Reg-DPO: SFT-Regularized Direct Preference Optimization with GT-Pair for Improving Video Generation

Jie Du[1]   **Xinyu Gong**[1]   **Qingshan Tan**[2]   **Wen Li**[1]   **Yangming Cheng**[1]   **Weitao Wang**[1]

Chenlu Zhan[1]   **Suhui Wu**[1*]   **Hao Zhang**[1*]   **Jun Zhang**[1*]

[1]ByteDance   [2]Shanghai Jiao Tong University

Abstract

Recent studies have identified Direct Preference Optimization (DPO) as an efficient and reward-free approach to improving video generation quality. However, existing methods largely follow image-domain paradigms and are mainly developed on small-scale models ($\sim$2B parameters), limiting their ability to address the unique challenges of video tasks—such as costly data construction, unstable training, and heavy memory consumption. To overcome these limitations, we introduce a GT-Pair that automatically builds high-quality preference pairs by using real videos as positives and model-generated videos as negatives, eliminating the need for any external annotation. We further present Reg-DPO, which incorporates the SFT loss as a regularization term into the DPO objective to enhance training stability and generation fidelity. Additionally, by combining the FSDP framework with multiple memory optimization techniques, our approach achieves nearly $3\times$ higher training capacity than using FSDP alone. Extensive experiments on both I2V and T2V tasks across multiple datasets demonstrate that our method consistently outperforms existing approaches, delivering superior video generation quality.

1 Introduction

In the field of video generation, Direct Preference Optimization (DPO) [1] has become a leading approach for enhancing generation quality due to its simplicity and efficiency. It directly optimizes generative models on paired preference samples (y_w, y_l) without requiring explicit reward models or complex policy gradients, thereby achieving an effective balance between training efficiency and computational cost. Recent studies have shown that DPO and its variants deliver strong results in text and image generation [2] and are increasingly being applied to video tasks [3–7]. However, most existing works directly transfer image-based paradigms and are typically built on small-scale models ($\sim$2B parameters). Compared with images, video generation presents greater challenges in spatiotemporal modeling and data complexity, making naive adaptation ineffective. Critical issues remain unresolved, including the high cost of data construction, instability during training, and substantial GPU memory consumption. To overcome these challenges, we propose a systematic framework that jointly addresses three key aspects—data construction, algorithmic stability, and memory optimization—as detailed below.

For data construction, DPO training typically relies on model-generated samples with external annotations to form preference pairs. However, in the video domain, this process faces two major obstacles: the lack of reliable automatic evaluators makes human scoring costly, and subtle perceptual differences between generated samples result in few high-quality pairs. To address these challenges, we introduce the GT-Pair (Section 3.1), which directly uses ground-truth videos (GT) as positive sam-

---

ples and model outputs as negatives, standardizing their formats to meet training requirements. This approach automatically produces high-quality preference pairs without external labeling, achieving greater efficiency, lower cost, and stronger discriminability.

For algorithmic stability, the complex spatiotemporal dependencies in videos, combined with inherent limitations of DPO, often result in rapid convergence and pronounced distribution shift during training. To address this, we propose Reg-DPO (Section 3.2), which incorporates an SFT-based regularization term into the DPO loss. A dynamic weighting scheme balances preference learning with distribution consistency, enhancing both training stability and model performance. Experiments demonstrate that SFT regularization concurrently supervises the output distributions of positive and negative samples, enabling the model to learn preference signals while constraining distribution shift within a controlled range. Model performance consistently improves with increasing SFT weight until saturation. Across Image-to-Video (I2V) and Text-to-Video (T2V) tasks, Reg-DPO achieves state-of-the-art results according to both GSB human evaluations and VBench automatic metrics, providing a simple yet effective enhancement over standard DPO.

For memory optimization, current video generation models often exceed 10B parameters [8–10], and the multi-frame nature of video inputs, together with DPO training—which requires a frozen reference model and paired data—leads to extremely high GPU memory usage. To mitigate this, we combine the FSDP framework [11] with Flash Attention [12], Context Parallelism [13], and a Pair Parallelism strategy, achieving significant memory savings. Compared with FSDP alone, our approach triples effective training capacity and scales efficiently with GPU count. Implementation details are provided in Appendix A.2.

In summary, the main contributions of this work are as follows:

- GT-Pair: An efficient data construction method for video DPO training that generates high-quality preference data without external annotations.

- Reg-DPO: Incorporates SFT regularization into DPO loss to stabilize training, address inherent DPO limitations, and improve video generation quality.

- Memory Optimization Scheme: Combines FSDP with multiple memory-saving techniques to triple effective training capacity.

- Systematic Experiments: Empirically validate SFT regularization and assess GT-Pair and Reg-DPO on I2V and T2V tasks across multiple datasets.

- Unified Training Framework: A unified and memory-optimized framework that supports I2V/T2V tasks, SFT/DPO modes, LoRA and full-parameter fine-tuning, as well as multi-resolution training. The data and code will be released later.

2 Related Work

2.1 Preference Optimization for Video Generation

In video generation, aligning model outputs with human preferences is a key research focus. Early studies adopt the RLHF framework from language and image models [14–26], using preference data to train a reward model and policy gradients for alignment. Initial methods apply image-based reward models [18, 27–30] frame by frame [31–34], ignoring temporal dependencies and cross-frame consistency. Later works develop specialized video reward models, integrating multiple video quality metrics [35, 36] into a unified score [3, 5], but these work mainly for short videos and underperform on long ones [10, 37, 38]. More recent approaches, like VideoAlign [39], VisionReward [40], and LiFT [41], train vision-language models on large human-annotated datasets for more reliable preference prediction. Yet they require massive annotations and costly training, making development expensive and slow, while RLHF's reliance on reward models adds complexity, instability, and risk of reward hacking [3, 4].

In contrast, DPO [1] streamlines preference alignment by learning directly from paired samples, avoiding separate reward models and complex policy optimization. DPO and its extensions excel in text and image generation [2, 42–56] and are being explored for video [3–7]. However, these remain preliminary, mainly transferring image-based DPO to small models ($\approx$2B parameters). Video generation involves complex spatiotemporal dependencies and higher memory costs, and as models
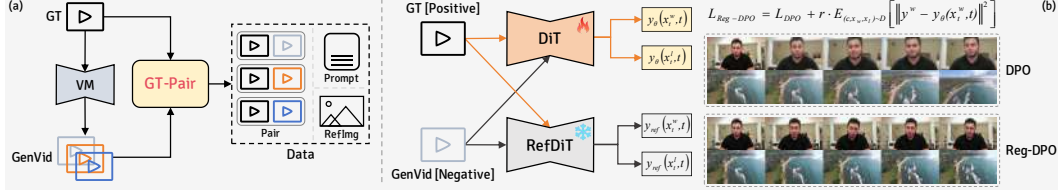
Figure 1: Overall framework. GT denotes the real video (Ground Truth), GenVid the generated video, RefImg the reference image, VM the video model, and RefDiT the reference DiT. Fig. 1a illustrates the GT-Pair strategy, where real videos serve as positive samples and generated videos as negatives, enabling high-quality preference pairs without any external annotations. Fig. 1b presents the overall pipeline of video DPO training. Owing to the instability of DPO (Section 3.2.2), training often collapses, producing grid-like blurring artifacts. In contrast, Reg-DPO incorporates an SFT regularization term that stabilizes training and yields significantly clearer videos.

scale beyond 10B parameters [8–10], existing DPO pipelines face challenges in data, algorithms, and memory. Efficient DPO strategies tailored to video tasks are therefore urgently needed.

2.2 Integration of DPO and SFT

Recent studies [57–64] explore combining DPO with SFT to enhance model performance, primarily in NLP. However, differences in probabilistic modeling and objectives hinder direct transfer of these findings to visual generation frameworks such as Diffusion and Flow Matching. In visual generation, some works adopt similar strategies: SSPO [63] uses stochastic checkpoint replay and self-sampling to adaptively switch between DPO and SFT, while ReDPO [64] linearly combines the two losses following NLP approaches [62]. In contrast, our work systematically analyzes DPO training instability from theoretical and visual perspectives, introducing SFT-based regularization. We further study its impact on training dynamics and video quality and propose a novel dynamic weighting scheme, establishing a new paradigm for DPO in visual generation.

3 Methods

3.1 GT-Pair

In DPO training, the quality of preference pairs is a key determinant of model performance. Existing pipelines typically rely on model-generated samples annotated manually or automatically to construct these pairs. However, this process faces two major challenges in the video domain. First, accurate video annotation models are costly and difficult to train, leading most optimization tasks to still depend on human annotators. Reliable judgments often require multiple experts to evaluate each generated video across several dimensions. The long temporal span and complex semantics of videos make annotation highly time-consuming and labor-intensive; in our experiments, professional annotators spent approximately 20 s per 5-second video, requiring about 10 days to label 9,788 pairs. Second, subtle perceptual differences between generated samples yield few high-quality preference pairs—those with clear positive–negative contrast and strong positive quality—accounting for only 10.9% of our data. Consequently, large-scale video generation is needed to obtain sufficient training samples, further escalating construction costs.

Motivated by the observation that real videos outperform generated ones in visual quality, temporal consistency, and semantic completeness, we propose GT-Pair (Fig. 1a), which constructs high-quality preference pairs by treating real videos (GT) as positives and their GT-conditioned generated counterparts as negatives. Considering the spatiotemporal complexity of video tasks, GT-Pair involves five stages (see Appendix A.1 for details): acquisition $\rightarrow$ filtering $\rightarrow$ normalization $\rightarrow$ video generation $\rightarrow$ pair construction. High-quality real videos are first selected as GT samples; low-quality content (static scenes, frequent cuts, or excessive effects) is filtered out. Videos are then normalized in resolution, aspect ratio, frame rate, and frame count, with reference frames and textual prompts generated. Corresponding videos are generated using these references, and each GT video is paired with its generated counterpart for DPO training.

Compared with methods that rely solely on generated videos to create paired data via manual annotation, GT-Pair offers the following advantages: (1) High data quality. GT videos inherently surpass generated videos, producing large gaps between positive and negative samples and avoiding annotation bias. As shown in Fig. 9, GT-Pairs surpass manually annotated ones in positive sample quality and discriminability, exhibiting superior visual consistency, semantic completeness, and motion smoothness. (2) Low cost and high scalability. GT-Pair requires no manual or automated annotation, using only real and generated videos, enabling efficient and scalable construction of large-scale, high-quality preference pairs.

3.2 Reg-DPO

3.2.1 Preliminaries

DPO [1] is a preference optimization algorithm derived from RLHF [14, 65, 66], which can be jointly formulated from the objective functions of PPO [21] and the reward model (see Appendix A.3 for detailed derivation). Its objective is

$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log\sigma(\beta\cdot s(\theta))],$$

$$s(\theta) = \underbrace{[\log\pi_\theta(y_w\mid x) - \log\pi_\theta(y_l\mid x)]}_{\Delta_\theta(y_w,y_l)} - \underbrace{[\log\pi_{\mathrm{ref}}(y_w\mid x) - \log\pi_{\mathrm{ref}}(y_l\mid x)]}_{\Delta_{\mathrm{ref}}(y_w,y_l)} \qquad (1)$$

Here, π_θ and π_{ref} denote the current and reference policies, respectively, and $\beta > 0$ controls the learning intensity of preference information. $\Delta_\theta(y_w, y_l)$ measures the log-probability gap between positive (y_w) and negative (y_l) samples under the current policy, while $\Delta_{\mathrm{ref}}(y_w, y_l)$ is the corresponding gap under the reference policy. $s(\theta)$ reflects how well the model aligns with preference information: larger values indicate stronger alignment. When $s(\theta) > 0$ and increases further, $\mathcal{L}_{\mathrm{DPO}}$ decreases accordingly. Since π_{ref} is typically frozen, $\Delta_{\mathrm{ref}}(y_w, y_l)$ can be regarded as constant. Therefore, minimizing $\mathcal{L}_{\mathrm{DPO}}$ essentially corresponds to enlarging $\Delta_\theta(y_w, y_l)$.

Subsequent studies [39, 2, 67] extend DPO to diffusion-based frameworks such as DDPM [68] and Rectified Flow [69], where the loss is formulated as:

$$L_{\mathrm{DPO}} = -\mathbb{E}_{(c,x_w,x_l)\sim\mathcal{D}}[\log\sigma(-\beta\cdot s(\theta))],$$

$$s(\theta) = \underbrace{(\|y^w - y_\theta(x_t^w,t)\|^2 - \|y^l - y_\theta(x_t^l,t)\|^2)}_{\Delta_\theta(x^w,x^l)} - \underbrace{(\|y^w - y_{\mathrm{ref}}(x_t^w,t)\|^2 - \|y^l - y_{\mathrm{ref}}(x_t^l,t)\|^2)}_{\Delta_{\mathrm{ref}}(x^w,x^l)}$$

$$(2)$$

Here, y^w and y^l denote the target outputs for the positive (x_w) and negative (x_l) samples, corresponding to noise in DDPM or vector fields in Rectified Flow. x_t^w and x_t^l represent intermediate noisy states at time step t. y_θ and y_{ref} represent the current and reference models, typically initialized identically. A negative and smaller $s(\theta)$ indicates that the current model distinguishes positive and negative samples better than the reference model, resulting in a smaller L_{DPO}. Since y_{ref} is typically frozen, $\Delta_{\mathrm{ref}}(x_w, x_l)$ can be regarded as constant. Therefore, minimizing L_{DPO} essentially corresponds to reducing $\Delta_\theta(x_w, x_l)$.

Moreover, the $Win\ Gap = (\|y^w - y_\theta(x_t^w,t)\|^2 - \|y^w - y_{\mathrm{ref}}(x_t^w,t)\|^2)$ captures the improvement of the current model over the reference model on positive samples; ideally it is negative, reflecting higher accuracy and implicit preference assignment. Conversely, the $Lose\ Gap = (\|y^l - y_\theta(x_t^l,t)\|^2 - \|y^l - y_{\mathrm{ref}}(x_t^l,t)\|^2)$ measures performance on negative samples; ideally positive, indicating lower prediction accuracy relative to the reference model, hence implicitly reducing their assigned probabilities. Equivalently, $s(\theta)$ can also be written as the difference between these two terms, i.e., $s(\theta) = Win\ Gap - Lose\ Gap$.

3.2.2 DPO lacks supervision over the sample distribution, leading to intrinsic instability.

The DPO objective (Eq.1) only constrains the variation of $\Delta_\theta(y_w, y_l)$ rather than directly regulating the distributions of $\pi_\theta(y_w \mid x)$ and $\pi_\theta(y_l \mid x)$, implying a lack of explicit control over the sample distribution. Ideally, π_θ should increase $\log\pi_\theta(y_w \mid x)$ while suppressing $\log\pi_\theta(y_l \mid x)$ to enlarge $\Delta_\theta(y_w, y_l)$. Yet, in Eq.1, π_θ may also increase this gap by jointly amplifying or shrinking both terms, as long as their difference grows, leading to intrinsic instability.

Eq.2 follows the same design philosophy as Eq.1 but expresses probability indirectly through prediction error—smaller error indicates higher probability:

$$\pi_{\theta/\mathrm{ref}}(y_{w/l} \mid x) \iff \|y^{w/l} - y_{\theta/\mathrm{ref}}(x_t^{w/l}, t)\|^2$$

Consequently, Eq.2 still constrains only the variation of $\Delta_\theta(x^w, x^l)$, without explicitly regularizing the individual reconstruction errors $\|y^w - y_\theta(x_t^w, t)\|^2$ and $\|y^l - y_\theta(x_t^l, t)\|^2$. The current model y_θ can thus minimize loss either by reducing the positive error while enlarging the negative one, or by simultaneously scaling both, thereby inheriting the same instability. Moreover, as shown in Appendix A.4, because entropy naturally favors increase, weaker constraints on $\|y^w - y_\theta(x_t^w, t)\|^2$ and $\|y^l - y_\theta(x_t^l, t)\|^2$ make both more likely to grow, especially when $y^{w/l}$ has higher dimensionality or a wider value range.

Taking the gradient of Eq.2 further reveals this instability (see Appendix A.5 for detailed derivation)

$$\nabla_\theta L_{\mathrm{DPO}} = -\mathbb{E}_{(c, x_w, x_l) \sim \mathcal{D}}[2\beta(1 - \sigma(S)) \cdot ((y^w - y_\theta(x_t^w, t))^\top \nabla_\theta y_\theta(x_t^w, t) \\ - (y^l - y_\theta(x_t^l, t))^\top \nabla_\theta y_\theta(x_t^l, t))], S = -\beta \cdot s(\theta) \tag{3}$$

We define the DPO gradient ratio (DGR) as:

$$DGR = \beta(1 - \sigma(S)) = \beta(1 - \frac{1}{1 + e^{\beta s(\theta)}}) \tag{4}$$

From Eq.2, DGR depends directly on $\Delta_\theta(x_w, x_l)$: smaller $\Delta_\theta(x_w, x_l)$ yield smaller $s(\theta)$ and hence lower DGR.

As shown in Fig. 3, DGR is upper-bounded by β and rapidly approaches zero as $s(\theta)$ decreases. A larger β yields a higher initial DGR but also induces a sharper decline, leading to gradient magnitude oscillations and optimization instability. In practice, DGR quickly diminishes to near zero, causing gradient vanishing and premature convergence, while the loss drops abruptly. As analyzed in Appendix A.4, due to the absence of constraints on the model's own output distribution, such rapid convergence tends to drive the model to enlarge the prediction error gap ($\Delta_\theta(x_w, x_l)$) between positive and negative samples by simultaneously increasing their individual errors ($\|y^w - y_\theta(x_t^w, t)\|^2$ and $\|y^l - y_\theta(x_t^l, t)\|^2$), thereby achieving rapid decay of the loss. Throughout this process, y_θ progressively diverges from y_{ref}, ultimately resulting in model collapse. A more detailed analysis of this phenomenon is provided in Section 4.3.

3.2.3 Reg-DPO: Stabilizing DPO via SFT Regularization

To address the inherent instability of DPO, we propose Reg-DPO (Fig. 1b). Prior analysis reveals that DPO's instability mainly stems from its reliance on the prediction error gap between positive and negative samples, while lacking direct supervision on the samples themselves. This allows the current model to reduce loss even when deviating substantially from the reference model.

Ideally, the model should reinforce the fitting of positive samples and suppress that of negative ones, thereby improving generation quality and widening the preference gap. Motivated by this insight, we introduce an explicit constraint on positive samples to enhance generative precision and training stability while preserving preference discrimination.

Specifically, we add a constant $r > 0$ to the positive-sample gradient term in Eq.3, imposing persistent supervision on positive samples. This ensures that optimization consistently moves toward positive directions even when DGR rapidly decays. The modified gradient is formulated as:

$$\nabla_\theta L_{\mathrm{Reg-DPO}} = -\mathbb{E}_{(c, x_w, x_l) \sim \mathcal{D}}[2(DGR + r)(y^w - y_\theta(x_t^w, t))^\top \nabla_\theta y_\theta(x_t^w, t) \\ - 2DGR(y^l - y_\theta(x_t^l, t))^\top \nabla_\theta y_\theta(x_t^l, t)] \tag{5}$$

where r controls the strength of positive-sample supervision. This modification is equivalent to introducing an SFT-style regularization weighted by r into Eq.2, yielding the Reg-DPO loss:

$$L_{\mathrm{Reg-DPO}} = \mathbb{E}_{(c, x_w, x_l) \sim \mathcal{D}}[-\log \sigma(-\beta \cdot s(\theta)) + r \cdot \|y^w - y_\theta(x_t^w, t)\|^2], \tag{6}$$

Reg-DPO thus integrates preference learning with sample-level supervision through the additional SFT regularization term $(r \cdot \|y^w - y_\theta(x_t^w, t)\|^2)$, leading to more stable training and a higher probability of generating preferred outputs.

Comprehensive experiments (Section 4.3) demonstrate that although the SFT term explicitly supervises only the output distribution of positive samples, it effectively enforces joint supervision over both positive and negative distributions, thereby alleviating the training instability of vanilla DPO and yielding smoother metric trajectories. Increasing r further strengthens this stabilizing effect and improves performance, with diminishing returns observed beyond $\sim 1 \times DGR$. Considering both stability and cross-scenario adaptability, we adopt a dynamic setting of $r = DGR$ as the default configuration for Reg-DPO.

4 Experiments

4.1 Experiment Settings

Datasets. We construct two high-quality datasets (details in Appendix A.1): Human-I2V and OpenVid-12K. Human-I2V targets human activity videos and comprises two 4,932-pair training subsets—Human-I2V-GT-Pair (GT-Pair) and Human-I2V-Annotated_Pair (manual annotation)—and a test set Human-I2V-Test with 495 samples. OpenVid-12K, derived from OpenVid-1M [70], covers diverse categories such as humans, animals, plants, machinery, and animation. It includes two 12K-pair training subsets—OpenVid-12K-I2V and OpenVid-12K-T2V—and a test set OpenVid-12K-Test with 500 samples.

Model Evaluation. We evaluate models using two methods: GSB and VBench (details in Appendix A.6). GSB is a manual evaluation contrasting outputs of model A and B using three labels—Good (A is better), Same (equivalent), and Bad (A is worse)—to compute the $Score$: $[Good + Same]/[Same + Bad]$. $Score > 1$ indicates A outperforms B, with larger values reflecting stronger advantages. Compared to automated metrics, GSB results better align with human preferences, offering higher reliability for video generation. We evaluate GSB along Video Quality (VQ), assessing overall visual appeal and smoothness, and Prompt Follow (PF), measuring consistency with the input.

VBench [36] is a standard automated evaluation tool. On our custom test sets, only six dimensions— subject consistency (Subj.Cons.), background consistency (Bkgd.Cons.), motion smoothness (Mot.Smooth.), dynamic degree (Dyn.Deg.), aesthetic quality (Aesth.Qual.), and imaging quality (Img.Qual.)—are applicable, with higher values denoting better performance. Methods are ranked per dimension; a Total_Score, aggregated across all six, enables direct comparison.

Training details are provided in Appendix A.7, and parameter study is provided in Appendix A.8.

4.2 Effectiveness of the GT-Pair

To validate the effectiveness of the GT-Pair, we use Wan2.1-I2V-14B-720P [10] as the baseline and conduct DPO training on Human-I2V. Four experiments are designed: (1) Base: baseline results; (2) GT-Pair: trained on Human-I2V-GT-Pair, assessing the GT-Pair; (3) Annotated_Pair: trained on Human-I2V-Annotated_Pair, assessing the manual annotation strategy; and (4) GT+Annotated: trained on the combined datasets (doubling the data volume) to evaluate joint effects.

GSB results (Table 1) show that GT-Pair substantially outperforms all others. Compared with Base, it achieves 278 Good labels in VQ with a VQ Score of 1.63, and a PF Score of 1.27, indicating marked improvement in both video quality and prompt adherence. Against Annotated_Pair, GT-Pair remains superior (VQ/PF Scores = 1.44/1.24), and even surpasses the mixed dataset (1.39/1.12), demonstrating that GT-Pair data alone yields more effective enhancement.

VBench results (Table 2) present a consistent ranking: GT-Pair > GT+Annotated > Base > Annotated_Pair. GT-Pair attains the best results across four dimensions, with a Total_Score of 19—far exceeding Annotated_Pair's 10—highlighting its strong advantage in improving multi-dimensional video quality. In contrast, Annotated_Pair exhibits instability, with most metrics and its Total_Score falling below the Base, suggesting vulnerability to label noise and ambiguous preference signals. The mixed dataset achieves a Total_Score of 17, outperforming Base and Annotated_Pair yet remaining inferior to GT-Pair, implying that data mixing introduces noise that weakens overall effectiveness.

In summary, GT-Pair consistently achieves the best results in both GSB and VBench, significantly improving the quality of generated videos.

Table 1: Effectiveness of the GT-Pair — GSB. Results for VQ and PF are reported in the format Good/Same/Bad.

Comparison (A vs B)	VQ	VQ Score	PF	PF Score
GT-Pair vs Base	278/75/141	1.63	168/246/80	1.27
GT-Pair vs Annotated_Pair	217/183/95	1.44	130/324/41	1.24
GT-Pair vs [GT+Annotated]	240/110/142	1.39	137/261/94	1.12

Table 2: Effectiveness of the GT-Pair — VBench. The best results are highlighted in bold, and the second-best results are underlined. Symbols indicate changes relative to the Base model: ↑ denotes improvement, ↓ denotes degradation, and – denotes no change. Numbers in square brackets [ ] represent the ranking scores of each method for the corresponding dimension (given n experimental groups, the top-ranked method receives n points, and the lowest-ranked method receives 1 point). Total_Score is the sum of ranking scores across the six dimensions.

Method	Subj.Cons.	Bkgd.Cons.	Mot.Smooth.	Dyn.Deg.	Aesth.Qual.	Img.Qual.	Total_Score
Base	0.93251 - [3]	0.93287 - [3]	0.98586 - [3]	0.52121 - [1]	0.54445 - [3]	0.74797 - [1]	14
GT-Pair	0.93108 ↓ [2]	0.93195 ↓ [1]	**0.98651** ↑ [4]	**0.56768** ↑ [4]	**0.54580** ↑ [4]	**0.75182** ↑ [4]	**19**
Annotated_Pair	0.92937 ↓ [1]	0.93276 ↓ [2]	0.98479 ↓ [2]	0.54595 ↑ [2]	0.54217 ↓ [1]	0.74872 ↑ [2]	10
GT+Annotated	**0.93311** ↑ [4]	**0.93453** ↑ [4]	0.98409 ↓ [1]	0.55758 ↑ [3]	0.54439 ↓ [2]	0.75033 ↑ [3]	17

4.3 Analysis of the SFT Regularization Term

To investigate the impact of the SFT regularization term in Reg-DPO, we train on OpenVid-12K-I2V with Wan2.1-I2V-14B-720P as the baseline and evaluate on OpenVid-12K-Test. Guided by DGR magnitude, we conduct 11 experiments: two controls (Base and DPO), five fixed-weight settings ($r = 25, 50, 125, 300, 500$), and four dynamic-weight settings ($r = DGR/5, DGR/2, DGR, DGR \times 2$). During training, we monitor key metrics (see Eqs. 2, 4, and 6 for explanations): DGR, r, $DGR + r$, $Win\ Gap$, $Lose\ Gap$, $s(\theta)$, and L_{DPO}. Checkpoints are saved every 300 steps and evaluated via VBench.

Results indicate that vanilla DPO suffers severe instability, causing continuous performance degradation and occasional collapse. Introducing the SFT regularization term suppresses this instability, yielding smoother training curves. As r increases, stabilization strengthens; beyond $\sim 1 \times DGR$, performance no longer declines but gradually improves without collapse.

Fig. 2 shows that during DPO training, DGR and L_{DPO} decline sharply over the same interval, reflecting a rapid negative surge in $s(\theta)$ (see Fig. 6). This corresponds to aggressive overfitting of preference signals: the model quickly enlarges the prediction error gap between positive and negative samples ($\Delta_\theta(x_w, x_l)$). Although this suggests apparent preference learning convergence, $Win\ Gap$ and $Lose\ Gap$ simultaneously rise sharply, indicating that predictions for both sample types deviate substantially from the reference distribution. This aligns with section 3.2.2: DPO optimizes only relative differences between positive and negative samples, lacking per-sample output constraints, which inherently destabilizes training. Fig. 7 further shows continual performance decline, and Fig. 8 visually confirms grid-like blurring in generated videos as training progresses, illustrating DPO-induced collapse.

With SFT regularization term, DPO instability is mitigated: sharp drops in DGR and L_{DPO} are smoothed, and abnormal increases in $Win\ Gap$ and $Lose\ Gap$ are controlled, indicating reduced distributional deviation from the reference. When r exceeds $\sim 1 \times DGR$, $Win\ Gap$ becomes negative, implying superior positive-sample predictions. As shown in Fig. 7, SFT's regularization strengthens with increasing r, and model performance steadily improves rather than declines. Fig. 8 confirms that Reg-DPO ($r = DGR$) maintains stability and consistently clear outputs throughout training, in stark contrast to vanilla DPO. Moreover, Table 10 demonstrates that higher r enables Reg-DPO to surpass both Base and DPO across multiple evaluation metrics, achieving a markedly higher Total_Score.

7

This improvement arises because SFT regularization preserves high gradient coefficients for both positive $(2(DGR + r))$ and negative $(2DGR)$ samples (see Eq. 5 for details), preventing the rapid decay to zero as observed in vanilla DPO (see Fig. 2 and Fig. 5 for comparison). As a result, convergence—i.e., the reduction of Eq. 5— requires minimizing both DGR and individual prediction errors $(y^w - \boldsymbol{y}_\theta(\boldsymbol{x}_t^w, t))$ and $(y^l - \boldsymbol{y}_\theta(\boldsymbol{x}_t^l, t))$. This mechanism provides supervision for both positive and negative outputs, effectively addressing the inherent instability of vanilla DPO. In other words, although SFT explicitly targets only positive samples, it implicitly constrains negative ones, enforcing simultaneous learning of preference differences and distributional consistency. Consequently, training stability and overall performance are substantially enhanced.

Moreover, as shown in Fig. 2 and Table 10, results across groups converge once r exceeds $\sim 1 \times DGR$. To avoid over-regularization that may hinder preference learning, and given task-dependent variations in DGR, we set $r = DGR$ as the default in Reg-DPO. This choice ensures both the effectiveness of the SFT regularization term and its generalizability across different scenarios.
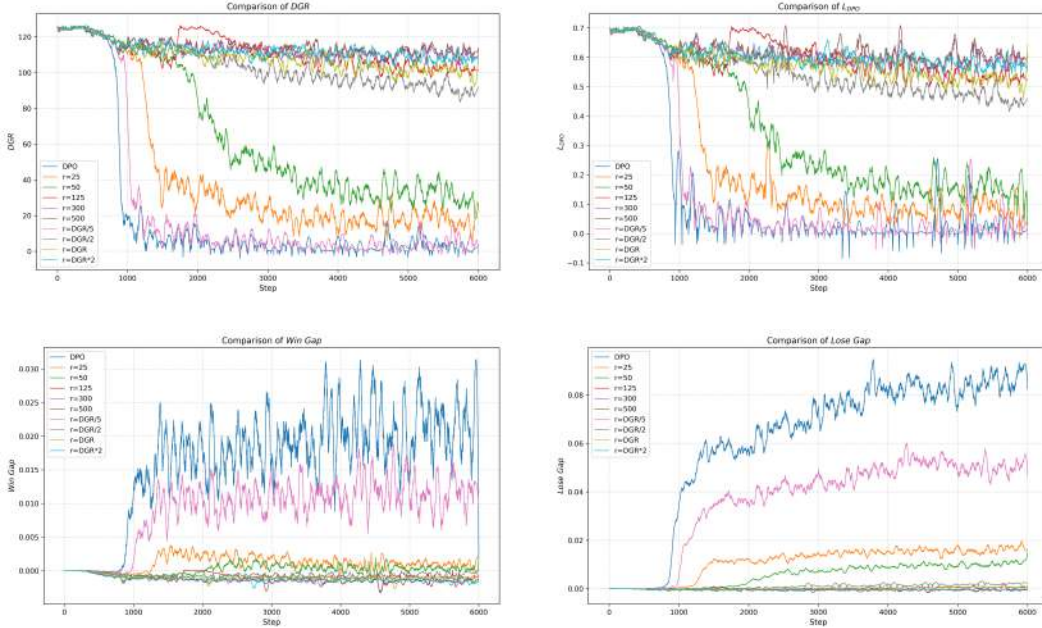

Figure 2: Training metric curves of different experimental groups. The magnitude relationship of r across groups is $DGR/5(\approx 25 \to 0) < 25 < 50 < DGR/2(\approx 65 \to 40) < DGR(\approx 125 \to 90) < 125 < DGR \times 2(\approx 255 \to 210) < 300 < 500$. The variation curves of r, $DGR + r$, and $s(\theta)$ are shown in Figs. 4, 5, and 6.

4.4 Evaluation of Reg-DPO

To evaluate the effectiveness of the Reg-DPO, we conduct experiments under both I2V and T2V modes.

For I2V, we use Wan2.1-I2V-14B-720P as the baseline model and perform training and evaluation on two datasets: 1) Human-I2V-GT-Pair for training and Human-I2V-Test for testing; 2) OpenVid-12K-I2V for training and OpenVid-12K-Test for testing. For T2V, we use Wan2.1-T2V-14B [10] as the baseline model, training on OpenVid-12K-T2V and evaluating on OpenVid-12K-Test.

Four experimental groups are designed per dataset to compare the performance of different training strategies: (1) Base: the baseline model without additional training; (2) SFT: model trained with supervised fine-tuning (SFT); (3) DPO: model trained with DPO; and (4) Reg-DPO: model trained with Reg-DPO.

4.4.1 I2V

The experimental results demonstrate that Reg-DPO consistently outperforms the baseline and alternative training strategies across both GSB (Tables 3 and 4) and VBench (Tables 5 and 6), validating its effectiveness and robustness for I2V generation.

In GSB results, Reg-DPO achieves the best performance on both datasets, significantly improving video quality and instruction adherence. On Human-I2V-GT-Pair (Table 3), it attains the largest gain over Base, with a VQ Score of 2.05 and a PF Score of 1.73, outperforming SFT (1.49/1.32) and DPO (1.57/1.26). On OpenVid-12K-I2V (Table 4), it likewise yields the greatest improvement over Base and maintains a clear lead over both SFT and DPO.

In VBench results, Reg-DPO again delivers the best overall performance with balanced gains across multiple dimensions, reflecting stronger optimization capability and robustness. On Human-I2V-GT-Pair (Table 5), it is the only method surpassing Base in all dimensions, ranking first in three and second in the others, with a Total_Score of 21, notably higher than DPO (15) and Base/SFT (12). On OpenVid-12K-I2V (Table 6), Reg-DPO ranks first in two dimensions and second in three, achieving a Total_Score of 18, exceeding DPO (15) and Base/SFT (14).

Notably, SFT exhibits a clear "dynamic decay" phenomenon in experiments: the Dyn.Deg. drops markedly, and the model tends to generate static or micro-motion videos. In contrast, Reg-DPO maintains balanced improvements across dimensions, effectively avoiding dynamic decay while preserving video quality and temporal coherence—demonstrating superior optimization stability and robustness.

Table 3: Experimental results on Human-I2V-GT-Pair — GSB

Comparison (A vs B)	VQ	VQ Score	PF	PF Score
Reg-DPO vs Base	308 / 76 / 111	2.05	277 / 101 / 117	1.73
Reg-DPO vs SFT	258 / 95 / 142	1.49	186 / 222 / 87	1.32
Reg-DPO vs DPO	262 / 104 / 129	1.57	193 / 189 / 113	1.26

Table 4: Experimental results on OpenVid-12K-I2V — GSB

Comparison (A vs B)	VQ	VQ Score	PF	PF Score
Reg-DPO vs Base	242 / 157 / 101	1.55	172 / 285 / 43	1.39
Reg-DPO vs SFT	203 / 173 / 124	1.27	102 / 342 / 56	1.12
Reg-DPO vs DPO	169 / 232 / 99	1.21	150 / 281 / 69	1.23

Table 5: Experimental results on Human-I2V-GT-Pair - VBench

Method	Subj.Cons.	Bkgd.Cons.	Mot.Smooth.	Dyn.Deg.	Aesth.Qual.	Img.Qual.	Total_Score
Base	0.93251 - [2]	0.93287 - [2]	0.98586 - [2]	0.52121 - [2]	0.54445 - [2]	0.74797 - [2]	12
SFT	**0.93749** ↑ [4]	**0.93752** ↑ [4]	0.98375 ↓ [1]	0.46869 ↓ [1]	0.54427 ↓ [1]	0.74728 ↓ [1]	12
DPO	0.93108 ↓ [1]	0.93195 ↓ [1]	0.98651 ↑ [3]	**0.56768** ↑ [4]	0.54580 ↑ [3]	0.75182 ↑ [3]	15
Reg-DPO	0.93496 ↑ [3]	0.93380 ↑ [3]	**0.98665** ↑ [4]	0.54343 ↑ [3]	0.54963 ↑ [4]	**0.75758** ↑ [4]	**21**

4.4.2 T2V

Both GSB (Tables 8) and VBench (Tables 9) results demonstrate that Reg-DPO achieves the best overall performance, significantly improving video quality. In GSB results, Reg-DPO obtains 297 Good labels in VQ with a VQ Score of 1.91 over Base, indicating substantial enhancement. Compared with SFT and DPO, it maintains consistent advantages in both visual quality and instruction adherence. Moreover, the GSB results highlight distinct optimization tendencies: SFT emphasizes instruction adherence, DPO focuses on visual quality, whereas Reg-DPO effectively integrates both, achieving synergistic improvement in visual quality and instruction adherence. In VBench results, Reg-DPO attains the highest Total Score, surpassing DPO, SFT, and Base.

Table 6: Experimental results on OpenVid-12K-I2V — VBench

Method	Subj.Cons.	Bkgd.Cons.	Mot.Smooth.	Dyn.Deg.	Aesth.Qual.	Img.Qual.	Total_Score
Base	0.93159 - [2]	0.93246 - [1]	0.98540 - [2]	**0.68200** - [4]	0.55419 - [1]	**0.72488** - [4]	14
SFT	0.93052 ↓ [1]	**0.93791** ↑ [4]	**0.98802** ↑ [4]	0.59400 ↓ [1]	0.55810 ↑ [3]	0.72031 ↓ [1]	14
DPO	0.93280 ↑ [3]	0.93289 ↑ [2]	0.98530 ↓ [1]	**0.68200** - [4]	0.55600 ↑ [2]	0.72357 ↓ [3]	15
Reg-DPO	**0.93876** ↑ [4]	0.93420 ↑ [3]	0.98560 ↑ [3]	0.67200 ↓ [2]	**0.55970** ↑ [4]	0.72054 ↓ [2]	**18**

The experimental results for T2V and I2V exhibit notable differences, primarily due to the inherent generative characteristics of the two tasks and the properties of the OpenVid-12K-T2V dataset. We provide a detailed analysis of these differences in Appendix A.9.

4.5 Case Comparison

To assess the effectiveness of Reg-DPO in video generation, we conduct frame-by-frame comparisons of test results before and after optimization. The main improvements are summarized as follows: (1) Enhanced visual consistency: better semantic preservation and fewer irrelevant artifacts (Fig. 10-a), with videos exhibiting more uniform color and brightness over time (Fig. 10-f). (2) Reduction of micro-motion videos: the optimized model generates fluent, complete actions instead of nearly static scenes (Fig. 10-b); (3) Improved generation stability: fewer visual anomalies and crashes occur (Fig. 10-c); (4) Enhanced physical plausibility: generated content aligns with real-world logic, with correct object positioning and interactions (Fig. 10-d). (5) Stronger instruction adherence: generated videos more faithfully follow textual prompts (Fig. 10-e); (6) Improved temporal coherence: actions transition smoothly with abrupt motion largely eliminated (Fig. 10-g);

5 Conclusion

This work proposes a practical and efficient framework for video DPO training. In data construction, GT-Pair enables automatic generation of high-quality preference pairs without external annotations. In algorithm design, Reg-DPO integrates an SFT regularization term, substantially enhancing training stability and generation quality. In memory optimization, multiple strategies enable efficient DPO training for ultra-large models. Extensive experiments validate the effectiveness of the GT-Pair, the contribution of SFT regularization, and the superior performance of Reg-DPO in optimizing video generation.

References

[1] 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.

[2] Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, et al. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*, 2025.

[3] 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.

[4] 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.

[5] Jiacheng Zhang, Jie Wu, Weifeng Chen, Yatai Ji, Xuefeng Xiao, Weilin Huang, and Kai Han. Onlinevpo: Align video diffusion model with online video-centric preference optimization. *arXiv preprint arXiv:2412.15159*, 2024.

[6] Wenxu Qian, Chaoyue Wang, Hou Peng, Zhiyu Tan, Hao Li, and Anxiang Zeng. Rdpo: Real data preference optimization for physics consistency video generation. *arXiv preprint arXiv:2506.18655*, 2025.

[7] Shicheng Li, Lei Li, Kun Ouyang, Shuhuai Ren, Yuanxin Liu, Yuanxing Zhang, Fuzheng Zhang, Lingpeng Kong, Qi Liu, and Xu Sun. Temple: Temporal preference learning of video llms via difficulty scheduling and pre-sft alignment. *arXiv preprint arXiv:2503.16929*, 2025.

[8] 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.

[9] 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.

[10] 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.

[11] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

[12] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

[13] NVIDIA Corporation. *context_parallel package — Megatron Core Developer Guide*. NVIDIA, 01 2023. URL `https://docs.nvidia.com/megatron-core/developer-guide/latest/api-guide/context_parallel.html#context-parallelism-overview`. Megatron-Core version $\geq$0.5.0, Transformer Engine version $\geq$1.1 required for enabling context parallelism.

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[15] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[16] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. 2023.

[17] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

[18] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[19] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

[20] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

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

[22] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[23] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) 2023*. Neural Information Processing Systems Foundation, 2023.

[24] Chaofeng Chen, Annan Wang, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Enhancing diffusion models with text-encoder reinforcement learning. In *European Conference on Computer Vision*, pages 182–198. Springer, 2024.

[25] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *Advances in Neural Information Processing Systems*, 37:76177–76209, 2024.

[26] Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. In *European Conference on Computer Vision*, pages 108–124. Springer, 2024.

[27] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

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

[29] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[30] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024.

[31] Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback. *arXiv preprint arXiv:2412.02617*, 2024.

[32] Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*, 2024.

[33] 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.

[34] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

[35] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024.

[36] 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.

[37] 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.

[38] 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.

[39] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

[40] 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.

[41] 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.

[42] Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. Dspo: Direct score preference optimization for diffusion model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

[43] Ryotaro Kawata, Kazusato Oko, Atsushi Nitanda, and Taiji Suzuki. Direct distributional optimization for provable alignment of diffusion models. *arXiv preprint arXiv:2502.02954*, 2025.

[44] Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. *arXiv preprint arXiv:2505.11245*, 2025.

[45] Sanghyeon Na, Yonggyu Kim, and Hyunjoon Lee. Boost your human image generation model via direct preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23551–23562, 2025.

[46] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. Curriculum direct preference optimization for diffusion and consistency models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2824–2834, 2025.

[47] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13199–13208, 2025.

[48] Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8020–8030, 2025.

[49] Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18465–18475, 2025.

[50] Yunhong Lu, Qichao Wang, Hengyuan Cao, Xierui Wang, Xiaoyin Xu, and Min Zhang. Inpo: Inversion preference optimization with reparametrized ddim for efficient diffusion model alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28629–28639, 2025.

[51] Minghao Fu, Guo-Hua Wang, Liangfu Cao, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Chats: Combining human-aligned optimization and test-time sampling for text-to-image generation. *arXiv preprint arXiv:2502.12579*, 2025.

[52] Zijing Hu, Fengda Zhang, and Kun Kuang. D-fusion: Direct preference optimization for aligning diffusion models with visually consistent samples. *arXiv preprint arXiv:2505.22002*, 2025.

[53] Yunhong Lu, Qichao Wang, Hengyuan Cao, Xiaoyin Xu, and Min Zhang. Smoothed preference optimization via renoise inversion for aligning diffusion models with varied human preferences. *arXiv preprint arXiv:2506.02698*, 2025.

[54] Jie Ren, Yuhang Zhang, Dongrui Liu, Xiaopeng Zhang, and Qi Tian. Refining alignment framework for diffusion models with intermediate-step preference ranking. *arXiv preprint arXiv:2502.01667*, 2025.

[55] Jaa-Yeon Lee, Byunghee Cha, Jeongsol Kim, and Jong Chul Ye. Aligning text to image in diffusion models is easier than you think. *arXiv preprint arXiv:2503.08250*, 2025.

[56] X. Shen, Z. Li, Z. Yang, S. Zhang, Y. Zhang, D. Li, others, and Y. Tang. Directly aligning the full diffusion trajectory with fine-grained human preference. *arXiv preprint: 2509.06942*, 2025.

[57] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

[58] Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-Yu Liu, Jun Zhu, and Haoxiang Wang. Bridging supervised learning and reinforcement learning in math reasoning. *arXiv preprint arXiv:2505.18116*, 2025.

[59] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

[60] Bo Wang, Qinyuan Cheng, Runyu Peng, Rong Bao, Peiji Li, Qipeng Guo, Linyang Li, Zhiyuan Zeng, Yunhua Zhou, and Xipeng Qiu. Implicit reward as the bridge: A unified view of sft and dpo connections. *arXiv preprint arXiv:2507.00018*, 2025.

[61] Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*, 2025.

[62] Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. *Advances in Neural Information Processing Systems*, 37:138663–138697, 2024.

[63] D. Zhang, G. Lan, D. Han, W. Yao, X. Pan, H. Zhang, M. Li, P. Chen, Y. Dong, C. G. Brinton, and J. Luo. Bridging sft and dpo for diffusion model alignment with self-sampling preference optimization. *arXiv preprint arXiv:2410.05255*, 2024.

[64] Jisoo Kim, Wooseok Seo, Junwan Kim, Seungho Park, Sooyeon Park, and Youngjae Yu. Vip: Iterative online preference distillation for efficient video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17235–17245, 2025.

[65] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[66] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[67] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

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

[69] 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.

[70] 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.

[71] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[72] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7752–7762, 2025.

[73] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[74] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022.

[75] PKU-Yuan Lab, Tuzhan AI, et al. Open-sora-plan, 04 2024. URL `https://doi.org/10.5281/zenodo.10948109`. Open-source project for video generation, released by PKU-Yuan Lab and Tuzhan AI.

[76] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.

[77] Kling AI. Kling ai: Next-gen ai video & ai image generator. `https://app.klingai.com/cn/`, n.d. Accessed: 2025-10-26.

[78] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A Appendix

A.1 Data Construction

Due to the complex spatiotemporal characteristics of video data, video DPO training is particularly sensitive to data quality compared to text and image modalities. Consequently, data construction plays a critical role throughout the training process. Insufficient data quality—such as low distinguishability between positive and negative samples, low-quality positive samples, or mismatched data formats relative to the baseline model—can easily lead to training collapse. Ideal training data satisfy the following criteria: 1) a significant gap between positive and negative samples, ensuring clear distinguishability; 2) high-quality positive samples; 3) data format (resolution, aspect ratio, frame rate, number of frames, etc.) consistent with the baseline model. For instance, in this study, the baseline models are Wan2.1-I2V-14B-720P [10] and Wan2.1-T2V-14B [10]. Accordingly, based on the model generation specifications and task requirements, we unify the data format to 16 FPS, 81 frames, and a minimum side resolution of 720 pixels.

Based on this setup, we construct two datasets: Human-I2V, focusing primarily on human activities, and OpenVid-12K, derived from the publicly available OpenVid-1M [70], covering diverse video content. The overall data construction pipeline comprises five stages: data acquisition, data filtering, data normalization, data generation, and pair construction.

During the data acquisition stage, for Human-I2V, we select approximately 5K raw videos featuring human activities from multiple publicly available datasets [71–73] with the following criteria: frame rates between 16–45 FPS, duration of 5–6 seconds, and a minimum edge resolution no less than 512 pixels. Based on the task requirements, we only retain videos with aspect ratios close to 9:16, 3:4, or 1:1. For OpenVid-12K, we select about 20K videos from OpenVid-1M with consistent criteria. Since OpenVid-1M predominantly contains 16:9 videos, only videos approximating this aspect ratio are included. OpenVid-1M contains 1M high-quality text-video pairs, of which 433K have a resolution of 1920×1080. All videos are real-world footage sourced from ChronoMagic, CelebV-HQ [74], Open-Sora-plan [75], Panda-70m [76], and are further filtered and processed by the authors to enhance quality, content richness, and prompt detail.

In the data filtering stage, to ensure quality, we remove videos with static frames, frequent transitions, visual effects, excessive text, or image montages/slideshows. Even after initial selection, variations in aspect ratio, resolution, frame rate, and frame count prevent direct use for DPO training, necessitating further normalization. Specifically, we: 1) unify resolution by scaling the shortest edge to at least 720 pixels; 2) unify aspect ratio via center cropping to 9:16/3:4/1:1 (Human-I2V) or 16:9 (OpenVid-12K); 3) unify frame rate by downsampling to 16 FPS; 4) unify frame count by truncating excess frames to 81. After normalization, all real videos meet the specifications of 16 FPS, 81 frames, aspect ratios of 9:16/3:4/1:1 or 16:9, and a minimum edge resolution of 720 pixels. We refer to these as GT (Ground Truth) videos. We then extract the first frame of each GT video as a reference image and generate video prompts automatically using MLLM for data generation or model evaluation. Since OpenVid-1M provides detailed prompts, we do not generate additional prompts for its videos. Ultimately, Human-I2V contains 996 GT videos for training, with an additional 495 GT videos used solely for reference frame extraction and prompt generation as the test set, denoted Human-I2V-Test. OpenVid-12K retains 12K GT videos for training, with 500 GT videos reserved for reference and prompt generation as the test set, denoted OpenVid-12K-Test.

In the data generation stage, multiple video generation models are employed to produce videos for subsequent Pair construction. For Human-I2V, using GT reference frames and prompts, we employ five I2V models: Wan2.1-I2V-14B-720P [10], CogVideoX1.5-5B-I2V [38], Kling2.1 [77], HunyuanVideo-I2V [9], and SkyReels-V2-I2V-14B-720P [8], producing 996, 996, 968, 996, and 976 videos, respectively. For OpenVid-12K, we use Wan2.1-I2V-14B-720P and Wan2.1-T2V-14B to generate 12K videos in I2V and T2V modes, respectively. As output specifications vary across models, even after parameter tuning or source modifications, generated videos undergo additional normalization to match 16 FPS, 81 frames, aspect ratios of 9:16/3:4/1:1 or 16:9, and minimum edge resolution of 720 pixels.

During the Pair construction stage, for Human-I2V, we apply two strategies: first, the GT-Pair strategy, which pairs GT videos as positive samples with model-generated videos as negatives, creating 4,932 training pairs, denoted Human-I2V-GT-Pair. Second, the manual annotation strategy combines generated videos in pairs and employs three or more annotators to score them across five dimensions

(image quality, motion quality, reference consistency, instruction adherence, overall quality) on a 5-point scale, requiring approximately 10 days. This yields 9,788 pairs; samples with higher average scores serve as positive samples and lower scores as negatives. Of these, 8,692 pairs have an average score difference ≥ 1, 3,250 pairs ≥ 2, and 1,722 positive samples have average scores ≥ 4. Only 1,068 pairs satisfy both average score difference ≥ 2 and positive score ≥ 4, representing the ideal high-quality training subset ($\sim 10.9\%$). To match the Human-I2V-GT-Pair volume for comparative evaluation, we retain all pairs with score difference ≥ 2 and randomly select additional pairs with a difference of 1 to reach 4,932 pairs, denoted Human-I2V-Annotated_Pair. This illustrates that manual annotation is costly, time-consuming, and yields a low proportion of truly high-quality data. For OpenVid-12K, only the GT-Pair strategy is applied, pairing I2V and T2V generated videos with their corresponding GT videos, producing OpenVid-12K-I2V and OpenVid-12K-T2V, each containing 12K pairs. These subsets primarily evaluate the method's performance across comprehensive datasets and different video generation modalities.

A.2 Memory Optimization

In large-scale video generation models, Direct Preference Optimization (DPO) fine-tuning faces memory consumption as a critical bottleneck that limits both training efficiency and model scalability. Unlike image or text tasks, video DPO training requires simultaneously handling high-resolution inputs, long temporal frame sequences, and paired sample structures ([win, lose]), resulting in computational and storage demands far exceeding conventional generative tasks. For instance, the Wan2.1-I2V-14B-720P model [10] adopts a Diffusion Transformer (DiT) architecture with over 14B parameters. During training, considering model weights, gradients, and optimizer states (e.g., AdamW), the total parameter footprint increases substantially; additionally, the DPO framework requires loading a frozen reference model of the same scale for preference comparison. Altogether, approximately 70B parameters reside in memory concurrently. Even with float16 precision, the model and optimizer alone occupy around 130 GB of GPU memory. Moreover, video generation models typically include auxiliary modules such as a text encoder and video variational autoencoder, further increasing memory demand. The challenge intensifies as video samples consist of multiple frames, with input sizes tens to hundreds of times larger than typical image tasks, and DPO training requires parallel loading of paired video samples [win, lose]. Under these compounded factors, memory requirements grow exponentially, frequently triggering out-of-memory (OOM) errors and severely constraining the efficiency and scalability of large models.

To address these challenges, we implement a systematic memory optimization strategy based on Fully Sharded Data Parallel (FSDP) [11]. The strategy integrates six key techniques: Flash Attention [12], Context Parallelism [13], fully parallelized pair computation, prompt pre-encoding, model offloading, and computational graph and memory reclamation optimizations, aiming to comprehensively enhance training efficiency and scalability for ultra-large video generation models. After full optimization, our approach achieves over threefold improvement compared to the FSDP-only baseline. For example, using the Wan2.1-I2V-14B-720P model, this scheme stably supports single-card training of video pairs with 720×1280 resolution and 200 frames per video on an 8× NVIDIA H100 (80GB) setup. This framework provides a reusable, system-level optimization paradigm for future high-resolution, long-sequence video generation training. The following sections detail each of the key optimization strategies.

1. Flash Attention
Flash Attention is an efficient attention computation algorithm that optimizes data read/write flows by fully leveraging the GPU memory hierarchy (e.g., SRAM and HBM), achieving significant speedup and memory efficiency. Its core principle is to partition inputs and perform local attention computations within on-chip caches, combined with tiling and recomputation to reduce memory accesses. This method reduces memory complexity from $O(N^2)$ to $O(N)$ while maintaining computational accuracy, providing 2–4× acceleration and substantial memory savings, thus supporting longer contexts and larger-scale models for efficient training. In video generation tasks, attention complexity grows quadratically with the number of frames. We adopt Flash Attention to significantly reduce intermediate memory buffers while maintaining numerical stability, effectively mitigating the memory pressure caused by multi-frame inputs.

2. Context Parallel
Context Parallel (CP) is an efficient parallelization strategy along the sequence dimension, repre-

senting an enhanced form of sequence parallelism. CP partitions all inputs and activations along the sequence dimension and employs Flash Attention2 [12] for local attention computation. GPUs exchange KV data in a ring structure to achieve cross-block self-attention aggregation. This method preserves computational correctness while significantly reducing memory usage and improving parallel efficiency for long-sequence attention. In this study, we apply CP to the temporal dimension (frame dimension) of videos, enabling parallel computation and memory sharing along the time axis. Under the same batch size, memory consumption decreases proportionally with the number of partitions, providing robust support for training long-sequence videos.

3. Full Parallel Processing of Pairs

In DPO training, sequential computation of positive and negative samples ([win, lose]) generates redundant gradient graphs and causes memory spikes during backpropagation, particularly under FSDP. To address this, we redesign the training pipeline to execute both sample paths fully in parallel during the forward pass. Experiments show that this optimization reduces peak memory usage by approximately 50% on average within the FSDP framework while maintaining model performance and substantially increasing system throughput.

4. Pre-encoding Prompts

In video generation training, textual prompts require tokenization and encoding into semantic vectors via a text encoder. Dynamic encoding incurs significant computational and memory overhead and results in repeated computation. We perform offline pre-encoding of all prompts before training, converting texts directly into fixed-dimensional vectors. During training, these vectors are loaded from cache without invoking the text encoder again. This strategy effectively eliminates memory usage from forward passes through the text encoder and accelerates the overall data loading stage.

5. Model Offloading for Frozen Modules

In DPO training, modules such as the image VAE, video VAE, and frozen reference model participate only in forward computation without requiring gradient updates. To prevent these modules from occupying GPU memory unnecessarily, we implement a module-level offloading mechanism: after forward computation, module weights are immediately moved to CPU memory and reloaded to GPU only when needed. This approach reduces peak memory by approximately 10 GB on average. Although it introduces some I/O overhead (~ 1 minute per batch), the memory savings outweigh the cost in memory-constrained scenarios.

6. Graph and Memory Management

To further improve memory utilization, we refine computational graph construction and memory reclamation strategies, including:

- Explicitly setting `requires_grad_(False)` or enclosing tensors in `torch.no_grad()` for non-gradient tensors to prevent unnecessary graph creation;
- Using `.detach()` on intermediate variables where safe, shortening backpropagation paths;
- Actively deleting unused variables via `del` and calling `torch.cuda.empty_cache()` at the end of each iteration to reclaim memory promptly.

These optimizations significantly reduce memory fragmentation, lowering overall peak memory by approximately 3 GB, and enhance system stability and sustainable training capability.

A.3 Derivation of the objective function of DPO

DPO [1] is a preference optimization algorithm derived from RLHF [14, 65, 66], which can be jointly formulated from the objective functions of PPO [21] and the reward model. The PPO objective aims to maximize the expected reward while constraining the divergence between the current policy π_θ and the reference policy π_{ref} :

$$\max_{\pi_\theta} \mathbb{E}_{x \sim D, \, y \sim \pi_\theta(y|x)}[r_\phi(x,y)] - \beta D_{\text{KL}}[\pi_\theta(y|x) \| \pi_{\text{ref}}(y|x)] \tag{7}$$

where $r_\phi(x, y)$ denotes the reward function that evaluates the quality of (x, y), and $\beta > 0$ controls the trade-off between reward maximization and distributional stability. Solving Eq. 7 yields the optimal policy:

$$\pi_\theta(y|x) = \frac{1}{Z(x)} \pi_{ref}(y|x) \, e^{\frac{1}{\beta} r_\phi(x,y)} \tag{8}$$

where

$$Z(x) = \sum_y \pi_{ref}(y|x) \, e^{\frac{1}{\beta} r_\phi(x,y)}.$$

This reveals an implicit relationship between the reward function and the policy, showing that the policy itself encodes reward signals:

$$r_\phi(x,y) = \beta log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} + \beta log Z(x) \tag{9}$$

The reward model aims to maximize the reward difference between positive and negative samples, formulated as:

$$\mathcal{L}_{\text{RM}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(r_\phi(x,y_w) - r_\phi(x,y_l) \right) \right] \tag{10}$$

where

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

Substituting Eq. 9 into Eq. 10 yields the DPO objective:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right] \tag{11}$$

Expanding Eq. 11, we obtain:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log \sigma(\beta \cdot s(\theta))] \tag{12}$$

where

$$s(\theta) = \underbrace{\left[\log \pi_\theta(y_w \mid x) - \log \pi_\theta(y_l \mid x) \right]}_{\Delta_\theta(y_w,y_l)} - \underbrace{\left[\log \pi_{\text{ref}}(y_w \mid x) - \log \pi_{\text{ref}}(y_l \mid x) \right]}_{\Delta_{\text{ref}}(y_w,y_l)}$$

Detailed Derivation from Eq. 7 to Eq. 8

$$\begin{aligned}
&\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)} \left[r_\phi(x,y) \right] - \beta D_{\text{KL}} \left(\pi_\theta(y|x) \| \pi_{\text{ref}}(y|x) \right) \\
&= \max_{\pi_\theta} \mathbb{E}_{x\sim D} \mathbb{E}_{y\sim\pi_\theta(y|x)} \left[r_\phi(x,y) \right] - \beta \sum_{x\sim D, y\sim\pi_\theta(y|x)} \pi_\theta(y|x) \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \\
&= \max_{\pi_\theta} \mathbb{E}_{x\sim D} \mathbb{E}_{y\sim\pi_\theta(y|x)} \left[r_\phi(x,y) \right] - \beta \mathbb{E}_{x\sim D} \mathbb{E}_{y\sim\pi_\theta(y|x)} \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \\
&= \min_{\pi_\theta} \mathbb{E}_{x\sim D} \mathbb{E}_{y\sim\pi_\theta(y|x)} \left[\log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r_\phi(x,y) \right] \\
&= \min_{\pi_\theta} \mathbb{E}_{x\sim D} \mathbb{E}_{y\sim\pi_\theta(y|x)} \left[\log \frac{\pi_\theta(y|x)}{\frac{1}{Z(x)}\pi_{\text{ref}}(y|x)e^{\frac{1}{\beta} r_\phi(x,y)}} - \log Z(x) \right]
\end{aligned} \tag{13}$$

where

$$Z(x) = \sum_y \pi_{ref}(y|x) \, e^{\frac{1}{\beta} r_\phi(x,y)}$$

Define the policy:

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{ref}(y|x) \, e^{\frac{1}{\beta} r_\phi(x,y)} \tag{14}$$

Substituting $\pi^*(y|x)$ into Eq. 13

$$\begin{aligned}
&\min_{\pi_\theta} \mathbb{E}_{x\sim D} \mathbb{E}_{y\sim\pi_\theta(y|x)} \left[\log \frac{\pi_\theta(y|x)}{\frac{1}{Z(x)}\pi_{\text{ref}}(y|x)e^{\frac{1}{\beta} r_\phi(x,y)}} - \log Z(x) \right] \\
&= \min_{\pi_\theta} \mathbb{E}_{x\sim D} \mathbb{E}_{y\sim\pi_\theta(y|x)} \left[\log \frac{\pi_\theta(y|x)}{\pi^*(y|x)} - \log Z(x) \right] \\
&= \min_{\pi_\theta} \mathbb{E}_{x\sim D} \left[\mathbb{E}_{y\sim\pi_\theta(y|x)} \log \frac{\pi_\theta(y|x)}{\pi^*(y|x)} - \log Z(x) \right] \\
&= \min_{\pi_\theta} \mathbb{E}_{x\sim D} \left[\mathbb{D}_{\text{KL}} \left(\pi_\theta(y|x) \| \pi^*(y|x) \right) - \log Z(x) \right]
\end{aligned} \tag{15}$$

Eq. 15 shows that the minimization is achieved if and only if:

$$\pi_\theta(y|x) = \pi^*(y|x) = \frac{1}{Z(x)}\pi_{ref}(y|x)\, e^{\frac{1}{\beta}r_\phi(x,y)} \tag{16}$$

A.4 Analysis of the Directional Change of $\left\|y^w - y_\theta(x_t^w, t)\right\|^2$ and $\left\|y^l - y_\theta(x_t^l, t)\right\|^2$ in Eq. 2

Let the n-dimensional space containing y^w be denoted as Y, where the Euclidean distance between any two points is at most $2R$. Assume y^w is a fixed point and define $r = \left\|y^w - y_\theta(x_t^w, t)\right\|^2$. If the update of $y_\theta(x_t^w, t)$ is unconstrained, it can be viewed as a new uniform sample drawn within Y. According to Proposition 1, we can reasonably infer that when n or R is larger, or r is smaller, the new sample is more likely to fall outside a sphere centered at y^w with radius r. In other words, $\|y^w - y_\theta(x_t^w, t)\|^2$ is more likely to increase. Put differently, in higher-dimensional spaces (larger n) with larger spatial extent (larger R) and when the current model y_θ predicts y^w with smaller error (smaller r), the unconstrained gradient update will most likely move the new $y_\theta(x_t^w, t)$ away from y^w, i.e., the prediction error $\|y^w - y_\theta(x_t^w, t)\|^2$ tends to increase. Moreover, as $n \to \infty$, $R \to \infty$, or $r \to 0$, the probability of $\|y^w - y_\theta(x_t^w, t)\|^2$ increasing approaches 1. The change behavior of $\|y^l - y_\theta(x_t^l, t)\|^2$ follows the same pattern as that of $\|y^w - y_\theta(x_t^w, t)\|^2$.

Proposition 1 (Concentration of Volume in High-Dimensional Spheres)
Consider a closed sphere in an n-dimensional Euclidean space defined as

$$B^n(A, R) = \{x \in \mathbb{R}^n : \|x - A\| \le R\} \tag{17}$$

where A is the center and $R > 0$ is the radius. Let a random point y be uniformly distributed within $B^n(A, R)$, and denote its distance to the center as $\|y - A\|$. Then, for any fixed $0 < r < R$,

$$P(\|y - A\| \le r) = \left(\frac{r}{R}\right)^n \tag{18}$$

As r decreases or R or n increases, $P(\|y - A\| \le r)$ decreases, meaning the probability of the random point y falling outside the inner sphere of radius r increases. Specifically, as $r \to 0$, $R \to \infty$, or $n \to \infty$,

$$P(\|y - A\| \le r) = \left(\frac{r}{R}\right)^n \to 0 \tag{19}$$

indicating that y almost certainly falls outside the inner sphere of radius r from the center A.

Derivation of Eq. 18
Since y is uniformly distributed within $B^n(A, R)$, the probability distribution of its distance to the center can be expressed as the volume ratio:

$$P(\|y - A\| \le r) = \frac{V(B^n(r))}{V(B^n(R))}$$

where $V(B^n(R))$ denotes the volume of an n-dimensional sphere with radius R. Using the volume formula for an n-dimensional sphere,

$$V(B^n(R)) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} R^n$$

the volume ratio simplifies to

$$\frac{V(B^n(r))}{V(B^n(R))} = \left(\frac{r}{R}\right)^n$$

Therefore,

$$P(\|y - A\| \le r) = \left(\frac{r}{R}\right)^n$$

A.5 Gradient Derivation of Eq. 2

The DPO loss is defined as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(c, x_w, x_l) \sim \mathcal{D}}\left[\log \sigma\left(-\beta \cdot s(\theta)\right)\right] \tag{20}$$

where

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$s(\theta) = \underbrace{\left(\left\| y^w - y_\theta(x_t^w, t) \right\|^2 - \left\| y^w - y_{\text{ref}}(x_t^w, t) \right\|^2 \right)}_{\text{Win Gap}} - \underbrace{\left(\left\| y^l - y_\theta(x_t^l, t) \right\|^2 - \left\| y^l - y_{\text{ref}}(x_t^l, t) \right\|^2 \right)}_{\text{Lose Gap}}$$

Since y_{ref} is frozen, $\nabla_\theta y_{ref} = 0$. For clarity, we define $S = -\beta \cdot s(\theta)$ and omit $(c, x_w, x_l) \sim \mathcal{D}$ in the derivation.

Step1: Outer derivative

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{\text{DPO}} &= - \nabla_\theta \mathbb{E}[\log \sigma(S)] \\
&= - \mathbb{E}\left[\nabla_\theta \log \sigma(S)\right] \\
&= - \mathbb{E}\left[\frac{\sigma'(S)}{\sigma(S)} \nabla_\theta S\right]
\end{aligned} \tag{21}$$

Using $\sigma'(x) = \sigma(s)(1 - \sigma(x))$, we obtain:

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[(1 - \sigma(S)) \nabla_\theta S\right] \tag{22}$$

Step2: Derivative of S
For a single term $\|y - y_\theta\|^2$:

$$\begin{aligned}
\frac{\partial}{\partial \theta} \|\mathbf{y} - \mathbf{y}_\theta\|^2 &= 2(\mathbf{y}_\theta - \mathbf{y})^\top \nabla_\theta \mathbf{y}_\theta \\
&= -2(\mathbf{y} - \mathbf{y}_\theta)^\top \nabla_\theta \mathbf{y}_\theta
\end{aligned} \tag{23}$$

Since the reference model is frozen:

$$\frac{\partial}{\partial \theta} \text{Win Gap} = -2(\mathbf{y}^w - \mathbf{y}_\theta(\mathbf{x}_t^w, t))^\top \nabla_\theta \mathbf{y}_\theta(\mathbf{x}_t^w, t) \tag{24}$$

$$\frac{\partial}{\partial \theta} \text{Lose Gap} = -2(\mathbf{y}^l - \mathbf{y}_\theta(\mathbf{x}_t^l, t))^\top \nabla_\theta \mathbf{y}_\theta(\mathbf{x}_t^l, t). \tag{25}$$

Thus,

$$\nabla_\theta S = 2\beta[(y^w - y_\theta(x_t^w, t))^\top \nabla_\theta y_\theta(x_t^w, t) - (y^l - y_\theta(x_t^l, t))^\top \nabla_\theta y_\theta(x_t^l, t)] \tag{26}$$

Step3: Final DPO gradient

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[2\beta \left(1 - \sigma(S)\right) \left((y^w - y_\theta(x_t^w, t))^\top \nabla_\theta y_\theta(x_t^w, t) - (y^l - y_\theta(x_t^l, t))^\top \nabla_\theta y_\theta(x_t^l, t)\right)\right] \tag{27}$$

A.6 Evaluation metrics

For model evaluation, we adopt two assessment methods: GSB and VBench.

GSB is a manual evaluation approach that contrasts the outputs of model A and model B. It employs three labels—$Good$ (A is better), $Same$ (results are equivalent), and Bad (A is worse)—to compute the $Score$, defined as $[Good + Same]/[Same + Bad]$. The $Score$ greater than 1 indicates that A outperforms B, with larger values reflecting more pronounced advantages. Compared to automated metrics, GSB results align more closely with human preferences, offering higher reliability in video generation tasks. We conduct GSB evaluations along two dimensions: $Video\ Quality\ (VQ)$, which assesses overall visual appeal and smoothness, and $Prompt\ Follow\ (PF)$, which measures the consistency between the generated video and the input.

VBench [36] is a widely used automated evaluation tool in the video domain. On our custom test sets, it provides six evaluative dimensions: subject_consistency(Subj.Cons.), background_consistency(Bkgd.Cons.), motion_smoothness(Mot.Smooth.), dynamic_degree(Dyn.Deg.), aesthetic_quality(Aesth.Qual.), and imaging_quality(Img.Qual.). For all metrics, higher values indicate better performance. The subject_consistency evaluates whether the main subjects maintain consistent appearance across frames, reflecting the model's ability to preserve identity and detail. The

background_consistency measures temporal stability of the scene, indicating the model's capability to maintain environmental consistency over time. The motion_smoothness assesses the coherence and naturalness of motion, reflecting the model's grasp of physically plausible movement. The dynamic_degree quantifies the richness of motion, preventing the model from generating static frames and balancing consistency with dynamism. The aesthetic_quality evaluates each frame's composition, color, and overall visual appeal from an artistic perspective, reflecting high-level visual performance. The imaging_quality focuses on clarity and distortion, assessing low-level visual fidelity. In presenting results, we indicate relative changes against the baseline for each dimension as improvement (↑), decline (↓), or unchanged (–). Methods are ranked based on their performance across dimensions, with higher-ranking methods receiving higher scores. Finally, we compute a Total_Score by aggregating the six dimensions, facilitating direct comparison among different methods.

A.7 Training details

In our experiments, we employ two baseline models: Wan2.1-I2V-14B-720P [10] for the I2V task and Wan2.1-T2V-14B [10] for the T2V task. Both are built upon the Flow Matching framework. All experiments are conducted on $8 \times$ NVIDIA H100 (80 GB) GPUs with a per-GPU batch size of 1, using a CP = 4 parallelization strategy. Training utilizes the AdamW optimizer [78] with a learning rate of 5×10^{-6}, and adopts LoRA fine-tuning (rank = 16, α = 32) to improve stability and generation quality. A systematic analysis of key hyperparameters is provided in Appendix A.8.

A.8 Parameter Study

To systematically evaluate the impact of key hyperparameters on model training, we conduct a comprehensive study on Tuning Strategy, LoRA configuration parameters (LoRA Rank and LoRA Alpha), Learning Rate, and DPO preference coefficient (β). Among these, the training strategy has the most significant effect: in video DPO training, LoRA fine-tuning substantially outperforms full-parameter fine-tuning. Specifically, LoRA fine-tuning effectively prevents training collapse, whereas full-parameter fine-tuning often destabilizes the model due to large weight updates, potentially causing severe collapse (e.g., videos with grid artifacts, solid blue frames, or frequent flickering).

Additionally, the choice of LoRA Rank is critical for generation quality: a too-small Rank results in insufficient weight updates, preventing convergence and yielding outputs nearly identical to the baseline, while a too-large Rank can cause over-updating, which may also trigger model collapse.

Table 7 reports experimental results under different hyperparameter settings, with bolded values indicating the optimal configuration.

Table 7: Study of key hyperparameters.

Parameter	Settings
Tuning Strategy	Full_Parameter, **LoRA**
(LoRA Rank, LoRA Alpha)	(4, 8), (8, 16), **(16, 32)**, (32, 64), (64, 128)
Learning Rate	1e-6, **5e-6**, 1.5e-5, 5e-5
β	100, **500**, 1000, 2000

A.9 Analysis of Differences Between T2V and I2V Results

The results for T2V and I2V showed significant differences. In GSB results, Reg-DPO delivers strong gains in both VQ and PF in I2V compared to Base (Tables 3 and 4) but primarily improves VQ in T2V (Table 8). In VBench results, although models optimized using different approaches outperform the Base model across multiple metrics in the I2V setting (Tables 5 and 6), in the T2V setting, the optimized models exhibit a more pronounced polarization pattern (Table 9): static smoothness metrics (Subj.Cons., Bkgd.Cons., Mot.Smooth.) are slightly below Base, whereas dynamic richness metrics (Dyn.Deg., Aesth.Qual., Img.Qual.) improve substantially—especially in Dyn.Deg. (DD). These disparities stem from the inherent generative characteristics of I2V and T2V tasks and the properties of the OpenVid-12K-T2V dataset.

In T2V generation, the model operates solely under textual constraints, granting greater generative freedom. Consequently, the outputs align closely with input instructions and exhibit coherent visuals and smooth motion, causing static-smoothness metrics to approach saturation and leaving limited room for further gains through fine-tuning. However, due to the typically simpler content and smaller motion amplitude in T2V videos, dynamic-richness metrics—particularly DD—remain suboptimal, providing substantial improvement potential. Visual differences between I2V and T2V results are shown in Fig. 11.

The OpenVid-12K-T2V dataset includes positive samples from real videos with rich semantics, diverse scenes, and large motion amplitudes. Preference-based fine-tuning on this dataset markedly enhances realism, reduces AI artifacts, and increases content diversity and motion intensity (see Fig. 12 for example comparisons). Consequently, Reg-DPO yields videos that better match human preferences, achieving notable gains in video quality in GSB and pronounced improvements in dynamic richness metrics in VBench, with DD showing the most significant boost.

Table 8: Experimental results on OpenVid-12K-T2V — GSB

Comparison (A vs B)	VQ	VQ Score	PF	PF Score
Reg-DPO vs Base	297 / 91 / 112	1.91	145 / 273 / 82	1.18
Reg-DPO vs SFT	225 / 155 / 120	1.38	114 / 321 / 65	1.13
Reg-DPO vs DPO	172 / 205 / 123	1.15	181 / 255 / 64	1.37

Table 9: Experimental results on OpenVid-12K-T2V — VBench

Method	Subj.Cons.	Bkgd.Cons.	Mot.Smooth.	Dyn.Deg.	Aesth.Qual.	Img.Qual.	Total_Score
Base	**0.96681** - [4]	**0.95955** - [4]	**0.98919** - [4]	0.34200 - [1]	0.56288 - [1]	0.67584 - [1]	<u>15</u>
SFT	<u>0.96515</u> ↓ [3]	<u>0.95442</u> ↓ [3]	0.98837 ↓ [2]	<u>0.40800</u> ↑ [3]	0.57484 ↑ [2]	0.70004 ↑ [2]	<u>15</u>
DPO	0.95843 ↓ [1]	0.94580 ↓ [1]	0.98831 ↓ [1]	**0.49600** ↑ [4]	**0.58979** ↑ [4]	<u>0.72111</u> ↑ [3]	14
Reg-DPO	0.96211 ↓ [2]	0.94846 ↓ [2]	<u>0.98905</u> ↓ [3]	0.40600 ↑ [2]	<u>0.58651</u> ↑ [3]	**0.72828** ↑ [4]	**16**

A.10   Supplementary Materials
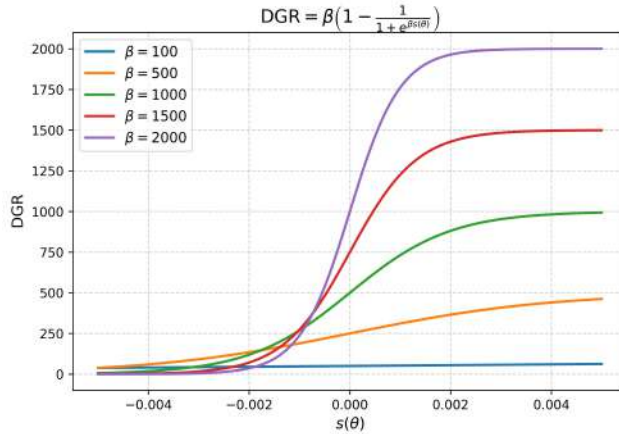
Figure 3: Curve Variation of DGR under Different β Settings. Since DGR is jointly influenced by σ and β, its upper bound is β, and its value drops sharply toward 0 as $s(\theta)$ decreases. A larger β results in a higher initial DGR, but the curve declines more steeply. This functional property of DGR can induce highly volatile gradient changes, making training stability difficult to maintain. In practice, DGR quickly approaches 0, causing the gradient to decay rapidly toward 0, which stalls parameter updates and leads to premature convergence. Moreover, due to the lack of effective constraints on the output distributions of positive and negative samples, such abrupt convergence can easily drive y_θ predictions far from y_{ref}, ultimately causing model collapse.
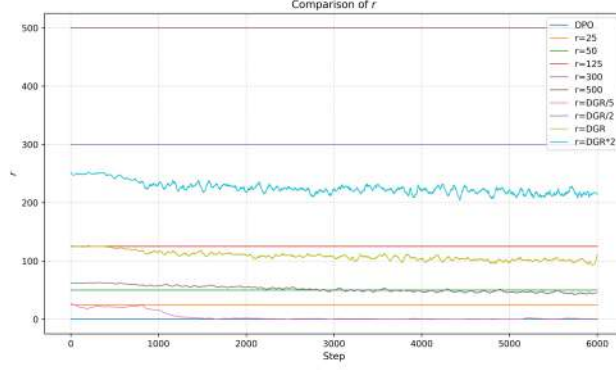
Figure 4: Curves of r across experimental groups. The relative magnitudes of r can be roughly ordered as: $DGR/5(\approx 25 \to 0) < 25 < 50 < DGR/2(\approx 65 \to 40) < DGR(\approx 125 \to 90) < 125 < DGR \times 2(\approx 255 \to 210) < 300 < 500$.
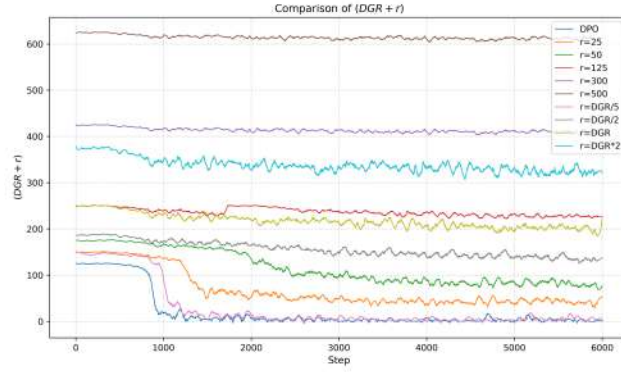

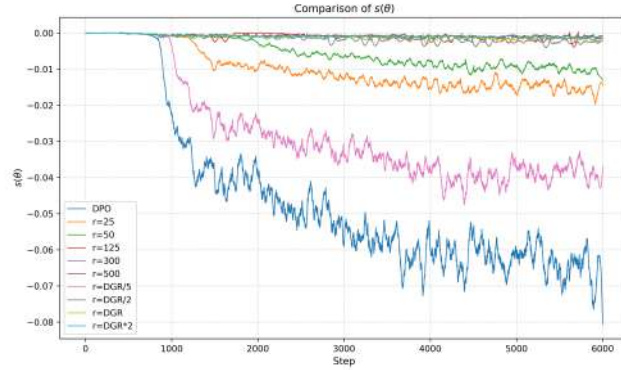Figure 5: Curves of $DGR + r$ across experimental groups.

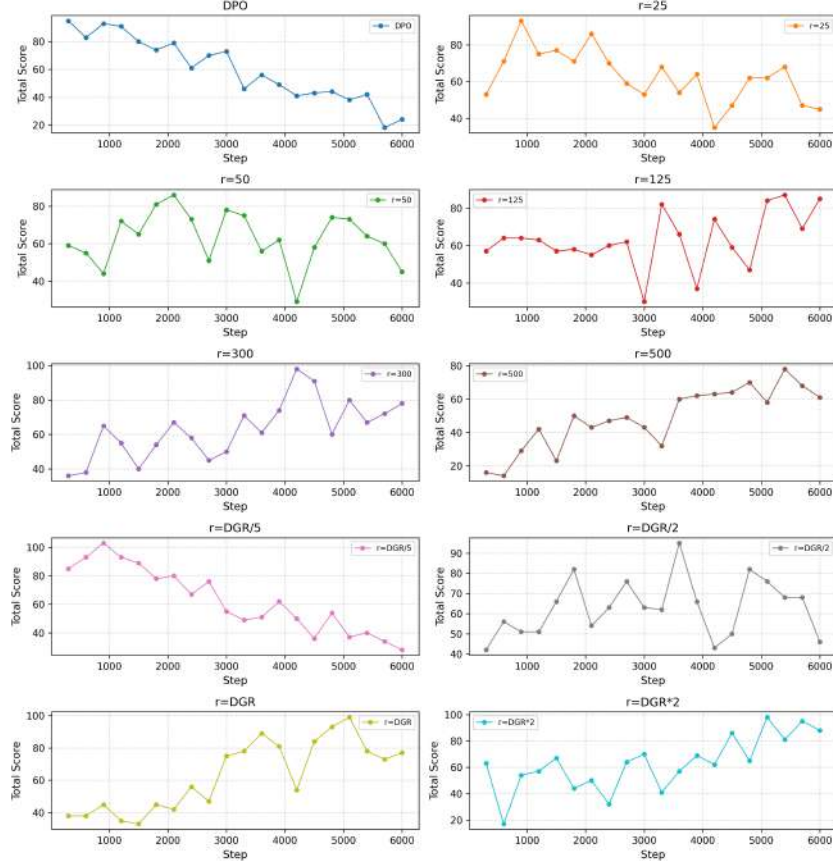Figure 6: Curves of $s(\theta)$ across experimental groups.

Figure 7: Performance curves of models across different experimental groups during training. Model checkpoints are saved every 300 steps, and their VBench results on the test set are evaluated to compute the Total_Score, reflecting overall performance. The Total_Scores of different checkpoints within the same experimental group are plotted as line charts to illustrate performance dynamics throughout training. Based on the training curves shown in Fig. 2, the experimental groups can be categorized into three types: 1.Highly unstable training (severe metric fluctuations): $DPO, r = 25, DGR/5$; 2.Moderately unstable training (noticeable fluctuations): $r = 50, DGR/2$; 3.Stable training (smooth metric variation): $r = 125, 300, 500, DGR, DGR \times 2$. As shown in Fig. 7, the highly unstable group experiences a continuous performance decline throughout training. In the moderately unstable group, as r increases, the SFT regularization term starts to take effect. Although the curves still exhibit some oscillations, they converge significantly better than the previous group, and model performance at different checkpoints no longer deteriorates with training steps, instead showing horizontal fluctuations—indicating a clear stabilizing effect from the SFT regularization. For the stable group, the training curves remain smooth, and model performance steadily improves as training progresses. These results demonstrate that increasing r strengthens the regularization effect of SFT on DPO, effectively mitigating its training instability, enhancing robustness, and leading to consistent performance gains.

Table 10: VBench results of the best checkpoints for each experimental group. As shown in Table 10, when r is small (e.g., $r = 25$ or 50), the model shows minor improvements over Base or DPO in certain metrics but no significant gain in the overall score (Total_Score), and in some cases even a decline. This suggests that under low r, the SFT regularization provides insufficient constraint to stabilize the generative distribution. Training remains dominated by the DPO objective, causing the model to overfit preference signals while neglecting the modeling of the sample distribution, which ultimately degrades overall performance. As r increases, the models exhibit superior performance across more dimensions and achieve higher rankings. The Total_Score statistics clearly show that the best checkpoint performance improves steadily with larger r. When r reaches a threshold ($\sim$ $1 \times DGR$), the model gains a substantial advantage—for instance, at $r = 125$, the total score reaches 44, significantly outperforming pure DPO's 30. This indicates that the SFT regularization effectively constrains the generative distribution, stabilizing training and preventing excessive distributional drift. Beyond this threshold, the performance trends gradually plateau.

Method	Subj.Cons.	Bkgd.Cons.	Mot.Smooth.	Dyn.Deg.	Aesth.Qual.	Img.Qual.	Total_Score
Base	0.93159 - [1]	0.93246 - [1]	0.98540 - [6]	0.68200 - [9]	0.55419 - [2]	**0.72488** - [11]	30
DPO	0.93280 ↑ [2]	0.93289 ↑ [2]	0.98530 ↓ [4]	0.68200 - [9]	0.55600 ↑ [3]	0.72357 ↓ [10]	30
$r = 25$	0.93288 ↑ [3]	0.93410 ↑ [5]	0.98540 - [6]	0.68600 ↑ [10]	0.55221 ↓ [1]	0.71085 ↓ [2]	27
$r = 50$	0.93654 ↑ [5]	**0.93721** ↑ [11]	0.98516 ↓ [3]	0.67400 ↓ [6]	0.55691 ↑ [4]	0.71044 ↓ [1]	30
$r = 125$	0.93749 ↑ [8]	0.93469 ↑ [10]	0.98578 ↑ [8]	0.63200 ↓ [2]	**0.56268** ↑ [11]	0.72019 ↓ [5]	44
$r = 300$	0.93702 ↑ [7]	0.93321 ↑ [3]	**0.98666** ↑ [11]	0.66800 ↓ [4]	0.56089 ↑ [10]	0.71956 ↓ [4]	39
$r = 500$	**0.93902** ↑ [11]	0.93420 ↑ [9]	0.98628 ↑ [10]	0.62200 ↓ [1]	0.55900 ↑ [5]	0.72243 ↓ [9]	**45**
$r = DGR/5$	0.93499 ↑ [4]	0.93416 ↑ [7]	0.98497 ↓ [1]	0.68000 ↓ [7]	0.56035 ↑ [8]	0.72187 ↓ [8]	35
$r = DGR/2$	0.93659 ↑ [6]	0.93416 ↑ [7]	0.98508 ↓ [2]	**0.69800** ↑ [11]	0.56072 ↑ [9]	0.71834 ↓ [3]	38
$r = DGR$	0.93876 ↑ [10]	0.93420 ↑ [9]	0.98560 ↑ [7]	0.67200 ↓ [5]	0.55970 ↑ [6]	0.72054 ↓ [6]	43
$r = DGR \times 2$	0.93848 ↑ [9]	0.93336 ↑ [4]	0.98608 ↑ [9]	0.65800 ↓ [3]	0.56030 ↑ [7]	0.72154 ↓ [7]	39

Figure 8: Comparison of generation results at different training steps between DPO (step = 0, 600, 2000, 6000) and Reg-DPO ($r = DGR$) (step = 0, 2000, 4000, 6000). As shown in Figure 8, DPO produces relatively clear video outputs at the early stage of training (step = 600). However, after severe metric fluctuations (step = 2000; see Fig. 2 for detailed metric trends), the model rapidly collapses, generating videos with mosaic-like grid artifacts. This degradation intensifies as training continues (step = 6000). In contrast, Reg-DPO maintains remarkable training stability: its outputs remain consistently clear across all checkpoints without any sign of collapse, forming a sharp contrast with DPO. These observations demonstrate that the design of Reg-DPO effectively mitigates the instability-induced model collapse encountered in standard DPO training. Prompt instructions and reference image details are provided in Table 11.
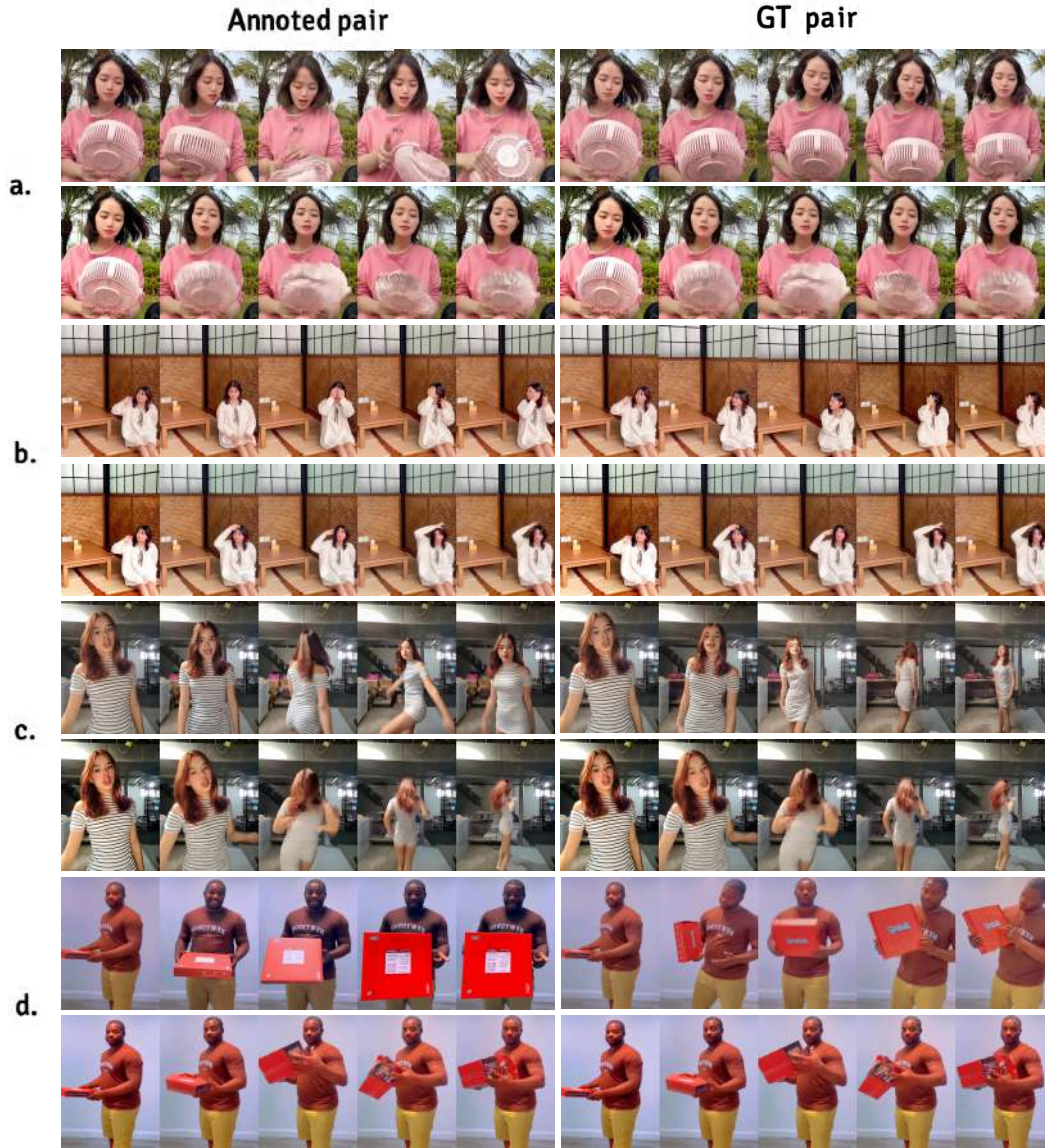
Figure 9: Comparison of Pair Construction Between Manual Annotation and GT-Pair Strategies. In each pair, the upper image represents the positive sample, and the lower image represents the negative sample. As shown in Fig. 9, Pairs constructed using the GT-Pair strategy outperform manually annotated pairs in both positive sample quality and discriminability, with positive samples demonstrating better visual consistency, semantic completeness, and motion smoothness. The main differences in the presented examples are as follows: (a) integrity of the fan; (b) image clarity and number of fingers; (c) completeness of clothing stripes and body-twisting posture; (d) preservation of box shape and plausibility of hand structure.

Figure 10: Case Comparison. The main differences in the presented examples are as follows: (a) stronger content consistency, with no extra umbrella appearing; (b) improved motion dynamics, completing the full action as instructed; (c) clothing patterns remain intact without collapsing; (d) the cup is no longer floating in midair and is correctly placed on the table as instructed; (e) the "camera remains static" instruction is followed, and the hand–football interaction becomes more natural; (f) enhanced color consistency; (g) the pattern on the gift bag no longer flickers. Prompt instructions and reference image details are provided in Table 12.

Table 11: Prompt and reference image information for Fig. 8.

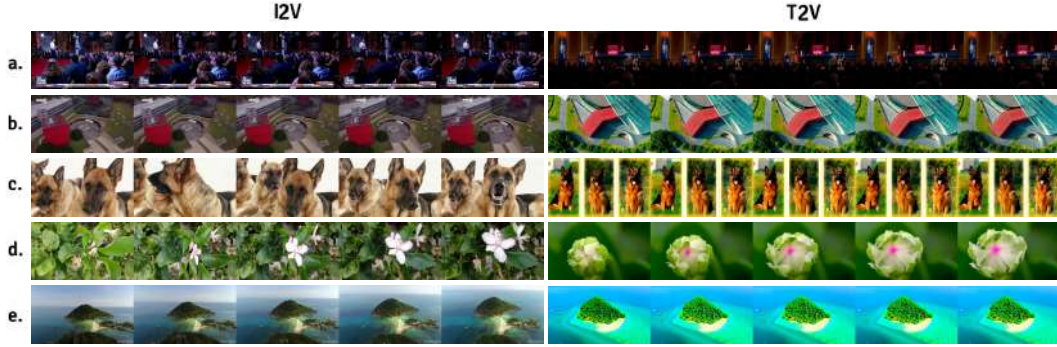| Num	Prompt	Reference Image
a	The video features a man with a beard and short hair, wearing a black t-shirt. He is seated in a room with a white door frame and a window in the background. The man's facial expression changes from a neutral look to a frown, suggesting a shift in his emotions or the conversation. The room appears to be a living area, with a chair and a table visible. The lighting in the room is bright, and the overall atmosphere is casual and relaxed.	
b	The video is a scenic aerial view of a tropical beach. The first frame shows a clear blue sky above the beach, with the ocean stretching out to the horizon. The second frame shows the beach itself, with white sand and palm trees lining the shore. The third frame shows the ocean, with waves crashing against the shore. The style of the video is a combination of aerial photography and time-lapse, capturing the beauty of the beach and the ocean in different times of the day. The video is likely meant to showcase the beauty of the beach and the ocean, and to promote tourism to the area.	

Figure 11: Comparison of generation results between I2V and T2V. I2V results are generated by Wan2.1-I2V-14B-720P, while T2V results are generated by Wan2.1-T2V-14B. As shown in Fig. 11, in the T2V setting, the model operates without reference-image constraints, granting greater generative freedom. Consequently, it often produces temporally coherent videos with visuals that closely align with textual instructions, demonstrating strong instruction-following ability. However, the generated scenes are typically simpler, with smaller motion amplitudes, and may occasionally exhibit static frames. In contrast, in the I2V setting, the presence of reference-image constraints restricts the generative space. Complex reference images and text–image inconsistencies frequently occur, and these factors collectively lead to reduced temporal consistency, causing artifacts such as frame jumps or visual breakdowns. Nonetheless, the resulting videos tend to feature richer, more realistic content, with more diverse and larger-scale motions. Prompt instructions and reference image details are provided in Table 13.

Figure 12: Comparison of model performance before and after Reg-DPO optimization on the OpenVid-12K-T2V dataset. OpenVid-12K-T2V is constructed based on the GT-Pair strategy, where positive samples are sourced from real videos characterized by high realism, diverse content, complex scenes, and large motion amplitudes. After preference optimization, the generated videos exhibit a significant improvement in realism, a reduced sense of artificiality, and a clear tendency toward higher content complexity and dynamic intensity. Prompt instructions are provided in Table 14.

Table 12: Prompt and reference image information for Fig. 10.

Num	Prompt	Reference Image
a	Outdoors, captured in a medium shot, a young woman sits elegantly on a blue structure under soft evening lighting, holding a transparent umbrella with a jellyfish print. She gently rests the umbrella by her shoulder, allowing its unique design and elegant curves to be visible. The atmosphere is calm and whimsical, with a serene nighttime backdrop enhancing the product's charm.	
b	Outdoors, captured in a medium shot, a woman stands confidently in a natural desert-like setting, wearing a casual black graphic t-shirt paired with light blue denim shorts. She casually places one hand lightly on the side of her shorts while the other arm rests naturally, showcasing the relaxed fit and effortless style of the outfit. The camera remains static, ensuring the entire outfit stays fully visible while emphasizing the casual and laid-back vibe. The atmosphere is warm and carefree, with soft natural lighting enhancing the scene.	
c	Indoors, captured in a medium shot, a woman sits comfortably in a woven chair, holding a ceramic mug with a dragonfly design. She wears a black T-shirt with colorful text and graphics that stand out vividly against the neutral background. The camera remains static as she gently lifts the mug to take a sip, showcasing the relaxed fit and casual style of the T-shirt. She then lightly adjusts her posture, allowing the shirt's design to remain fully visible and emphasizing its vibrant colors. The overall atmosphere is cozy and inviting, with soft lighting enhancing the scene's warmth.	
d	Indoors, captured with a medium shot, a cheerful man stands in a brightly lit kitchen, holding a glass of freshly squeezed orange juice in one hand while smiling. The juicer is positioned prominently on the countertop beside a cutting board with sliced oranges, creating a balanced composition. The man gently places the glass on the counter and gestures toward the juicer with a relaxed motion, emphasizing its presence in the scene. The camera remains static throughout, ensuring the juicer and the surrounding elements stay fully visible. The atmosphere is light and refreshing, conveying a sense of ease and satisfaction.	
e	Outdoors, captured in a medium shot, a child stands confidently on a sunny field, wearing a sports jersey and a medal around their neck, holding a soccer ball under one arm. The camera remains steady, fully showcasing the child and the medal in the frame. The child then raises the soccer ball slightly with a proud smile, emphasizing the celebratory and sporty vibe. The scene conveys a joyful and energetic atmosphere, with bright natural lighting enhancing the vibrant colors of the jersey and medal.	
f	Indoors, captured in a medium shot, a woman stands elegantly against a warm orange backdrop, holding a vibrant ombre-patterned handbag in one hand while wearing a navy sleeveless dress and black heels. She slightly tilts her wrist to present the handbag while holding it at her side, highlighting its structured design and bold colors. The overall atmosphere is chic and sophisticated, exuding subtle confidence and timeless style.	
g	Indoors, captured in a medium shot, a young woman sits gracefully on a bench, holding a colorful gift bag with both hands. The vibrant design of the bag stands out against her polka-dot dress, creating a visually appealing contrast. She gently tilts the bag slightly to one side, showcasing its sturdy handles and elegant design. The camera remains static, ensuring the bag and her relaxed posture are fully visible. The atmosphere is light and cheerful, with soft natural lighting enhancing the scene's warmth.	

Table 13: Prompt and reference image information for Fig. 11.

Num	Prompt	Reference Image
a	The video is a news segment from Fox News, focusing on a political event in Des Moines, Iowa. The style of the video is a live broadcast with a panel discussion. The panel consists of a moderator and two guests, all dressed in business attire. The audience is seated in rows, attentively listening to the discussion. The setting is a large hall with a stage, and the atmosphere is formal and serious. The video captures the essence of a political debate, with the panelists engaging in a thoughtful exchange of ideas.	
b	The video is an aerial view of a modern architectural complex. The style is a blend of modern and minimalist design, with a focus on geometric shapes and clean lines. The complex features a large red cube, a glass building with a reflective facade, and a curved driveway leading to a tunnel. The surrounding area includes a parking lot, a grassy field, and a few trees. The video captures the contrast between the man-made structures and the natural environment. The overall impression is one of order and precision, with a sense of tranquility provided by the open space and greenery.	
c	The video shows a German Shepherd dog in three different positions. In the first frame, the dog is sitting with its tongue out, looking to the left. In the second frame, the dog is standing with its tongue out, looking to the right. In the third frame, the dog is sitting with its tongue out, looking straight ahead. The dog has a black and tan coat and is the main subject of the video. The style of the video is a simple, straightforward depiction of the dog in various poses.	
d	The video captures the blooming of a flower in a close-up shot. The flower, with its white petals and pink center, is surrounded by green leaves. The flower opens up in a time-lapse sequence, showcasing the transformation from a closed bud to a fully bloomed flower. The background is blurred, drawing focus to the flower, and the overall style of the video is a close-up, time-lapse shot of a flower blooming.	
e	The video is a drone shot of a tropical island. The island is lush and green, with a sandy beach on one side. The water surrounding the island is a vibrant blue, and there are boats visible in the distance. The sky is clear and blue, suggesting a sunny day. The style of the video is aerial, providing a bird's eye view of the island and its surroundings. The video captures the beauty of the island and its natural environment.	

Table 14: Prompt information for Fig. 12.

Num	Prompt
a	The video depicts a professional setting where two men are engaged in a discussion at a table. The first man, wearing a gray suit and tie, is seated on the left side of the table, while the second man, dressed in a black suit and tie, is seated on the right. They are both holding papers and appear to be reviewing them. The table is white and is set with two cups, suggesting that they might be having a coffee meeting. The background features a building with a balcony, indicating that they might be in an urban environment. The overall style of the video suggests a formal and professional atmosphere.
b	The video captures a bustling city street scene on a cloudy day. The street is lined with tall buildings, some of which have balconies. The buildings are a mix of architectural styles, with one featuring a distinctive tower. The street is busy with pedestrians walking on the sidewalks and cars driving on the road. The traffic lights are visible, indicating a controlled intersection. The overall atmosphere of the video is urban and dynamic, with the city's hustle and bustle captured in the frames.
c	The video captures a dynamic scene on a race track. The main focus is on a blue car, which is seen from the front and side in different frames. The car is in motion, driving down the track with its headlights on. The driver is visible in the side mirror, wearing a black cap and focused on the road ahead. In the background, there are other cars on the track, adding to the sense of competition and speed. The track itself is wide and well-maintained, with clear markings and a smooth surface. The style of the video is realistic and action-oriented, capturing the intensity of the race and the skill of the drivers. The camera angles and movements suggest a sense of speed and movement, while the clear visibility of the cars and track details provide a sense of realism. The overall impression is of a high-stakes racing event, with the blue car as the main protagonist.
d	The video shows a man with a beard and a black cap, sitting in a dark room with a black couch. He is holding a smartphone in his hands, looking at the screen with a surprised expression. The man is wearing a blue t-shirt and appears to be in a relaxed posture. The room has a dimly lit ambiance, with the focus on the man and his smartphone. The style of the video is casual and candid, capturing a moment of surprise or interest in the man's expression.
e	The video features a man with a beard and glasses, wearing a black cap and a blue shirt. He is standing in front of a fence with green foliage in the background. The man appears to be speaking, as his mouth is open. The style of the video is casual and informal, with a focus on the man's facial expression and the surrounding environment. The lighting in the video is natural, suggesting that it was taken during the day. The man's attire and the setting suggest that the video may have been taken in an outdoor or semi-outdoor location.
f	The video shows a close-up view of a car engine, focusing on the various components and parts. The style of the video is informative, providing a detailed look at the inner workings of the engine. The engine is complex, with numerous parts and components visible, including the air filter, spark plugs, and fuel lines. The video captures the intricate design and engineering of the engine, showcasing the precision and craftsmanship involved in its construction. The engine is shown in different angles and perspectives, allowing viewers to fully appreciate its complexity and functionality. The video is likely intended for educational purposes, providing viewers with a comprehensive understanding of how a car engine works.
g	The video is a drone shot of a tropical island. The island is lush and green, with a sandy beach on one side. The water surrounding the island is a vibrant blue, and there are boats visible in the distance. The sky is clear and blue, suggesting a sunny day. The style of the video is aerial, providing a bird's eye view of the island and its surroundings. The video captures the beauty of the island and its natural environment.