

SUPPLEMENTARY MATERIAL: DUAL-IPO: DUAL-ITERATIVE PREFERENCE OPTIMIZATION FOR TEXT-TO-VIDEO GENERATION

A Human Annotation	16
A.1 Annotator selection	16
A.2 Annotation Guideline	16
A.3 Dataset distribution	17
A.4 Annotation training and interface	18
B Additional Quantitative Analysis	18
B.1 All evaluation dimension results	18
B.2 Impact of Reward Model Scale on Accuracy	18
B.3 Impact of Reward Model Size on RL	19
C Additional Qualitative Comparison	19
D Use of LLM	19
E Societal Impacts	19
F Limitations	20

A HUMAN ANNOTATION

Prior to the final manual annotation, we undertook several steps to ensure the quantity, quality, and efficiency of the annotation process. In building the dataset, we focused on the diversity and complexity of prompts, creating annotated data across multiple dimensions and categories. We meticulously selected candidates for the annotation task and provided them with in-depth training sessions. We then designed a user-friendly interface and a comprehensive annotation program to streamline the process. Additionally, we developed a detailed annotation guide that covers all aspects of the task and highlights essential precautions. To further ensure consistency and accuracy among annotators, we conducted multiple rounds of annotation and implemented random sampling quality checks.

A.1 ANNOTATOR SELECTION

The effectiveness of annotated data hinges significantly on the capabilities of the individuals responsible for the annotation. As such, at the onset of the annotation initiative, we placed a strong emphasis on meticulously selecting and training a team of annotators to ensure their competence and objectivity. This process entailed administering a comprehensive evaluation to potential candidates, which was designed to gauge their proficiency in ten critical domains: domain expertise, tolerance for visually disturbing content, attention to detail, communication abilities, reliability, cultural and linguistic competence, technical skills, ethical awareness, aesthetic discernment, and motivation.

Given that the models under assessment could generate videos containing potentially distressing or inappropriate material, candidates were informed of this possibility in advance. Their participation in the evaluation was predicated on their acknowledgment of this condition, and they were assured of their right to withdraw at any time. Based on the results of the evaluation and the candidates' professional backgrounds, we sought to form a team that was diverse and well-rounded in terms of expertise and capabilities across the ten assessed areas. Ultimately, we selected a group of 10 annotators—comprising five men and five women, all of whom possessed bachelor's degrees. We conducted follow-up interviews with the selected annotators to reaffirm their suitability for the annotation task.

A.2 ANNOTATION GUIDELINE

When it comes to assessing human preferences in videos, the evaluation process is structured around three fundamental aspects: **semantic consistency**, **motion smoothness**, and **video faithfulness**. Each of these dimensions is meticulously examined independently to provide a comprehensive understanding of the video's overall quality. **Semantic consistency** ensures that the video's content aligns logically and contextually, while **motion smoothness** evaluates the smoothness and naturalness of movements within the video. **Video faithfulness**, on the other hand, assesses how well the video maintains a realistic and high-quality appearance. By breaking down the evaluation into these distinct yet interconnected aspects, we can achieve a more nuanced and accurate assessment of human preferences in video content.

- **Semantic Consistency** The consistency between the text and the video refers to whether elements such as the topic category (e.g., people or animals), quantity, color, scene description, and style mentioned in the text align with what is displayed in the video.
- **Motion Smoothness** The motion smoothness refers to whether the movements depicted are smooth, coherent, and logically consistent, particularly in terms of their adherence to physical laws and real-world dynamics. This includes evaluating if the actions unfold in a seamless and believable manner, without abrupt or unrealistic transitions, and whether they align with the expected behavior of objects or individuals in the given context.
- **Video Faithfulness** Faithfulness refers to the accuracy and realism of the visual content in the video. It examines whether depictions of people, animals, or objects closely match their real-world counterparts, without unrealistic distortions like severed limbs, deformed features, or other anomalies. High fidelity ensures the visual elements are believable and consistent with reality, enhancing the video's authenticity..

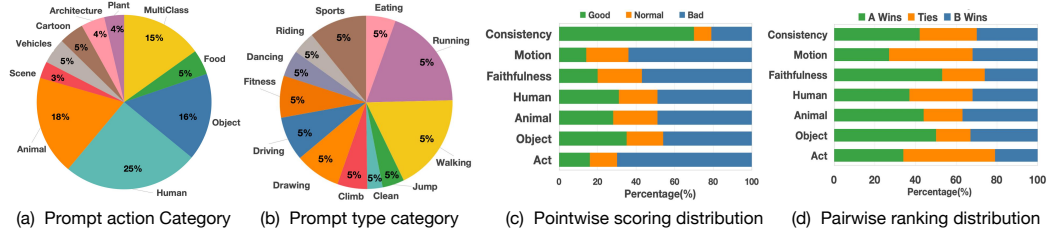


Figure A1: The distribution statistics of Human Preference Dataset on prompts in (a) and (b) as well as scoring/ranking in (c) and (d).

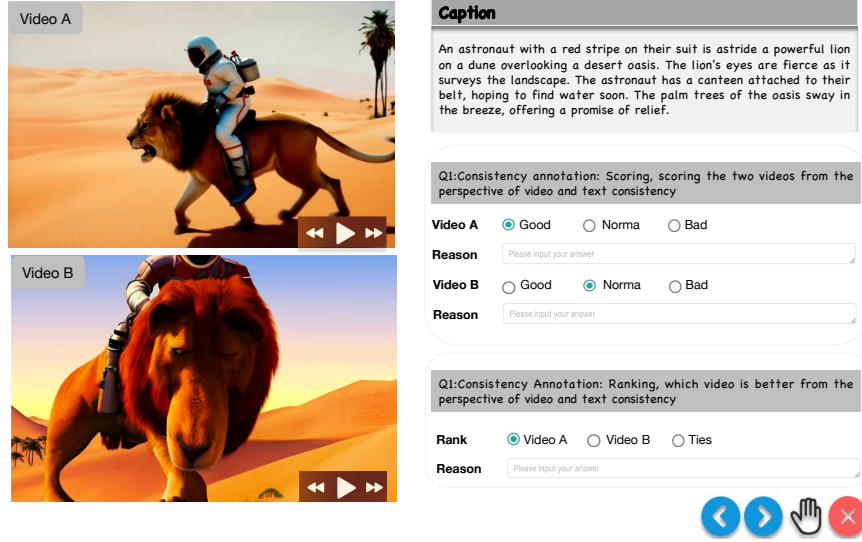


Figure A2: Our user interface presents one sample at a time to annotators, featuring four distinct icons for different functions.

Pointwise Annotation For each dimension, annotators assign one of three ratings:

- **Good:** No observable flaws; fully satisfies the criterion.
- **Normal:** Minor imperfections that do not severely impact perception.
- **Bad:** Significant violations that degrade quality or realism.

Annotations are performed independently across dimensions to avoid bias.

Pairwise Annotation Given two videos $\mathcal{V}_A$ and $\mathcal{V}_B$, annotators compare them under each dimension:

- For **semantic consistency**, determine which video better aligns with the intended context.
- For **motion smoothness**, identify the video with more natural and continuous motion.
- For **video faithfulness**, select the video exhibiting higher realism and fewer artifacts.

A ranked preference ($\mathcal{V}_A \succ \mathcal{V}_B$ or $\mathcal{V}_B \succ \mathcal{V}_A$) is recorded separately for each dimension. Ties are permitted only if differences are indistinguishable.

A.3 DATASET DISTRIBUTION

As shown in Fig. A1, the dataset includes over ten categories, such as Human, Architecture, Vehicles, and Animals, as well as more than ten different actions of human and animal. For each

prompt, the video generation model produces four variants with different random seeds. The dataset is iteratively expanded through an optimized regeneration strategy to improve preference alignment.

A.4 ANNOTATION TRAINING AND INTERFACE

We convened a comprehensive training session centered on our detailed user guidelines to ensure that the annotation team fully comprehends our goals and standards. During the training, we covered the purpose, precautions, standards, workload, and remuneration related to the annotation process. Additionally, we formally communicated to the annotators that the human annotation is strictly for research purposes and that the annotated data might be made publicly available in the future. We reached a mutual agreement with the annotators regarding the standards, workload, remuneration, and intended use of the annotated data. The criteria for assessing video faithfulness, text-video alignment, and motion smoothness are universal, which means that individual standards should not vary significantly. Consequently, we annotated several sample videos using our carefully designed annotation platform to ensure consistency among annotators. An overview of the developed annotation platform is shown in A2. Through this training, we also equipped the annotators with the necessary knowledge to conduct impartial and detailed human evaluations of video faithfulness, text-video alignment, and motion smoothness.

Model	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
CogVideoX-2B	96.78	96.63	98.89	97.73	59.86	60.82	61.68	83.37
CogVideoX-2B-KTO ₁	96.77	97.01	99.01	97.84	64.33	61.55	62	85.41
CogVideoX-2B-KTO ₂	96.81	97.23	99.21	98.03	68.76	62.35	61.77	85.63
CogVideoX-2B-KTO ₃	96.79	97.48	99.35	98.17	69.46	62.28	62.87	85.77

Model	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal style	Overall Consistency
CogVideoX-2B	62.63	98.00	79.41	69.90	51.14	24.8	24.36	26.66
CogVideoX-2B-KTO ₁	63.25	98.54	81.67	67.83	54.15	25.13	24.86	27.03
CogVideoX-2B-KTO ₂	63.44	98.97	83.24	70.16	54.35	25.47	26.44	27.37
CogVideoX-2B-KTO ₃	63.51	99.03	82.33	70.23	54.41	25.48	25.39	27.66

Table A1: Quantitative results on video assessment metrics. Display of comparison results between all indicators and baseline on vbench through multiple iterations.

B ADDITIONAL QUANTITATIVE ANALYSIS

B.1 ALL EVALUATION DIMENSION RESULTS

We present the evaluation metrics of VBench benchmark in 16 different dimensions in the Tab. A1. The results indicate that the performance shows a continuous upward trend in multiple training iterations, indicating that the training is dynamically stable. Although small fluctuations are observed, they can be ignored and will not affect the overall positive progress of the indicator.

Model Scale	Pairwise Accuracy	Pointwise Accuracy
Dual-IPO-Reward-13B	78.41	82.95
Dual-IPO-Reward-40B	81.33	85.57

Table A2: Accuracy of Reward Models on Validation Sets Across Different Scales. The reward models used in this study are finetuned from the ViLALin et al. (2023) multimodal large language model. This table evaluates their performance at different scales (13B and 40B) in terms of accuracy on pairwise and pointwise validation datasets. The results demonstrate that larger model scales yield higher accuracy for both pairwise ranking and pointwise scoring, indicating improved alignment with human preferences as the model size increases.

B.2 IMPACT OF REWARD MODEL SCALE ON ACCURACY

To evaluate the impact of model scale on reward performance, we experimented with two ViLA-based critic models of 13B and 40B parameters. Fine-tuned on paired and pointwise annotated datasets, these models were assessed on validation tasks, as shown in Tab. A2.

Results show that increasing the model size from 13B to 40B significantly improves accuracy, with the 40B model achieving 80.73% in paired ranking and 86.57% in pointwise scoring—2.92% and

2.62% higher than the 13B model. This suggests that larger critic models better capture human preferences by representing complex multimodal relationships.

Our ablation study further highlights the importance of model scale. While smaller models perform reasonably well, larger models significantly boost accuracy, benefiting downstream reinforcement learning tasks. These findings suggest that scaling the critic model is a simple yet effective way to enhance preference alignment and improve video generation quality.

Models	Total Score	Quality Score	Semantic Score	Subject Consist.	Background Consist.
CogVideoX-2B	80.91	82.18	75.83	96.78	96.63
w/IPO-Reward-13B	82.42	83.53	77.97	96.81	97.23
w/IPO-Reward-40B	82.52	83.63	78.07	96.83	97.37

Models	Temporal Flicker	Motion Smooth.	Aesthetic Quality	Dynamic Degree	Image Quality
CogVideoX-2B	98.89	97.73	60.82	59.86	61.68
w/IPO-Reward-13B	99.21	98.03	62.35	68.76	61.77
w/IPO-Reward-40B	99.19	98.09	62.38	68.92	62.01

Table A3: Evaluate the impact of the critic model on performance, verify the performance of the critic model results for different model sizes of 13B and 40B, a more accurate critic model leads to improved results.

B.3 IMPACT OF REWARD MODEL SIZE ON RL

We conducted a comparative analysis of reward models of different sizes in the iterative preference optimization training process. As shown in Tab. A3, larger reward models provide more precise reward estimations, leading to improved alignment with human preferences. This enhanced reward accuracy contributes to better optimization during preference alignment training, ultimately resulting in superior model performance. Our findings highlight the importance of scaling the critic model to achieve more reliable preference optimization and reinforce the benefits of leveraging larger critic architectures for improved alignment outcomes.

C ADDITIONAL QUALITATIVE COMPARISON

We provide a more detailed qualitative comparison in Fig. A3 and Fig. A4, which demonstrates the superiority of our model over the baseline. This visualization highlights key improvements in fidelity and overall quality. By comparing outputs side by side, we illustrate the advantages of our model in handling complex scenarios, reinforcing its effectiveness in real-world applications.

D USE OF LLM

Large Language Models (LLMs) were employed solely as auxiliary tools to improve the efficiency of manuscript preparation. Their use included assisting in retrieving related literature, editing and formatting $\LaTeX$ mathematical expressions, polishing the language style, debugging minor code or visualization scripts, and formatting references.

E SOCIETAL IMPACTS

The emergence of advanced text-to-video generation pipelines and optimized training strategies marks a major breakthrough in AI-driven content creation. These innovations significantly lower barriers to video production, making it more efficient, cost-effective, and widely accessible. Industries such as education, entertainment, marketing, and accessibility technology stand to benefit immensely, as creators can generate high-quality videos with minimal effort and resources.

However, alongside these advancements come ethical concerns, particularly regarding potential misuse. One of the most pressing risks is the creation of highly realistic yet misleading content, such as deepfakes, which could be exploited for misinformation, reputational harm, or social manipulation. Additionally, automated video curation may unintentionally reinforce biases present in training datasets, leading to the perpetuation of stereotypes or discriminatory narratives.

On a larger scale, the widespread adoption of AI-driven video generation could disrupt traditional media industries, impacting professions like video editing, animation, and scriptwriting. The proliferation of hyper-realistic AI-generated content also raises concerns about media authenticity, making it increasingly difficult for audiences to distinguish between real and synthetic footage. Furthermore, if access to these technologies remains limited to well-funded organizations, existing digital divides could widen, leaving smaller creators at a disadvantage.

To mitigate these risks, proactive measures must be taken. Implementing transparency features—such as metadata labels to distinguish AI-generated content—can enhance accountability. Ensuring diverse and ethically curated datasets is crucial to reducing bias and promoting fairness in outputs. Policymakers, industry leaders, and researchers must work together to develop regulatory frameworks that balance innovation with ethical responsibility. Additionally, providing educational resources and affordable access to AI tools can help democratize their benefits, preventing technological disparities.

The rapid evolution of generative AI necessitates ongoing ethical oversight. Continuous evaluation of its societal impact, coupled with adaptive safeguards, is essential to minimizing harm while maximizing positive outcomes. Addressing these challenges requires a multidisciplinary approach, fostering a responsible AI ecosystem that empowers creators without compromising ethical integrity.

F LIMITATIONS

While our approach represents a significant advancement in text-to-video generation, several challenges persist. Addressing these issues presents valuable opportunities for future research and development. Below, we outline key limitations and potential solutions:

Computational Constraints Despite improvements in training efficiency, large-scale text-to-video models still demand substantial computational power. This limitation may hinder accessibility, particularly for individuals and organizations with limited resources. Future research should focus on developing lightweight architectures, model compression techniques, and efficient inference strategies to enhance accessibility in resource-constrained environments.

Dataset Representativeness and Cultural Adaptability The effectiveness of our model is highly dependent on the quality and diversity of its training dataset. While significant efforts were made to curate a representative dataset, certain cultural contexts, niche topics, and underrepresented communities may not be adequately covered. Additionally, biases present in pretraining data could lead to unintended stereotypes or inappropriate outputs. Expanding dataset coverage, improving curation techniques, and integrating fairness-aware training strategies will be crucial to enhancing contextual accuracy and mitigating bias.

Multimodal Hallucination and Interpretability Since our evaluation models are fine-tuned multimodal large language models (MLLMs), they are susceptible to hallucination—generating content that appears plausible but is factually incorrect or cannot be inferred from the input text and images. Furthermore, MLLMs inherit biases from their pretraining data, which may lead to inappropriate responses. While careful curation of supervised fine-tuning (SFT) data helps alleviate these issues, they are not fully resolved. In addition to these concerns, MLLMs also suffer from opacity, interpretability challenges, and sensitivity to input formatting, making it difficult to understand and control their decision-making processes. Enhancing model explainability and robustness remains a crucial area for future research.

Human Annotation Limitations Human evaluation plays a critical role in assessing model performance, but it is inherently subjective and influenced by annotator biases, perspectives, and inconsistencies. Errors or noisy labels may arise during annotation, potentially affecting evaluation

reliability. To address this, we conducted multiple rounds of trial annotations and random sampling quality inspections to improve inter-annotator consistency. Additionally, we designed user-friendly annotation guidelines and interfaces to streamline the process and enhance accuracy. However, variations in annotation methodologies across different platforms and annotators can still lead to discrepancies, limiting the generalizability of our results. Moreover, human annotation remains time-consuming and resource-intensive, constraining scalability. Future efforts could explore semi-automated evaluation approaches or crowdsourced methods to improve efficiency while maintaining quality.

Ethical Considerations and Regulatory Compliance Ensuring responsible AI deployment remains an ongoing challenge, even with safeguards in place to prevent misuse. The rapid advancement of AI-generated media demands continuous revisions to ethical policies and regulatory frameworks. Proactively engaging with policymakers, industry leaders, and researchers is crucial for developing guidelines that uphold accountability, transparency, and fair usage. Embedding metadata or digital watermarks in AI-generated content can improve traceability and help differentiate synthetic media from authentic sources. Furthermore, raising public awareness and education on AI-generated content help users recognize potential risks. Developers should prioritize responsible AI innovation by ensuring algorithmic fairness while minimizing bias. Additionally, fostering cross-industry collaboration and promoting global standards will contribute to a more secure and trustworthy AI ecosystem.



A man with a beard is playing the piano in a concert hall. He wears a black tuxedo and a bow tie. His fingers dance on the keys as he plays a beautiful melody.



A director, dressed in a dark blazer and casual jeans, sits at a modern desk cluttered with scripts and notes. He types intently on a sleek laptop, his brow furrowed in concentration. The room is dimly lit by a single desk lamp, casting a warm glow over his focused expression. Behind him, a large window frames a city skyline at dusk, adding a sense of urgency and creativity to the scene.



A woman sits in a cozy, well-lit room, her hands deftly working a long, creamy white scarf. She wears a warm, olive-green sweater and a gentle smile, her eyes focused on the intricate pattern of her knitting. The soft, natural light streaming through a nearby window highlights the texture of the yarn and the calm, serene atmosphere of the space. Her concentration and skill are evident as she effortlessly manipulates the needles, creating a beautiful, handmade scarf.



An actor dressed in rugged, casual attire, including a dark jacket, jeans, and hiking boots, walks alongside a dingo in a natural, wooded setting. The dingo, with its distinctive reddish-brown fur and alert eyes, moves gracefully beside the actor, who looks calm and engaged. The background features lush greenery and dappled sunlight filtering through the trees, creating a serene and harmonious atmosphere. The actor and the dingo walk in sync, their bond evident in their synchronized steps and mutual awareness of each other.

Figure A3: Enhanced qualitative comparison. We conduct a qualitative comparison between CogVideo-2B and Dual-IPO-2B to evaluate their performance



Figure A4: **Enhanced qualitative comparison.** We conduct a qualitative comparison between CogVideo-2B and Dual-IPO-2B to evaluate their performance