10] National University of Singapore [11] DeciLix Lab
[†] Email: nancyzhzhai@gmail.com
[*] Co-corresponding: youhuali2-c@my.cityu.edu.hk; niyongxin@u.nus.edu

Abstract. Video generation is central to AI-powered content creation. The alignment with human preferences is one of the key metrics for measuring the quality of the generated videos. Despite significant progress in visual quality, three key challenges remain: 1) The reliability of reward signals is constrained by the quality of human preference data, which is often corrupted by subjective noise and bias. 2) Standard scalar reward models collapse multi-aspect human preference into a single value, leading to the loss of dynamic trade-offs across multiple dimensions of human preference. 3) In policy optimization, the widely adopted KL-divergence imposes only local constraints, failing to capture holistic human preference. To address these challenges, we propose a unified, preference-aware learning framework for video generation. First, we propose elite-guided filtering to calibrate preference data and construct reliable supervision for reward-model training. We then model video quality as a multidimensional reward distribution to capture the uncertain nature of human preference, and use the Wasserstein distance to align it with the empirical human preference distribution. Finally, we introduce Wasserstein-based distributional alignment in GRPO, guiding the policy's video generation to match the global structure of human video preference. Experiments on reward modeling and video generation show that our approach improves the reliability of reward signals and the perceptual consistency of generated videos. Our code is available at https://github.com/alignhs26/ahs.

Keywords: Video Generation, Preference Calibration, Reward Modeling, Human Preference Alignment

1 Introduction

Video generation has become a core technology in AI-driven content creation, with growing applications in film and television production, virtual interaction, and digital simulation. Recent methodological advances in deep generative modeling have substantially improved the temporal coherence and perceptual fidelity

Fig. 1: Our framework produces high-quality text-to-video generation that closely aligns with human preferences, demonstrating significant improvements in both perceptual plausibility and semantic fidelity.

of synthesized videos [8,11,41]. This progress is accelerating adoption in creative and interactive applications [5, 17, 19]. However, as baseline generation quality approaches basic plausibility requirements, further performance gains increasingly depend on alignment with human preferences. Given the inherently multi-dimensional and context-sensitive nature of human judgment, achieving effective preference alignment has emerged as a critical bottleneck that determines the quality of generated video content.

To address this challenge, human feedback-based alignment frameworks developed initially for language models, including reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO), have been adapted to video generation [18, 36]. Recent work has made several advancements to particularly address the inherent variability and expressive richness of human preferences, which often impede consistent recovery of underlying intent. One line of work enhances the granularity and structure of human feedback in annotation data, translating vague, holistic preference into structured and traceable supervisory signals to enable finer-grained and more transparent alignment [18, 32, 36]. Another line aims to reduce or eliminate manual annotation by either bootstrapping reward models from limited human labels [32] or reframing preference learning as a generative process, enabling scalable and evolvable preference acquisition [35].

Although existing studies have explored various strategies to improve alignment, they continue to face three key issues that prevent genuine human alignment: 1) The reliability of the reward signal is limited by the quality of human preference data, which is often affected by subjective noise and bias, leading to deviation from the accurate preference distribution [28]. 2) Existing reward models [7,37] typically assume that video quality can be represented by a single scalar value and optimize it using objective functions such as pairwise ranking or binary classification. In practice, human preferences arise from multiple interrelated dimensions, including but not limited to semantic plausibility, visual naturalness, and temporal coherence. These dimensions exhibit complex couplings and trade-offs, and forcing them into a single scalar not only discards essential structural information but also frequently results in inconsistent preference representations [7,37]. 3) Additionally, the KL divergence objective used

in RLHF and GRPO [12, 24, 38] applies constraints only locally within each denoising step of the diffusion process, functioning as a pointwise and instantaneous mechanism for distributional alignment. It does not account for temporal dependencies or semantic evolution across steps, and therefore cannot reflect the holistic human preference for video. To avoid penalties from local distributional deviations, the policy tends to select low-risk denoising trajectories, often resulting in motion-conservative video outputs that fail to capture complex dynamic interactions. Although such samples may match the reference behavior in local statistical properties, they consistently diverge from human preference in terms of global visual plausibility and temporal coherence.

To address the alignment discrepancy stemming from the three issues discussed above, we propose a unified framework for human-aligned video preference learning that systematically tackles their root causes. Specifically, as shown in Fig. 2, our approach proceeds in three stages: 1) **Elite-Guided Preference Calibration:** To extract high-quality and reliable supervision signals from noisy and biased human annotations, we avoid treating all samples equally. Instead, we quantify the intrinsic reliability of each sample using consistency metrics. High-confidence samples identified through this process are used to train an initial scoring model, which is then employed as a scorer to re-evaluate the quality of low-confidence data and recalibrate the overall preference distribution, thereby providing higher-quality training signals. 2) **Preference-Aware Reward Modeling:** To address the limitation of reward models that rely on a single scalar and thus fail to capture the multidimensional couplings within human preference, we model video quality as a multidimensional reward distribution defined over multiple perceptual and semantic dimensions rather than a single value. Based on this formulation, we employ the Wasserstein distance as a distribution-level alignment objective to align the model-predicted reward distribution with the empirical distribution of human preference, ensuring that the predicted rewards more faithfully reflect the diversity of real human preference. 3) **Preference-Guided Policy Optimization:** We further extend the concept of distributional alignment to the policy optimization stage by introducing the Wasserstein distance as an alternative to the KL divergence within the single-step update framework of Group Relative Policy Optimization (GRPO) [24]. In this stage, the policy governs the generation of video sequences, and the human video-preference distribution serves as the reference for alignment. During each policy update, we align the global characteristics of the generated video distribution rather than merely matching local statistical densities. The Wasserstein distance acts as a robust behavioral alignment metric, guiding the policy's output distribution towards the empirical distribution of human video preferences.

Our contribution can be summarised as follows:

1. We propose a unified framework for human-aligned video preference learning that systematically addresses the misalignment across three levels: data reliability, reward representation, and policy alignment.

2. We propose an elite-guided calibration framework to evaluate annotation reliability. Building on this, we leverage this representation to achieve alignment in both reward modeling and policy optimization respectively.
3. We conduct comprehensive experiments on reward modeling and video generation across multiple metrics, verifying the effectiveness of our framework in aligning the learned reward distribution with human preference.

2 Related Work

Data calibration Human preference data underpin the alignment of large language models (LLMs) and video generation systems; however, its scalability is limited by the cost and inconsistency of human annotations. Recent studies address these challenges through data calibration, emphasizing both the selective use of high-quality data and the correction of noise in low-quality supervision. Automated surrogates, such as reward models (e.g., the helpful–harmless objective [1]) and lightweight alignment strategies [42], reduce reliance on costly human feedback by capturing alignment signals at scale. Meanwhile, data-centric approaches focus on quality-based selection and reweighting. DEITA [15], which quantifies complexity, quality, and diversity to identify high-impact samples, shows that 6K curated examples can outperform ten times more random data. Complementary studies on filtering, clustering, and IFD-based self-calibration [10, 20, 21] further show that alignment gains mainly come from data quality and calibration, rather than data volume.

Human Feedback Alignment Recent studies in visual generation emphasize the collection of human preference data, training reward models, and optimization of generative models. VideoScore [7] builds a large multi-dimensional human rating dataset and an automatic scorer to simulate feedback. VisionReward [37] models fine-grained, interpretable preferences such as composition, motion, and prompt consistency for better alignment. VideoAlign [13] directly applies preference optimization to video generation, while UnifiedReward [34] proposes a cross-task, cross-modal reward model for broader generalization. Overall, the field is moving toward fine-grained, multi-dimensional, and unified human preference alignment, though challenges remain in data scale and reward-model generalization.

Video Generation Video generation has advanced markedly, driven by diffusion models [2, 8, 22, 25] and flow-matching models [11, 16, 30, 40]. Despite steady gains in resolution and stability, generated results often remain misaligned with human preferences, manifesting as semantic misinterpretation, visual content deviating from intent descriptions, and mismatches between perceived quality and user judgments. To mitigate this alignment gap, human feedback-based approaches have been transferred from language to visual generation. Methods based on DPO [14, 36] apply Direct Preference Optimization to align both visual quality and semantic relevance. Methods based on GRPO [12, 38] adopt

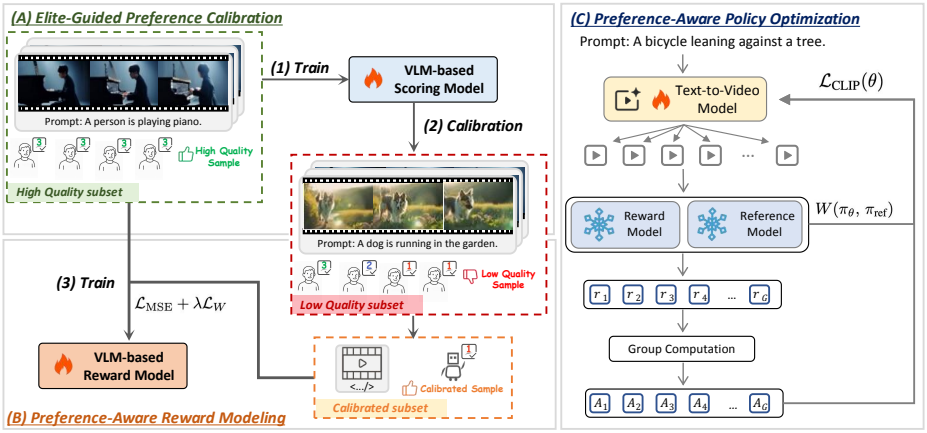


Fig. 2: Overview of our framework. (A) Elite-Guided Preference Calibration trains a scoring model on high-consistency annotations and calibrates the low-quality subset. (B) Preference-Aware Reward Modeling fine-tunes a VLM-based reward model with an MSE loss and a Wasserstein alignment loss. (C) Preference-Guided Policy Optimization applies Wasserstein-based GRPO to align group-wise video generation with the reference reward model.

Group Relative Policy Optimization across diffusion and flow-matching models, unifying reward models and stabilizing policy optimization for aligning human preferences in video generation. Regardless of the optimization algorithm, the pivotal factor remains the quality and distribution of human preference data, because aligning the model output to human subjective standards ultimately depends on capturing and modelling those preferences accurately.

3 Method

Our method proceeds in three stages (Fig. 2). Elite-Guided Preference Calibration (Sec. 3.1) first partitions annotations by intrinsic agreement, trains a scorer on the high-consistency subset, and calibrates low-consistency samples to expand a high-confidence training set. Preference-Aware Reward Modeling (Sec. 3.2) then fits a VLM-based reward model on the calibrated data. Preference-Guided Policy Optimization (Sec. 3.3) finally fine-tunes the video generator with Wasserstein-based GRPO, using the trained reward model to improve perceptual quality and text–video alignment.

3.1 Elite-Guided Preference Calibration

To extract high-quality and reliable supervision signals from noisy and biased human annotations, we propose the Elite-Guided Preference Calibration (EGPC) framework. The core idea is to treat human annotations with high inter-annotator

consistency as elite signals to construct a reliable scoring proxy, thereby enabling systematic calibration of the preference distribution at the full-dataset scale. The procedure consists of three stages:

High-Quality Subset Selection Given the original annotation dataset $\mathcal{D} = \{(v_i, p_i, \mathbf{x}_i)\}_{i=1}^{N}$, where v_i denotes a generated video, p_i represents the corresponding prompt, and $\mathbf{x}_i = \{x_{i1}, \ldots, x_{in}\}$ is the independent ratings provided by n human annotators on a discrete rating scale $\mathcal{S} = \{s_1, \ldots, s_k\}$.

We define a high-quality subset $\mathcal{D}_H$ based on annotation consistency. This subset aims to capture stable and reproducible human judgments, thereby providing trustworthy supervision for downstream modeling. Specifically, a sample i is included in $\mathcal{D}_H$ if it meets a standard of high consistency: that is, all annotators assign identical scores, or, if a discrepancy exists, there is at most a single outlier rating x_{ij} which satisfies the condition that its absolute difference from the mode of the ratings does not exceed 1, i.e., $|x_{ij} - \text{mode}(\mathbf{x}_i)| \leq 1$. Here, $\text{mode}(\mathbf{x}_i)$ denotes the most frequently occurring rating among the annotators for sample i. All remaining samples that do not meet this high consistency standard constitute the low-quality subset $\mathcal{D}_L = \mathcal{D} \setminus \mathcal{D}_H$. This selection process ensures that the samples in $\mathcal{D}_H$ possess a high degree of annotation consistency and reliability, thereby providing a more robust and trustworthy supervisory signal.

Elite Scorer Training Using $\mathcal{D}_H$, we train a parameterized scoring function $f_\theta : (v, p) \mapsto \hat{x} \in \mathbb{R}$ by minimizing the empirical risk between predictions and human ratings:

$$\theta^* = \arg\min_\theta \frac{1}{|\mathcal{D}_H|} \sum_{(v_i, p_i, \mathbf{x}_i) \in \mathcal{D}_H} \ell\big(f_\theta(v_i, p_i),\, \text{agg}(\mathbf{x}_i)\big), \tag{1}$$

where $\text{agg}(\cdot)$ is an aggregation operator, and ℓ is a regression loss. The resulting model f_{θ^*}, referred to as the Elite Scorer, aims to approximate the mapping underlying expert-consensus judgments.

Full-Set Preference Calibration The trained elite scorer is applied to the low-quality subset $\mathcal{D}_L$ to obtain calibrated scores: for each $(v_i, p_i, \mathbf{x}_i) \in \mathcal{D}_L$, compute $\hat{x}_i = f_{\theta^*}(v_i, p_i)$, and construct the calibrated subset $\mathcal{D}_L^* = \{(v_i, p_i, \hat{x}_i) \mid (v_i, p_i, \mathbf{x}_i) \in \mathcal{D}_L\}$. Finally, we merge the original high-quality samples with the calibrated ones to form an augmented high-quality supervision set:

$$\mathcal{D}_H^{\text{aug}} = \mathcal{D}_H \cup \mathcal{D}_L^*. \tag{2}$$

This augmented set preserves the structural integrity of expert consensus while substantially expanding the scale of effective supervision, thus providing a consistent preference signal foundation for subsequent reward modeling.

3.2 Preference-Aware Reward Modeling

To overcome the limitation of single-scalar rewards that fail to capture the complexity of multidimensional human preferences, we model video quality as a probability distribution over a K-dimensional score space rather than a single value. For each video–prompt pair (v_i, p_i), let $\mathbf{y}_i \in \mathbb{R}^K$ denote its calibrated multidimensional human-preference label. The empirical target distribution is represented as $\delta_{\mathbf{y}_i}$.

For a reward model parameterized by θ, we predict a conditional distribution $p_\theta(\mathbf{z} \mid v_i, p_i)$, where $\mathbf{z} \in \mathbb{R}^K$ denotes a possible multidimensional preference score. To ensure consistency between the predicted and empirical distributions, we introduce Wasserstein distance [4] as a complementary alignment objective. For a given sample $(v_i, p_i, \mathbf{y}_i)$, the distribution alignment loss is defined as:

$$\mathcal{L}_W(v_i, p_i) = W_p\big(p_\theta(\mathbf{z} \mid v_i, p_i),\, \delta_{\mathbf{y}_i}\big). \tag{3}$$

The total training objective combines conventional regression supervision with this distributional alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_W, \tag{4}$$

where $\lambda > 0$ controls the relative importance of the Wasserstein alignment term. The mean squared error (MSE) term is defined as:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{(v,p,\mathbf{y})\sim\mathcal{D}} \left\| f_\theta(v, p) - \mathbf{y} \right\|_2^2, \tag{5}$$

where $f_\theta(v, p)$ is the predicted multidimensional reward vector.

3.3 Preference-Guided Policy Optimization

To overcome the limitation of KL divergence in failing to capture holistic human preferences, we extend the concept of distributional alignment to the policy optimization stage by proposing Wasserstein-based Group Relative Policy Optimization (WGRPO). Specifically, we replace the KL divergence with the Wasserstein distance within the single-step update framework of GRPO. Unlike the KL divergence, which only matches local statistical densities, the Wasserstein distance aligns the global characteristics of the generated video distribution, serving as a behavioral alignment constraint that guides the policy's overall output to progressively approximate human preference.

Specifically, given a prompt $\mathbf{c}$, the old policy $\pi_{\theta_{\text{old}}}$ first samples a group of G candidate sequences $\{\mathbf{o}_1, \ldots, \mathbf{o}_G\}$ following the group sampling paradigm of DeepSeek-R1 [26], where each candidate $\mathbf{o}_i$ corresponds to a trajectory consisting of state–action pairs $\{(s_{t,i}, a_{t,i})\}_{t=1}^{T}$. For each timestep, we compute the importance ratio between the new and old policies as

$$\rho_{t,i} = \frac{\pi_\theta(a_{t,i}|s_{t,i})}{\pi_{\theta_{\text{old}}}(a_{t,i}|s_{t,i})}. \tag{6}$$

The reward r_i obtained for each trajectory is normalized within the group to form a group-relative advantage,

$$A_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G)}, \tag{7}$$

which stabilizes optimization, reduces reward-scale variance, and preserves the relative preference relationships among candidates. Building upon the PPO framework [23], the surrogate objective for one timestep (t, i) is defined as

$$L_{t,i}(\theta) = \min\big(\rho_{t,i}\, A_i,\ \text{clip}(\rho_{t,i},\, 1-\epsilon,\, 1+\epsilon)\, A_i\big), \tag{8}$$

where the clipping term limits the update magnitude and prevents excessively large policy shifts. Averaging over all timesteps of trajectory $\mathbf{o}_i$ yields the per-sample objective,

$$\mathcal{L}_i(\theta) = \frac{1}{T} \sum\nolimits_{t=1}^{T} L_{t,i}(\theta), \tag{9}$$

and the final group-level objective is obtained by taking the expectation over all groups sampled from the old policy,

$$\mathcal{L}_{\text{CLIP}}(\theta) = \mathbb{E}_{\substack{\{\mathbf{o}_i\}\sim\pi_{\theta_{\text{old}}} \\ \{a_{t,i}\}\sim\pi_{\theta_{\text{old}}}}} \left[\frac{1}{G} \sum\nolimits_{i=1}^{G} \mathcal{L}_i(\theta) \right]. \tag{10}$$

To further align the global behavioral characteristics of the new policy with human-preferred generation patterns, we incorporate a distributional alignment term based on the Wasserstein distance between the output distribution π_θ and the reference policy π_{ref}. The Wasserstein term measures the minimal transport cost required to transform one output distribution into another under a ground metric d on the output space $\mathcal{Y}$, serving as a behavioral alignment constraint that captures global semantic and perceptual consistency. Combining the clipped GRPO surrogate with a Wasserstein alignment penalty yields the following minimization objective:

$$\mathcal{L}_{\text{WGRPO}}(\theta) = -\mathcal{L}_{\text{CLIP}}(\theta) + \beta\, W(\pi_\theta,\, \pi_{\text{ref}}), \tag{11}$$

where $\beta > 0$ controls the strength of the distributional alignment penalty.

4 Experiments

4.1 Dataset

This experiment employs two publicly available benchmark datasets: VidProM [33] for text-conditional video generation, and Video-Bench [6] for reward modeling targeting human alignment.

Table 1: Comparison of sub-dimension scores before and after alignment for Wan, CogVideo and ModelScope. Fine-tuning with our two-dimensional calibrated reward models shows stronger human-preference alignment in generated videos. RM Type denotes Reward Model Type.

Model	RM Type	Motion smooth.	Dynamic degree	Spatial relation	Scene	Appear. style	Subject consist.	Back. consist.	Avg.
Wan2.1	Raw (Full)	88.25	35.58	61.23	69.65	42.21	88.89	91.32	68.16
T2V-1.3B	Calibrated	89.32	36.48	63.23	71.32	42.25	91.90	92.55	**69.58**
CogVideoX	Raw (Full)	86.53	31.33	54.52	42.15	75.58	87.35	87.32	66.39
2B	Calibrated	90.25	31.99	56.89	47.15	76.95	86.45	88.57	**68.32**
ModelScope	Raw (Full)	81.53	27.77	59.33	40.39	70.58	82.12	85.33	63.86
	Calibrated	82.34	29.54	63.12	45.56	73.98	80.25	86.66	**65.92**

Video-Bench Video-Bench [6] provides a collection of annotated video–prompt pairs, supporting multi-dimensional quality assessment. The dataset comprises 8799 unique video samples, generated by four open-source and three commercial text-to-video models. Each sample is rated independently by four annotators across nine dimensions, specifically:

- Five alignment dimensions (e.g., Action, Color Consistency), assessing semantic fidelity between generated content and the textual instruction;
- Four video quality dimensions (e.g., Motion Effects, Aesthetic Quality), evaluating perceptual quality and physical plausibility.

VidProM For policy training, we curate 50,000 text prompts from VidProM [33], a collection of community-generated prompts from the official Pika Labs Discord channel from October 2023 to March 2024. The selected prompts cover diverse video-generation scenarios, actions, and visual styles, providing varied text conditions for policy learning.

4.2 Experimental Setting

Dataset Calibration Following the procedure in Sec. 3.1, we first isolate a high-consistency subset from Video-Bench. The subset is selected via inter-annotator agreement and annotation stability and serves as the foundation for training an elite scoring model. Using Qwen2-VL-2B [31] as the backbone, this model learns to predict video quality aligned with human perceptual judgments. Subsequently, we reweight the remaining samples by the model's confidence and perform hard filtering to discard low-scoring outliers, thereby suppressing tail-distributed noise and yielding a calibrated training distribution.

Reward Modeling The reward model is implemented by fine-tuning Qwen2-VL-2B [31] with LoRA (rank = 8) [9], batch size 32, and learning rate 2×10^{-6}. All other hyperparameters follow VideoAlign [13]. Consistent with practices in

Table 2: Comparison of Reward Models on the pairwise accuracy on the Action and Motion Effects dimensions under two data scales (Equal and Full).

Scale	Action		Motion Effects	
	w/ Ties	w/o Ties	w/ Ties	w/o Ties
Raw(Equal)	54.65	74.24	67.68	80.30
Curated(Equal)	**59.60**	**78.79**	**68.69**	**84.85**
Raw(Full)	58.59	80.30	65.66	86.36
Calibrated	**60.61**	**83.33**	**71.72**	**87.88**

previous work [7, 12], we evaluate on a held-out set of 100 high-fidelity video pairs, reporting pairwise accuracy both with ties and excluding ties [3].

Preference-Guided Model Adaptation We adapt three open-source text-to-video base models, Wan2.1-T2V-1.3B [27], CogVideoX-2B [39], and ModelScope-T2V [29], using reward signals from the preference-aware reward model trained in Sec. 3.2. Training is conducted within the Video-Bench framework, ensuring fair comparison and reproducibility. We refer to this stage as preference-guided policy adaptation because it applies GRPO-style updates to align generated videos with calibrated human-preference rewards. To mitigate prompt leakage, we use GPT-4o to rewrite and extend the original test prompts. Each original prompt is paraphrased into three stylistically and semantically distinct variants, yielding a more robust and generalizable evaluation set.

Further Details All experiments are conducted on 8 NVIDIA H100 GPUs. For policy optimization, we use a sampling timestep $T = 20$ and a group size $G = 16$. We use LoRA with $\alpha = 64$ and $r = 32$.

4.3 Main Results

Reward Modeling Performance Table 2 reports pairwise accuracy for the reward model on Action Consistency and Motion Effectiveness. Training on curated calibrated data improves accuracy over training on the raw data, showing the benefit of calibration. With the full dataset, the model reaches 60.61 with ties and 83.33 without ties on Action Consistency, and 71.72 with ties and 87.88 without ties on Motion Effectiveness. These gains indicate that calibration helps the reward model evaluate both action consistency and motion effectiveness. As illustrated in Fig. 1, our method achieves superior alignment with high-quality human preference, yielding outputs with high perceptual fidelity and overall performance improvements.

Video Generation Quality Improvement Table 1 compares three open-source video generation models before and after optimization with our calibrated

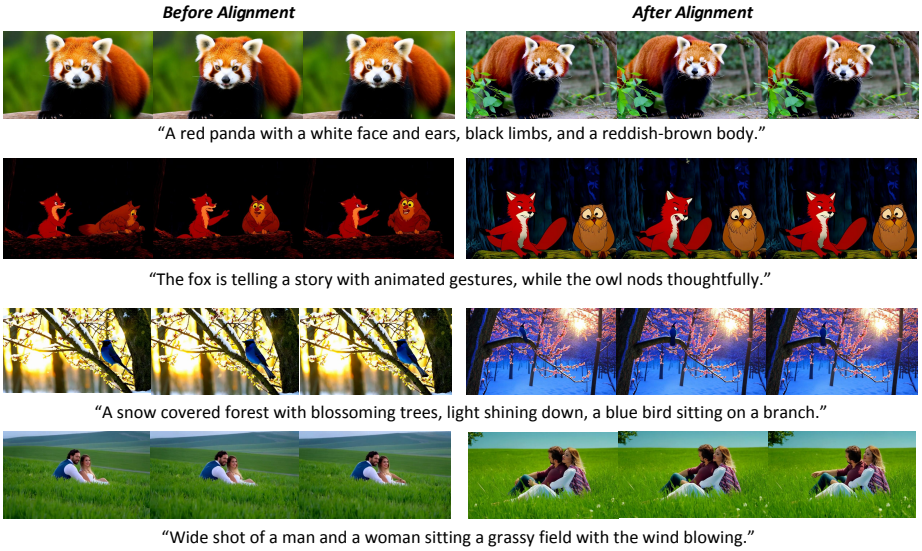


Fig. 3: Alignment effect comparison: The left column displays results from models before alignment, while the results after our alignment are shown on the right.

reward model. The optimized models improve mainly on motion-related metrics, including Motion Smoothness and Dynamic Degree. For example, CogVideoX improves from 86.53 to 90.25 in Motion Smoothness and from 31.33 to 31.99 in Dynamic Degree. The total composite score also increases by 2.1% for Wan2.1, 2.9% for CogVideoX, and 3.2% for ModelScope. The main exception is a small drop in Subject Consistency, suggesting a trade-off from selecting samples that favor dynamic quality over static content consistency.

4.4 Ablation Study

To validate the effectiveness of $\mathcal{L}_{\mathrm{W}}$, we conduct a two-part ablation study focusing on both Reward Modeling and Policy Optimization. The results in Tab. 3 and Tab. 4 demonstrate the benefits of choosing the Wasserstein Distance.

Table 3: Ablation study on Wasserstein distance loss for reward modeling.

$\mathcal{L}_{\mathrm{W}}$	Action		Motion Effects	
	w/ Ties	w/o Ties	w/ Ties	w/o Ties
✗	58.98	81.25	68.33	86.11
✓	**60.61**	**83.33**	**71.72**	**87.88**

Table 4: Ablation study on Wasserstein distance constraint for video generation.

	Motion smooth.	Dynamic degree	Spatial relation	Scene	Appear. style	Subject consist.	Back. consist.	Avg.
✗	89.10	36.29	59.78	69.01	41.32	88.98	91.33	67.97
KL	88.89	36.50	59.40	69.11	41.78	89.05	91.45	68.02
Wass.	89.32	36.48	63.23	71.32	42.25	91.90	92.55	**69.58**

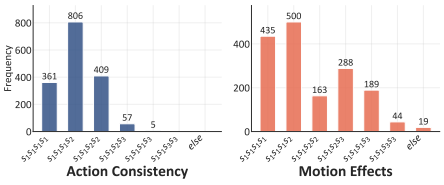
Reward Modeling Performance Table 3 shows the effect of $\mathcal{L}_{\mathrm{W}}$ on the reward model's pairwise accuracy for Action Consistency and Motion Effectiveness. Adding $\mathcal{L}_{\mathrm{W}}$ improves all reported metrics. For Motion Effectiveness, accuracy increases from 68.33 to 71.72, a gain of 3.39. For Action Consistency, accuracy improves from 58.98 to 60.61. Wasserstein distance provides a distribution-level constraint by measuring the transport cost between the predicted reward distribution and the human-preference reference distribution, rather than only penalizing local deviations. This property is well suited to video preferences that involve global semantic consistency, temporal coherence, and multi-dimensional quality attributes, helping the reward model better distinguish quality differences in complex dynamic content.

Policy Optimization for Video Generation Table 4 compares video generation performance under three regularization settings during reinforcement learning: no regularizer, KL Divergence, and Wasserstein Distance. The model optimized with Wasserstein Distance achieves the highest Total score of 69.58, exceeding both the non-regularized baseline, 67.97, and the KL-regularized model, 68.02. The main gains appear in dynamic and consistency-related metrics, including Motion Smoothness, 89.32, Spatial Relation, 63.23, and Background Consistency, 92.55. Although KL regularization performs slightly better on Dynamic Degree, Wasserstein Distance gives stronger results across multiple metrics related to motion and scene coherence. This suggests that it better aligns the video generation policy distribution with the target reward distribution during optimization, which is also reflected in the examples in Fig. 3.

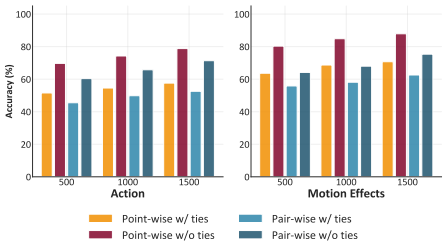
4.5 Further Analysis

Annotation Consistency and Data Quality Fig. 4a shows the distribution of sample consistency across different dimensions in the original dataset. In the action consistency dimension, besides 1,167 high quality samples, there are 471 low consistency samples; in the Motion Effects dimension, 703 noisy samples are nearly comparable in number to the 935 high quality samples. As shown, noisy samples account for a substantial proportion: 28.8% and 42.9% of the total samples in the Action and Motion dimensions, respectively.

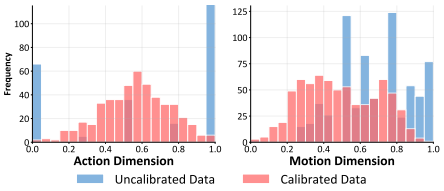
From empirical evidence, ratings from a single annotator are more susceptible to subjective bias and fail to capture a stable, population-level preference



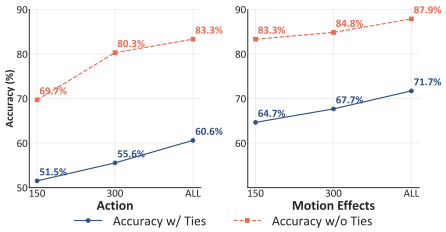
(a) Distribution of samples with different annotator-consistency levels for (left) action and (right) motion.



(b) Performance comparison across different data scales shows that under noisy conditions, point-wise training achieves more stable and reliable accuracy than pair-wise training.



(c) Score distribution before and after Calibration.



(d) Relationship between the number of calibrated training samples and accuracy on action consistency and motion effects Dimensions.

Fig. 4: Further analysis of data quality, training robustness, and calibration effects.

distribution. Even with multiple annotators, a subset of high-complexity samples still induces substantial inter-rater disagreement. Such noisy data exhibit high-variance human preferences that do not faithfully reflect genuine, high-quality preference signals, and therefore are unsuitable for model training.

Paradigm Comparison on Robustness to Noise In noisy samples, different annotators often assign different scores to the same instance. We investigate how this phenomenon influences the choice of training paradigms in Fig. 4b. Experimental results show that as the noise level increases, the performance of the pairwise training paradigm degrades substantially. In such cases, when training reward models with noisy data, the traditional pointwise training approach proves more effective in handling score inconsistencies. This is because pairwise training relies on accurate pairwise comparisons to optimize preference learning, but in noisy environments, these pairwise relationships can become distorted, hindering the optimization process. In contrast, the pointwise training method independently optimizes each sample's score without relying on pairwise comparisons, thereby exhibiting greater stability and robustness with noisy data.

Visualization on Calibration Effect We analyze the importance of score calibration, as illustrated in Fig. 4c, which shows the changes in score distributions of noisy data before and after calibration. For the original multi-annotator

Table 5: Policy optimization comparison using reward models trained on raw data and calibrated data. RM Type denotes Reward Model Type. We report results for DanceGRPO and FlowGRPO, where each method is evaluated with the original raw reward model and the reward model trained on the calibrated preference distribution.

Method	RM Type	Motion smooth.	Dynamic degree	Spatial relation	Scene	Appear. style	Subject consist.	Back. consist.	Avg.
Dance	Raw(Full)	88.10	36.40	63.90	71.20	42.10	89.50	91.70	68.99
GRPO	Calibrated	**88.30**	**37.20**	**65.60**	**72.10**	**43.30**	**90.00**	**93.30**	**69.97**
Flow	Raw(Full)	89.00	38.10	64.90	71.80	42.70	90.30	92.20	69.86
GRPO	Calibrated	**89.40**	**39.40**	**66.20**	**72.60**	**43.90**	**90.90**	**93.70**	**70.87**

scores, we use the average rating as the initial score label. In both the action and motion effects dimensions, we observe high-score samples whose actual quality is poor. The action consistency dimension also contains low-score samples with high actual quality. Due to the lack of consistency among annotators, it is impractical to assign a unified label to these samples. With a robust calibration model, we effectively mitigated such scoring inconsistencies. The calibrated score distributions become smoother and more continuous, substantially reducing the presence of outlier samples in the original distribution.

Which leads to performance improvement? Quality vs. Quantity. We investigate how increasing the proportion of calibrated samples affects model performance. First, we train the scoring model using absolutely high-quality samples, and then progressively calibrate noisy samples of different scales. As shown in Fig. 4d, when the number of calibrated samples increases from 150 to 300 and eventually covers all noisy data, the performance of the reward model trained on the progressively expanded calibrated dataset consistently improves. Specifically, in the action consistency dimension, the accuracy (with ties and without ties) increases by 9.1 and 13.6 points, respectively, while in the Motion dimension, it improves by 7.0 and 4.6 points, respectively.

Generalization Across Policy Optimizers Table 5 evaluates our calibrated distributional reward model with FlowGRPO [12] and DanceGRPO [38] on Wan2.1-T2V-1.3B. Replacing the raw reward model with the calibrated one consistently improves both optimizers, increasing the average score respectively. These results suggest that our calibration strategy provides a generally useful reward signal rather than an optimizer-specific gain.

5 Conclusion

In this paper, we propose a unified preference-aware video generation framework to align video generation with human perceptual judgments of quality. Our approach introduces three core mechanisms: the elite-guided filtering for reliable

preference data calibration, a multidimensional reward distribution to capture the nature of human preference and the Wasserstein-based distributional alignment integrated into GRPO to enforce alignment with the global structure of human preference. Experimental results show that our framework significantly improves the reliability of reward signals and enhances the consistency of human preference in the generated videos.

Acknowledgements

The work described in this paper was partially supported by InnoHK initiative, The Government of the HKSAR, and Laboratory for AI-Powered Financial Technologies.

References

1. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. CoRR (2022)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22563–22575. IEEE Computer Society (2023)
3. Deutsch, D., Foster, G., Freitag, M.: Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. arXiv preprint arXiv:2305.14324 (2023)
4. Feydy, J., Séjourné, T., Vialard, F.X., Amari, S.i., Trouvé, A., Peyré, G.: Interpolating between optimal transport and mmd using sinkhorn divergences. In: The 22nd international conference on artificial intelligence and statistics. pp. 2681–2690. PMLR (2019)
5. Fu, J., Ge, X., Zheng, K., Karatzoglou, A., Arapakis, I., Xin, X., Ni, Y., Jose, J.M.: Llmpopcorn: exploring llms as assistants for popular micro-video generation. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 12007–12011. IEEE (2026)
6. Han, H., Li, S., Chen, J., Yuan, Y., Wu, Y., Leong, C.T., Du, H., Fu, J., Li, Y., Zhang, J., Zhang, C., Li, L.j., Ni, Y.: Video-bench: Human-aligned video generation benchmark. arXiv preprint arXiv:2504.04907 (2025), cVPR 2025
7. He, X., Jiang, D., Zhang, G., et al.: Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In: EMNLP (2024)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
10. Li, M., Zhang, Y., Li, Z., Chen, J., Chen, L., Cheng, N., Wang, J., Zhou, T., Xiao, J.: From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 7595–7628 (2024)

11. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: 11th International Conference on Learning Representations, ICLR 2023 (2023)
12. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. Advances in neural information processing systems **38**, 40783–40818 (2026)
13. Liu, J., Liu, G., Liang, J., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
14. Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8009–8019 (2025)
15. Liu, W., Zeng, W., He, K., Jiang, Y., He, J.: What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In: ICLR (2024)
16. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
17. Meng, R., Zhang, X., Li, Y., Ma, C.: Echomimicv2: Towards striking, simplified, and semi-body human animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5489–5498 (2025)
18. Meng, X., Zhang, Z., Zhang, Z., Liao, J., Qin, L., Wang, W.: Identity-grpo: Optimizing multi-human identity-preserving video generation via reinforcement learning. arXiv preprint arXiv:2510.14256 (2025)
19. Ni, Y., Cheng, Y., Liu, X., Fu, J., Li, Y., He, X., Zhang, Y., Yuan, F.: A content-driven micro-video recommendation dataset at scale. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management. pp. 6486–6491 (2025)
20. Qin, Y., Yang, Y., Guo, P., Li, G., Shao, H., Shi, Y., Xu, Z., Gu, Y., Li, K., Sun, X.: Unleashing the power of data tsunami: A comprehensive survey on data assessment and selection for instruction tuning of language models. CoRR (2024)
21. Qin, Y., Yang, Y., Guo, P., et al.: Unleashing the power of data tsunami: A comprehensive survey on data assessment and selection for instruction tuning. Transactions on Machine Learning Research (2024)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
23. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
24. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
25. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: The Eleventh International Conference on Learning Representations (2022)
26. Team, D.: Deepseek-r1. `https://github.com/deepseek-ai/deepseek-r1` (2025)
27. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. CoRR (2025)

28. Wang, J., Liu, Y., Li, B.: Reinforcement learning with perturbed rewards. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 6202–6209 (2020)
29. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
30. Wang, J., Zeng, X., Qiang, C., Chen, R., Wang, S., Wang, L., Zhou, W., Cai, P., Zhao, J., Li, N., et al.: Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. arXiv preprint arXiv:2506.19774 (2025)
31. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
32. Wang, Q., Liu, J., Liang, J., Jiang, Y., Zhang, Y., Chen, J., Zheng, Y., Wang, X., Wan, P., Yue, X., et al.: Vr-thinker: Boosting video reward models through thinking-with-image reasoning. arXiv preprint arXiv:2510.10518 (2025)
33. Wang, W., Yang, Y.: Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. Advances in Neural Information Processing Systems **37**, 65618–65642 (2024)
34. Wang, Y., Zang, Y., Li, H., Jin, C., Wang, J.: Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236 (2025)
35. Wu, J., Gao, Y., Ye, Z., Li, M., Li, L., Guo, H., Liu, J., Xue, Z., Hou, X., Liu, W., et al.: Rewarddance: Reward scaling in visual generation. arXiv preprint arXiv:2509.08826 (2025)
36. Wu, Z., Kag, A., Skorokhodov, I., Menapace, W., Mirzaei, A., Gilitschenski, I., Tulyakov, S., Siarohin, A.: DenseDPO: Fine-grained temporal preference optimization for video diffusion models. NeurIPS (2025)
37. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., et al.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059 (2024)
38. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
39. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025)
40. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. CoRR (2024)
41. Zhang, L., Agrawala, M.: Packing input frame context in next-frame prediction models for video generation. arXiv preprint arXiv:2504.12626 (2025)
42. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al.: Lima: Less is more for alignment. Advances in Neural Information Processing Systems **36**, 55006–55021 (2023)