DynamicsBoost: Dynamic Plausible Video Generation via Annotation-Free Continuation Preference Optimization

Jiaxing Li[1,2] Jiepeng Wang[3] Junyao Gao[4] Yang Liu[2]
Eric Li[2] Bo An[1] Hao-Xiang Guo[2†]

[1]Nanyang Technological University, [2]Skywork AI,
[3]The University of Hong Kong, [4]Tongji University

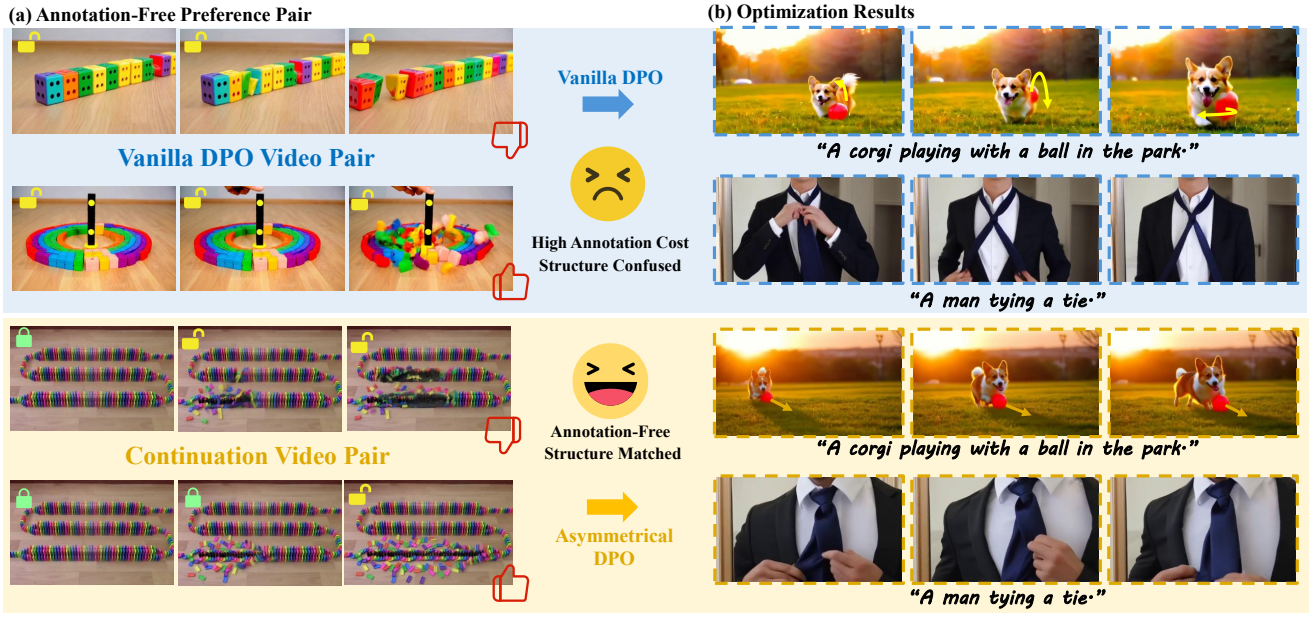

Figure 1. *DynamicsBoost* performs annotation-free preference alignment via video continuation, automatically producing structurally matched preference pairs and yielding video results with more faithful dynamics and stronger semantic consistency.

Abstract

Despite significant progress in text-to-video generation, current models still suffer from unrealistic dynamics, temporal inconsistency, and unstable semantic alignment. Existing preference alignment approaches rely on costly and often ambiguous human or VLM-based video preference annotation, which has become a major bottleneck for scaling data. To address this challenge, we propose an annotation-free preference alignment method that constructs accurate preference pairs through video continuation. We extend a pretrained video generation model into a continuation model and apply continuation with different amounts of reference frames while keeping the total video length fixed. As generated segments are inferior to ground-truth frames and and fixed-length continuations conditioned on more reference frames contain less generated content, they exhibit higher fidelity than those with fewer references, naturally inducing a preference order. We further introduce Asymmetrical DPO, which computes preference loss on all continuation regions except the shared prefix conditioning frames and normalizes it by their length, preventing spurious preference signals from leaking into the conditioned portion. Experiments across multiple benchmarks show that our method delivers significant improvements in dynamics realism, temporal coherence, and semantic alignment over existing DPO-based approaches, while fully eliminating the need for human preference labeling or auxiliary reward models.

1. Introduction

Recent advances in diffusion models have led to remarkable progress in text-to-video (T2V) generation, enabling

‡Corresponding authors.

models to synthesize videos with high-fidelity visual quality and basic spatiotemporal coherence [3, 4, 13, 30, 37]. Despite these achievements, current T2V models still exhibit notable limitations, including unrealistic dynamics, temporal inconsistency, and unstable semantic alignment. These drawbacks hinder the ability of modern video generative models to serve as true world-model–level systems capable of reliable temporal reasoning and prediction.

To address these limitations, recent research has begun exploring preference alignment for text-to-video generation. Existing approaches follow two paradigms inherited from language and image alignment: enhancing video quality through reinforcement learning methods that leverage reward models/functions. [4, 19, 35], or applying direct preference optimization (DPO) with positive–negative video pairs [33]. Despite their differences, both approaches fundamentally rely on labeled preference data: The former often requires collecting large-scale ranked video pairs for training reward models, while DPO avoids explicit RM training but still generates video pairs and requires human or Vision-Language-Model (VLM) preference judgment to decide which one is preferred. Unlike images, video preference annotation is costly and ambiguous: annotators must jointly evaluate visual fidelity, temporal consistency, motion dynamics, and semantic alignment, and preferences may vary across timestamps. Consequently, neither human annotators nor VLMs can generate accurate and consistent video preference labeling. The high cost and low efficiency of obtaining preference data have become one of the major bottlenecks in video preference alignment.

In our experiments, we observe that the data from *video continuation* task naturally exhibits a ordered structure under high structural consistency (Fig. 4). This inherent ordering provides an effective solution to this challenge and makes it possible to generate stable and fully automated preference data at scale.

Specifically, we first extend a pretrained video generation model into a latent-space continuation model capable of handling arbitrary continuation lengths. For a reference video of N frames, we extract the first N_1 frames and apply video continuation to synthesize the remaining $N - N_1$ frames, forming a new N-frame video. Similarly, we construct another continuation result using the first N_2 frames, where $N_2 > N_1$. Generated video segments are generally inferior to their ground-truth counterparts in dynamic videos. Since the first continuation contains a longer generated portion and less ground-truth conditioning than the second, we therefore infer first continuation is inferior to the second. We formally justify this assumption in Sec. 4.3. Then, a natural, structurally consistent preference ordering emerges without requiring any human judgment or external scoring, enabling scalable and annotation-free preference construction.

Building on these automatically constructed preference data, we introduce Asymmetrical DPO to ensure that preference signals act only on the generated regions rather than the shared conditioning frames. Unlike standard methods that apply preference loss over the entire video, we compute the DPO loss solely on the non-shared continuation regions where the two continuations differ. To avoid gradient instability caused by the varying length of these asymmetric regions across training steps, we further normalize the loss by the number of continuation frames. This asymmetrical design concentrates alignment signals precisely where preference differences arise, enabling effective preference learning without human annotations or reward models.

We validate our approach on multiple video generation benchmarks. Our method consistently achieves better motion realism, temporal coherence, and text alignment compared to prior DPO-based methods, without sacrificing visual quality. Ablation studies further show that continuation–based preference ordering strongly correlates with VLM-judged preference signals, confirming that continuation length is an effective supervision-free preference cue.

2. Related Work

2.1. Text-to-Video Generation

Text-to-video (T2V) generation aims to synthesize visually plausible and semantically aligned videos from textual descriptions. Early diffusion-based T2V models are built upon U-Net architectures [9, 11], where temporal modules such as 3D convolutions or temporal attention are inserted into 2D image-generation pipelines to capture motion dynamics [2, 5, 7]. Recent Diffusion Transformer (DiT) architectures [14, 24] replace U-Net convolutional backbones with fully Transformer-based designs, allowing stronger modeling of long-range spatial–temporal dependencies and better scalability. When scaled to massive video datasets [23, 31], DiT-based models such as Wan[30], Sora [3], SkyReels [4], CogVideoX [37], and HunyuanVideo [13] achieve substantial improvements in temporal coherence, motion fidelity, and scene consistency, pushing video generation towards world-model-style learning.

However, scaling alone does not guarantee controllability or alignment with user intent: large T2V models often fail to follow fine-grained semantic or aesthetic constraints, especially when user expectations diverge from the pretraining data distribution. To bridge this gap, recent work explores post-training strategies, including parameter-efficient fine-tuning (PEFT) [6, 10, 16], data-centric enhancement [8], and human preference alignment [20, 25, 33, 38], which rely on external supervision from human annotations or Vision-Language-Model (VLM) scoring. These approaches typically offer performance gains, yet require curated preference data that is expensive to obtain and inherently limited

in scale. In contrast, our method constructs preference pairs automatically via video continuation, enabling large-scale preference data expansion at minimal cost while delivering stronger performance improvement.

2.2. Preference Alignment for Video Generation

Recent work explores transferring the RLHF paradigm [26–28] to video generation [19, 20, 33, 35]. One line of approaches trains an explicit video reward model (RM), where models such as LiFT [32], VisionReward [34], and VideoAlign [19] collect large-scale video datasets and require expensive human preference annotations to fine-tune video-language models (VLMs) such that they can serve as video reward predictors. PISA [15] further designs proxy rewards using depth and optical flow for constrained physical scenarios. These RMs then provide preference feedback to guide alignment. Another line of work adopts Direct Preference Optimization (DPO), which avoids explicit RM training but still requires large amounts of pairwise preference data. In practice, constructing such win–loss pairs is costly and difficult: data must either be manually filtered [33] or labeled by pretrained VLMs acting as proxy annotators [19, 20, 32]. In contrast, we exploit the inherent ordered structure embedded in video continuation data to optimize pretrained video models. By leveraging the monotonic relationship between continuation length and video quality, we construct structurally matched preference pairs automatically and redesign the DPO objective into an Asymmetrical DPO formulation tailored to continuation properties. This yields stable, annotation-free preference supervision and enables high-quality post-training improvements.

3. Method

We aim to achieve scalable, annotation-free preference alignment for video generation. The key idea is to treat continuation as a natural source of preference: continuation videos with shorter generated segments are preferable to those with longer generated segments. We extend any pretrained text-to-video model into a latent continuation model to automatically construct preference pairs (Sec 3.2), and apply Asymmetrical DPO to align the model by optimizing only the continuation region (Sec 3.3). This yields efficient and targeted preference alignment without human annotations or reward models. The overall framework of DynamicBuff is illustrated in Figure 2.

3.1. Preliminaries

Flow Matching Models. Flow matching [17, 21] formulates generation as learning a velocity field that transports noise samples to data samples. Given a data sample $x_0 \sim X_0$ and a Gaussian noise sample $x_1 \sim X_1$, an intermediate state x_t is defined by linear interpolation. A neural network $v_\theta(x_t, t)$ predicts the velocity that moves x_t along the path toward the data sample. The training objective minimizes the difference between the predicted and target velocity:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\, x_0,\, x_1}\left[\|v_\theta(x_t, t) - (x_1 - x_0)\|_2^2\right]. \quad (1)$$

By integrating the learned velocity field, the model generates samples by transporting noise to data, enabling efficient and stable training for modern image and video generative models.

Diffusion-DPO and Flow-DPO. Direct Preference Optimization (DPO) can be applied to diffusion models by comparing the noise–prediction errors of a winning sample x^w and a losing sample x^l. Diffusion-DPO [29] optimizes the model such that the preferred sample yields a smaller prediction error. The loss is defined as:

$$\mathcal{L}_{\text{Diffusion-DPO}} = -\log\sigma\left(-\frac{\beta}{2}\,\Delta\mathcal{E}_\varepsilon\right), \quad (2)$$

where $\sigma(\cdot)$ is the sigmoid function, β is the DPO temperature that controls alignment strength, and $\Delta\mathcal{E}_\varepsilon$ denotes the difference in noise-prediction errors between winning and losing samples.

In rectified-flow models, the network predicts velocity instead of noise. Using the deterministic equivalence $\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2 = (1-t)^2\|v - v_\theta(x_t, t)\|^2$, Diffusion-DPO can be rewritten in velocity space, yielding Flow-DPO [18]:

$$\mathcal{L}_{\text{Flow-DPO}} = -\log\sigma\left(-\frac{\beta(1-t)^2}{2}\,\Delta\mathcal{E}_v\right), \quad (3)$$

where $\Delta\mathcal{E}_v$ is the velocity-prediction error difference. Thus, Flow-DPO is a mathematically equivalent formulation of Diffusion-DPO under the rectified-flow parameterization.

3.2. Continuation-Based Preference Pair

We construct preference data pairs through video continuation. For brevity, the following definitions operate in the video latent space by default. Given a ground-truth reference video $z^{ref} \in \mathbb{R}^{N\times H\times W\times C}$ with its corresponding text prompt, the video continuation task is formulated as a multi-frame conditioned video generation task, where the first N_1 frames $z^{cond} \in \mathbb{R}^{N_1\times H\times W\times C}$ serve as conditioning frames, and the remaining $N - N_1$ frames $z^{gen} \in \mathbb{R}^{(N-N_1)\times H\times W\times C}$ serve as target generated frames. By concatenating z^{cond} and z^{gen} along the frame dimension, we obtain a new video of N frames, which we call the N_1-*frame continuation* of z^{ref}. For a continuation result of a dynamic video, the generated portion is always inferior to its reference counterpart. Therefore, for another N_2-*frame continuation* of z^{ref} where $N_2 > N_1$, since its generated portion is shorter than that of the N_1-*frame continuation*, we can infer that its overall quality is better than the N_1-*frame continuation*. In this way, we can construct accurate preference data pairs without human annotation.

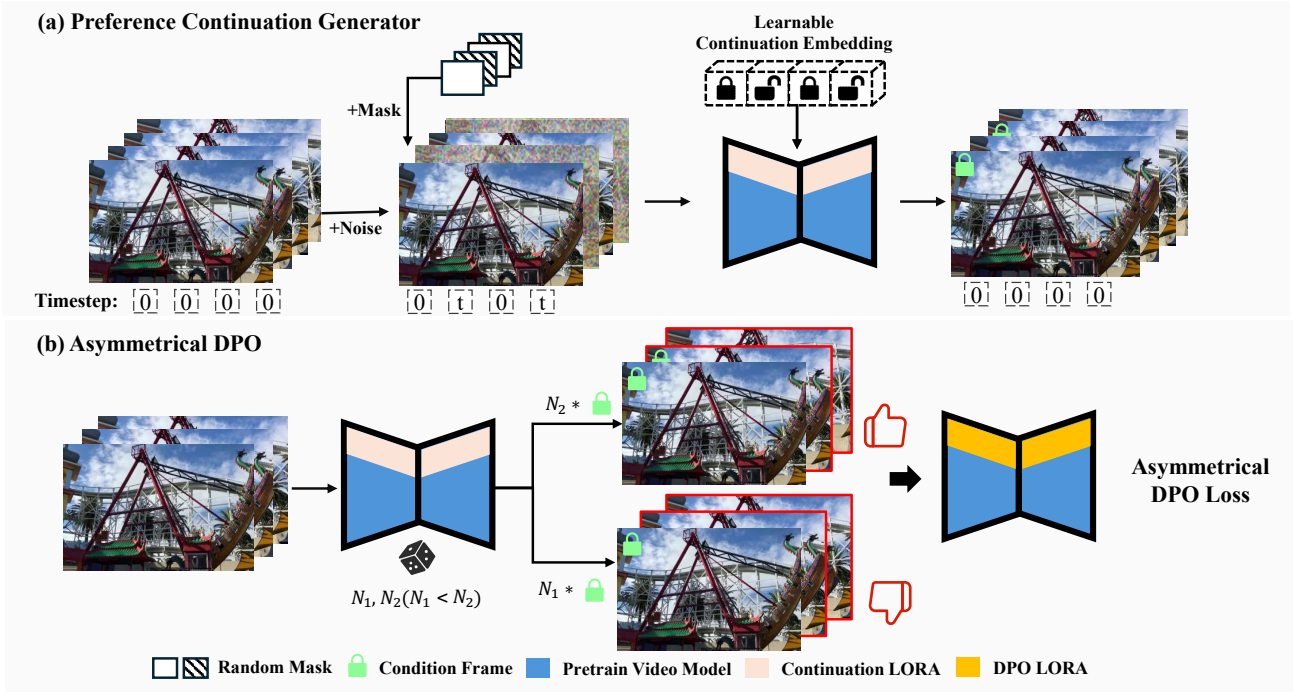

Figure 2. **Overview of the DynamicsBoost pipeline.** We extend a pretrained video generator into a continuation-based preference model (a), then sample continuation pair of different lengths and optimize them with Asymmetrical DPO (b). This pipeline automatically constructs high-quality, scalable and structurally aligned ordered data and yields video generations with improved fidelity and dynamic plausibility.

We extend the pretrained text-to-video model to support arbitrary-frame continuation by enabling per-frame conditioning through masked timesteps and frame-level task prompts. Given a video latent sequence $z_0 \in \mathbb{R}^{N\times H\times W\times C}$, we sample noise $z_1 \sim \mathcal{N}(0, I)$ and a timestep $t \sim \mathcal{U}(0, 1)$. We draw a binary mask $M \in \{0, 1\}^N$ over the frame dimension, where $M_i = 1$ denotes a conditioned frame and $M_i = 0$ denotes a target frame.

Timestep masking is first applied, where conditioned frames receive a zero timestep and target frames retain the sampled noise level,

$$t' = t \cdot (1 - M), \tag{2}$$

Based on the masked timestep, non-uniform noising is then applied to construct the model input,

$$z'_t = (1 - t')z_0 + t'z_1. \tag{1}$$

Next, we inject learnable frame-level task prompts, which encode whether each frame is conditioned or to be generated. Let P_{cond} and P_{noisy} denote trainable embeddings for conditioned and target frames, respectively. The frame-aligned prompt sequence is:

$$P_{\text{task}} = M \odot P_{\text{cond}} + (1 - M) \odot P_{\text{noisy}}, \tag{3}$$

The task-prompt embeddings are then concatenated with the context embeddings to form the final input sequence.

The flow model v_Φ takes $(z'_t, t', P_{\text{task}})$ and predicts the rectified-flow velocity $v_\Phi(z'_t, t', P_{\text{task}})$. Supervision is computed only on target frames to avoid degrading conditioned content:

$$\mathcal{L} = \mathbb{E}\Big[\|(1-M)\odot\big((z_1-z_0)-v_\Phi(z'_t,\, t',\, P_{\text{task}})\big)\|^2\Big]. \tag{4}$$

We optimize only the LoRA adapters and the task prompt embeddings while freezing the pretrained backbone:

$$\{\theta^*_{\text{LoRA}},\, P^*_{\text{cond}},\, P^*_{\text{noisy}}\} = \underset{\theta_{\text{LoRA}},\, P_{\text{cond}},\, P_{\text{noisy}}}{\operatorname{argmin}} \mathcal{L}. \tag{5}$$

After training, the model supports arbitrary-length continuation from the reference video. Sampling continuation results with different lengths naturally yields a preference pair, turning the continuation model into a *preference data generator* and enabling scalable, annotation-free supervision.

3.3. Asymmetrical DPO

To leverage the inherent preference signal introduced by latent continuation, we propose Asymmetrical DPO for preference alignment on continuation outputs. Instead of comparing two entire generated videos as in standard DPO, we exploit the fact that our continuation model can produce videos with different continuation lengths. For a given

conditioning video, we randomly sample two continuation lengths $N_1 < N_2$ and generate an N_1-*frame continuation* losing data sample z_0^l and an N_2-*frame continuation* winning data sample z_0^w.

The two sequences can be partitioned into three contrastive regions: (1) *Shared prefix* (0 to N_1), where both samples contain identical ground-truth frames; (2) *Asymmetric region* (N_1 to N_2), where the winning sample uses ground-truth frames while the losing sample uses model-generated ones; and (3) *Fully generated region* (N_2 to N), where both samples consist solely of model outputs.

Unlike standard Diffusion/Flow-DPO, which accumulates preference loss over the entire video, we constrain preference supervision to the continuation region only (i.e., *Regions (2)–(3)*). The resulting asymmetrical Flow-DPO objective is defined as:

$$\mathcal{L}_{\text{AsymDPO}} = -\frac{1}{N - \min(N_1, N_2)} \log \sigma\big(-\beta \cdot \Delta\mathcal{E}\big), \quad (4)$$

where $\sigma(\cdot)$ is the sigmoid, β is the DPO temperature, and $\Delta\mathcal{E}$ accumulates error differences only over continuation frames:

$$\Delta\mathcal{E} = \sum_{i=\min(N_1,N_2)}^{N} \big(\|v_i^w - v_\theta(x_{t,i}^w, t)\|^2 - \|v_i^l - v_\theta(x_{t,i}^l, t)\|^2\big). \quad (5)$$

Here, v^w, v^l denote target velocity fields (or noise) of winning and losing continuations, while v_θ is the predicted velocity. The normalization factor in Eq. 4 ensures that different continuation lengths remain comparable and prevents gradient imbalance during training.

This asymmetrical formulation focuses preference learning strictly on the frames where the model generates new content, avoiding gradient dilution caused by shared conditioning frames. Moreover, the preference ordering emerges automatically from continuation length, requiring neither human annotations nor reward models. Asymmetrical DPO is fully compatible with Flow-DPO and introduces no additional training cost or architectural modifications.

4. Experiments

4.1. Experiment Setup

Implementation Details. We build our pipeline on a DiT-based latent flow text-to-video model[30]. We curate 100K dynamic, high-quality videos from OpenVid [23]: 80K are used for continuation training and 20K serve as conditioning inputs for alignment. Videos are standardized to 49 frames at a resolution of 288×512. For continuation training, we use LoRA with rank 196 and $\alpha = 196$, a learning rate of 8×10^{-5}, a batch size of 16, and train for 20K steps. For alignment, we first apply a LoRA-based SFT cold start[1], using the 20K conditioning videos as supervision targets. Training is performed with a learning rate of 1×10^{-5}, batch size 32, and LoRA rank 128 for 0.6 epoch. We then perform Asymmetric DPO, continuing from the same LoRA weights, with a learning rate of 5×10^{-6}, batch size 32, LoRA rank 128, and $\beta = 800$ for 1 epoch. Positive samples condition on the first 80–100% of video frames, while negative samples condition on the the latent prefix from the first frame up to 60%. Prefix lengths are randomly sampled from these ranges at each training step. All experiments are conducted on eight NVIDIA A800 GPUs.

Evaluation Settings. We evaluate our model using evalution datasets from three major benchmarks for a comprehensive assessment. (1) **VBench** [12] and **VideoGen-Eval** [36] are general-purpose benchmarks that measure overall video generation quality, including fidelity and semantic correctness. We adopt the subsets released by VideoAlign [19], containing 946 and 400 prompts respectively, to assess general video generation capability. (2) **PhysGenBench** [22] is designed to evaluate motion rationality and physical coherence in generated videos. We use all 160 prompts in this benchmark to assess whether our continuation-pair strategy enhances dynamic consistency and motion quality. For evaluation metrics, we follow the VBench evaluation protocol, which jointly assesses video quality, motion quality, and semantic consistency.

4.2. Comparison

We set the pretrained model, SFT, Flow-DPO [19], Flow-StructuralDPO [33], and Flow-DenseDPO [33] as our baselines. Starting from the pretrained model, we perform supervised fine-tuning on 20k conditional video generation samples for 1.5 epochs to obtain the SFT model. Since Flow-StructuralDPO and Flow-DenseDPO do not provide official implementations, we re-implement them strictly based on the procedures and pseudocode described in the original papers. For Flow-StructuralDPO, we apply Guided Paired Video Generation during the sampling stage of Flow-DPO. For Flow-DenseDPO, each generated video is evenly divided into five short segments, and each segment is independently scored by a VLM to obtain dense preference signals. To ensure fair comparison, all DPO-based methods are initialized with the same SFT cold start used in our pipeline, and all of them employ VideoReward [19] as the reward model.

Table 1 and Table 2 present the quantitative results on VBench, VideoGen-Eval, and PhysGenBench. We observe that applying SFT on the pretrained model with a small-scale dataset leads to slight performance degradation on certain metrics, while DPO-based optimization quickly restores and further improves these metrics. Our method achieves notable gains in video dynamics, visual quality, and semantic alignment, obtaining the best performance

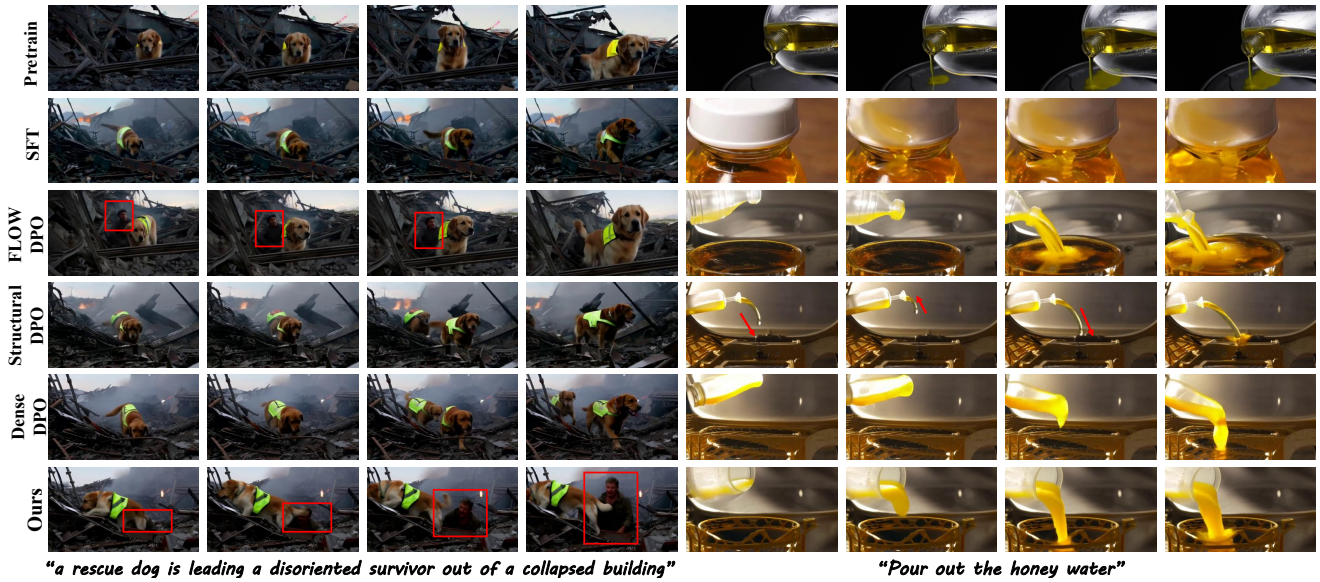

Figure 3. **Qualitative comparison with baselines.** Our method produces results that are more semantically consistent and dynamically coherent compared to other approaches.

Table 1. Quantitative comparison with baselines on VBench. Best in **bold**; second best underlined.

Method	Aesthetic Quality ↑	Imaging Quality ↑	Background Consistency ↑	Motion Smoothness ↑	Dynamic Degree ↑	Overall Consistency ↑
Pretrain	55.51	65.12	96.71	98.81	34.72	24.48
SFT	55.80	64.52	96.85	96.84	34.15	24.02
Flow-DPO	59.14	63.31	**98.15**	97.36	31.22	25.22
Flow-StructuralDPO	58.08	65.06	97.02	96.15	35.25	24.18
Flow-DenseDPO	58.95	**67.91**	97.11	98.16	40.10	25.02
Ours	**59.92**	66.81	97.53	**99.21**	**44.92**	**25.64**

across most metrics. Figure 3 further supports these findings. In the left example, only our method and Flow-DPO successfully generate characters consistent with the text prompt. In the right example, other baselines exhibit physically implausible motion during the pouring process: Structural DPO shows reversed water flow, while Dense DPO produces water leaking from the cup bottom. In contrast, our approach maintains both dynamic and physical consistency throughout the sequence.

4.3. Analysis and Ablation

Are Continuation Pairs Effective Preference Data? We use our continuation model and perform conditional continuation on our original DPO training data, generating 49-frame videos with 13 latent frames. We vary the number of provided latent reference frames N_{cond} in $\{0, 1, 2, 4, 8, 13\}$. The setting $N_{\text{cond}} = 0$ corresponds to text-to-video generation, while $N_{\text{cond}} = 13$ uses the full ground-truth reference. We randomly sample 100 videos from the dataset and evaluate them on VBench and VideoReward.

The upper part of Table 3 presents our results. Except for the pure T2V setting, all continuation configurations exhibit a consistent trend—providing more reference frames leads to higher VBench and VideoReward scores. This demonstrates that partial video context effectively enhances visual fidelity, motion stability, and semantic coherence. Such a monotonic improvement indicates that continuation pairs inherently satisfy the preference assumption, where continuations with more reference frames are consistently favored over those with fewer ones, validating them as reliable preference data for DPO supervision. Figure 4 further visualizes this behavior: except for the case without reference frames (i.e., pure T2V), the generated videos largely preserve structural and semantic consistency with the ground truth. When the number of reference frames is small, the videos exhibit lower visual quality and dynamic rationality. As the number of continuation frames increases, these aspects gradually improve, resulting in more stable and visually consistent outputs.

Effect of Reference Frame Length on DPO Training. To analyze how reference-frame length affects DPO training, we vary the continuation lengths for the negative sample (N_1) and the positive sample (N_2), while keeping the total latent length fixed at N. We evaluate five configurations

Table 2. Quantitative comparison with baselines on PhysGenBench and VideoGen-Eval. Best in **bold**; second best underlined.

Model	PhysGenBench						VideoGen-Eval					
	Aesth. Qual.	Imag. Qual.	Backgr. Cons.	Mot. Smooth.	Dyn. Deg.	Overall Cons.	Aesth. Qual.	Imag. Qual.	Backgr. Cons.	Mot. Smooth.	Dyn. Deg.	Overall Cons.
Pretrain	42.86	49.58	94.78	98.49	39.37	22.05	52.94	61.78	95.34	98.54	52.25	22.62
SFT	43.62	52.27	95.06	98.96	35.62	22.39	55.28	63.67	94.60	98.96	47.25	23.68
Flow-DPO	46.21	52.62	**97.11**	98.91	43.45	23.03	56.20	**66.24**	**97.23**	98.46	49.45	24.12
Flow-StructuralDPO	45.38	53.98	95.88	98.40	43.37	22.87	57.28	63.91	95.60	98.62	46.50	23.91
Flow-DenseDPO	46.44	52.95	95.23	99.20	**45.31**	22.58	56.12	62.39	95.95	98.93	43.75	23.20
Ours	**48.21**	**55.14**	96.85	**99.44**	44.32	**23.68**	**59.41**	64.23	96.49	**99.38**	**56.31**	**25.12**

Table 3. Ablation study on the Continuation Process. Best results are in **bold**, second best are underlined.

Ablation Setting	VBench Metrics (↑)						VideoReward Metrics (↑)		
	Aesthetic Quality	Imaging Quality	Background Consistency	Motion Smoothness	Dynamic Degree	Overall Consistency	VQ	MQ	TA
Conditioning/Total Frame Length									
0/13 (T2V)	44.83	**66.94**	94.24	98.88	43.97	21.93	-0.868	-0.571	**-0.323**
1/13	36.10	40.41	90.67	98.64	38.96	20.94	-1.220	-0.872	-0.376
2/13	37.04	42.52	91.34	98.71	53.01	21.84	-1.187	-0.925	-0.393
4/13	43.81	57.98	92.36	98.69	48.98	21.97	-0.902	-0.787	-0.353
8/13	45.34	60.51	93.72	99.04	39.96	**22.19**	-0.871	-0.599	-0.357
13/13 (GT)	**46.64**	63.09	**94.42**	**99.94**	**57.03**	21.91	**-0.859**	**-0.495**	-0.343
Continuation Model Design (using 4/13 conditioning frames)									
Full setting (4/13)	**43.81**	**57.98**	**92.36**	98.69	48.98	**21.97**	**-0.902**	**-0.787**	**-0.353**
w/o continuation prompt (4/13)	40.21	51.23	91.23	98.26	**51.36**	21.43	-1.002	-0.829	-0.364
w/o timestep mask (4/13)	39.86	49.29	92.11	**98.75**	44.17	21.28	-1.015	-0.901	-0.397

covering both randomized and fixed continuation strategies: (1) $N_1 \in [0, 0.6N]$, $N_2 \in [0.8N, N]$; (2) $N_1 \in [1, 0.6N]$, $N_2 \in [0.8N, N]$; (3) $N_1 = 0$, $N_2 = N$; (4) $N_1 = 1$, $N_2 = N$; (5) $N_1 \in [1, 0.6N]$, $N_2 = N$. These configurations allow us to isolate how the continuation-length gap between negative and positive samples influences preference learning. Results are presented in the upper part of Table 4.

Across the five continuation strategies, we observe two key trends. First, comparing Setting 1 vs. 2 and Setting 3 vs. 4 shows that using pure T2V samples ($n_2 = 0$) as negative examples brings only minor gains. This is consistent with our findings in Table 3: T2V outputs differ significantly in structure from continuation samples, producing weak preference signals and thus insufficient gradient guidance for DPO. Second, comparing Setting 2, Setting 4, and Setting 5 reveals that fixing the continuation length—especially for the negative sample—harms model generalization. In contrast, our bidirectional random continuation strategy (Setting 2), which samples both N_1 and N_2 from ranges, consistently achieves the best performance. These results indicate that diverse continuation-length gaps yield more informative preference supervision and are crucial for stable DPO optimization.

Ablation of Asymmetrical DPO. In this section, we perform an ablation on the effective comparison range used in our asymmetrical DPO loss. For a reference video of N frames and a corresponding data pair with a N_1-*frame continuation* negative sample and a N_2-*frame continuation* positive sample, we follow the partitioning in Sec. 3.3 and divide the sequences into three regions: (1) *Shared prefix* (0 to N_1); (2) *Asymmetric region* (N_1 to N_2); and (3) *Fully generated region* (N_2 to N). We compare three loss configurations: computing DPO only on *Region (3)*, on *Regions (2)–(3)*, and on all *Regions (1)–(3)*, the last of which reduces to the standard DPO formulation. The quantitative results are reported in Table 4.

The results show that computing the DPO loss on *Regions (2)–(3)* yields the best performance across all VBench metrics, followed by applying the loss only on *Region (3)*. In contrast, including *Region (1)* consistently leads to degraded results. This suggests that enforcing preference learning on segments where the two videos are identical may introduce noisy or uninformative gradients, while restricting optimization to segments where the positive and negative samples diverge provides cleaner and more effective supervision.

Ablation on Continuation Model Design. We conduct an ablation study on the continuation model design, comparing the full configuration with variants that remove the learnable continuation task prompt or omit timestep masking. With a total latent length of 13 frames, all experiments

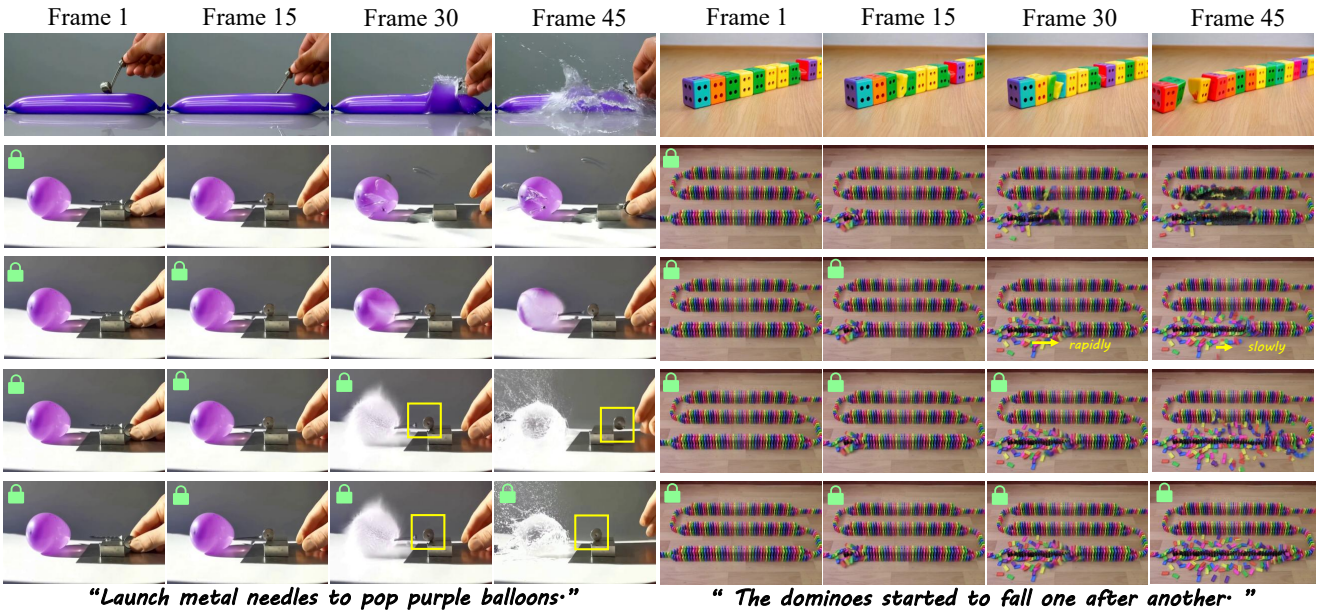

Figure 4. **Continuation results under different numbers of reference frames (indicated at the top-left of each example).** Non-zero reference settings produce continuations that better preserve structural consistency with the original video, and quality improves monotonically as the number of reference frames increases—approaching the ground truth when all frames are used. In contrast, the pure T2V setting exhibits significant structural deviation, breaking this monotonic relationship.

Table 4. Ablation study on the Asymmetrical DPO Process. Best results are shown in **bold**, and the second best are underlined.

Ablation Setting	Aesthetic Quality	Imaging Quality	Background Consistency	Motion Smoothness	Dynamic Degree	Overall Consistency
Continuation Sampling Strategy						
$N_1 \in [0,\ 0.6N],\ N_2 \in [0.8N,\ N]$	59.30	66.61	97.24	98.89	43.92	25.31
$N_1 \in [1,\ 0.6N],\ N_2 \in [0.8N,\ N]$	**59.92**	66.81	**97.53**	**99.21**	44.92	**25.64**
$N_1 = 0,\ N_2 = N$	57.09	65.08	96.76	96.11	34.32	24.42
$N_1 = 1,\ N_2 = N$	58.71	66.59	96.47	97.48	35.72	24.94
$N_1 \in [1,\ 0.6N],\ N_2 = N$	59.41	**66.94**	96.90	98.32	**46.41**	25.44
Loss Regions Used						
Regions (1)–(3)	58.10	66.42	96.63	97.21	42.48	22.15
Regions (2)–(3)	**59.92**	**66.81**	97.53	**99.21**	**44.92**	**25.64**
Region (3)	59.31	66.23	**97.71**	98.63	38.21	24.10

fix the continuation length to 4 latent frames. The lower part of Table 3 presents our results. Removing either component consistently degrades performance, particularly in motion smoothness and overall consistency, indicating that both the learnable prompt and timestep masking are essential for effective continuation modeling.

5. Conclusion

In this work, we address the critical bottleneck of preference data acquisition in video preference alignment by introducing an annotation-free method that automatically constructs preference pairs through video continuation. Our key insight is that video continuation naturally induces a preference ordering: for a fixed continuation length, continuations conditioned on more reference frames contain less generated content and therefore exhibit higher fidelity than those conditioned on fewer references. This observation enables scalable preference construction without human annotation or reward models. We further propose Asymmetrical DPO, which concentrates alignment signals on the continuation region by applying preference loss only where continuations differ. Extensive experiments demonstrate that our method achieves superior motion realism, temporal coherence, and text alignment compared to prior DPO-based approaches, validating the effectiveness of continuation-based preference ordering as a supervision-free signal for video alignment. Additional experimental results and discussion will be provided in the supplementary material.

References

[1] Trl: Transformer reinforcement learning. https://github.com/huggingface/trl. HuggingFace, 2024. 5

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 2

[4] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 2

[5] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 2

[6] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European conference on computer vision*, pages 39–55. Springer, 2024. 2

[7] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[8] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024. 2

[9] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 2

[10] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 5

[13] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[14] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2

[15] Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025. 3

[16] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *Advances in neural information processing systems*, 37:75692–75726, 2024. 2

[17] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[18] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 3

[19] 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. *arXiv preprint arXiv:2501.13918*, 2025. 2, 3, 5

[20] Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8009–8019, 2025. 2, 3

[21] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[22] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 5

[23] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 2, 5

[24] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[25] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion align-

ment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024. 2

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 3

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3

[29] B Wallace, M Dang, R Rafailov, L Zhou, A Lou, S Purushwalkam, S Ermon, C Xiong, SR Joty, and N Naik. Diffusion model alignment using direct preference optimization. in 2024 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8238, 2023. 3

[30] 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. 2, 5

[31] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025. 2

[32] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024. 3

[33] Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025. 2, 3, 5

[34] 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. *arXiv preprint arXiv:2412.21059*, 2024. 3

[35] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 2, 3

[36] Yuhang Yang, Ke Fan, Shangkun Sun, Hongxiang Li, Ailing Zeng, FeiLin Han, Wei Zhai, Wei Liu, Yang Cao, and Zheng-Jun Zha. Videogen-eval: Agent-based system for video generation evaluation. *arXiv preprint arXiv:2503.23452*, 2025. 5

[37] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[38] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6463–6474, 2024. 2