

# SAMPLE BY STEP, OPTIMIZE BY CHUNK: CHUNK-LEVEL GRPO FOR TEXT-TO-IMAGE GENERATION

Yifu Luo<sup>1, 2\*†</sup>, Penghui Du<sup>2\*</sup>, Bo Li<sup>2†</sup>, Sinan Du<sup>1,2</sup>, Tiantian Zhang<sup>1</sup>, Yongzhe Chang<sup>1</sup>, Kai Wu<sup>2✉</sup>, Kun Gai<sup>2</sup>, Xueqian Wang<sup>1✉</sup>

<sup>1</sup>Tsinghua University,

<sup>2</sup>Kolors Team, Kuaishou Technology

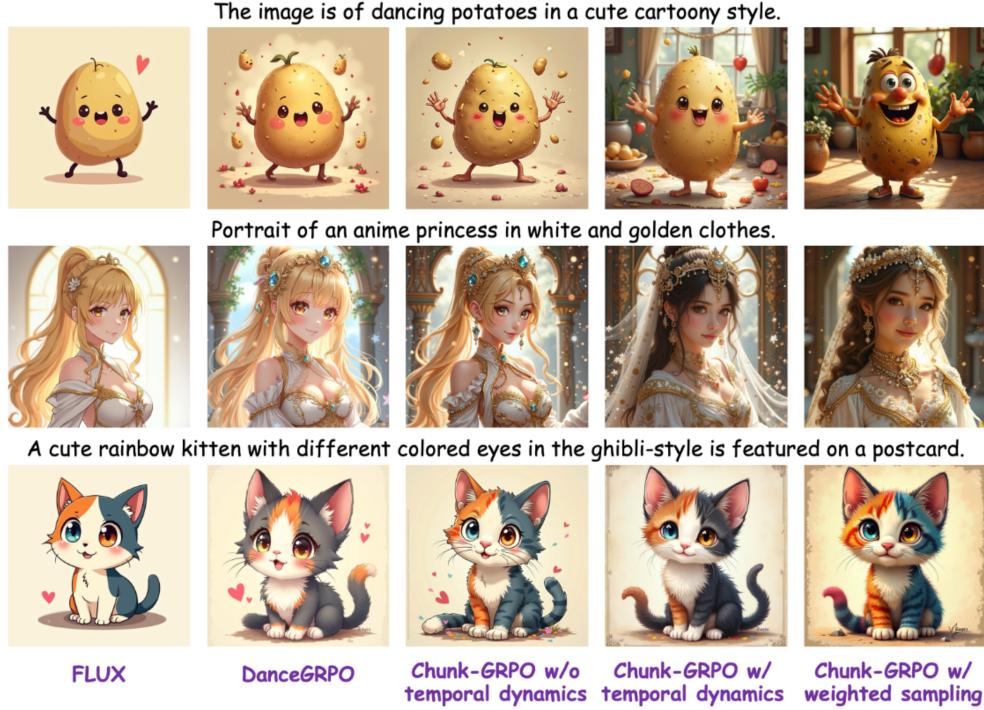


Figure 1: Chunk-GRPO significantly improves image quality, particularly in structure, lighting, and fine-grained details, demonstrating the superiority of chunk-level optimization.

## ABSTRACT

Group Relative Policy Optimization (GRPO) has shown strong potential for flow-matching-based text-to-image (T2I) generation, but it faces two key limitations: inaccurate advantage attribution, and the neglect of temporal dynamics of generation. In this work, we argue that shifting the optimization paradigm from the step level to the chunk level can effectively alleviate these issues. Building on this idea, we propose Chunk-GRPO, the first chunk-level GRPO-based approach for T2I generation. The insight is to group consecutive steps into coherent ‘chunk’s that capture the intrinsic temporal dynamics of flow matching, and to optimize policies at the chunk level. In addition, we introduce an optional weighted sampling strategy to further enhance performance. Extensive experiments show that Chunk-GRPO achieves superior results in both preference alignment and image quality, highlighting the promise of chunk-level optimization for GRPO-based methods.

\* Equal Contribution. † Project Lead. ✉Corresponding Authors.

‡ Work done during internship in Kolors Team, Kuaishou Technology.

## 1 INTRODUCTION

Reinforcement learning (RL)(Sutton et al., 1998; Schulman et al., 2017) has recently found success beyond traditional domains, particularly in the reasoning of Large Language Models (LLMs)(Jaech et al., 2024; Guo et al., 2025). Inspired by these advances, recent works(Xue et al., 2025; Liu et al., 2025b; Wang & Yu, 2025) have explored applying RL to text-to-image (T2I) generation for aligning specific preferences. In this context, Group Relative Policy Optimization (GRPO)(Shao et al., 2024; Guo et al., 2025) has emerged as a promising approach for flow-matching-based T2I generation (Lipman et al., 2022; Liu et al., 2023; Esser et al., 2024). GRPO-based methods typically sample a group of images from the same prompt, evaluate them using reward models, convert the rewards into group relative advantages, and then assign these advantages equally across all timesteps during optimization.

While effective, this uniform assignment introduces two key limitations: (1) inaccurate advantage attribution, and (2) disregard for the temporal dynamics of generation. We first illustrate the former in Figure 2, and discuss temporal dynamics later. Consider two generation trajectories from the same prompt in Figure 2, each consisting of three timesteps. Although the final advantage correctly favors the better trajectory (Trajectory1), assigning this same advantage uniformly across all timesteps incorrectly assumes that every step in Trajectory1 is superior to its counterpart in Trajectory2. However, at timestep  $t = 1$  Trajectory2 is clearly better than Trajectory1, despite Trajectory1 achieving the higher overall reward.

To address this, we draw inspiration from action chunking (Zhao et al., 2023; Li et al., 2025b) in robotics, which predicts sequences of consecutive actions jointly rather than treating each step independently. In a similar spirit, we propose to group consecutive timesteps into ‘chunk’s, and optimize at the chunk level rather than the step level. This alleviates the issue of inaccurate advantage attribution, as we analyze in detail in Section 4.1. Related ideas have been explored in LLMs as Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025), where an entire token sequence is treated as a single unit (analogous to viewing the whole trajectory as one chunk). However, our preliminary studies reveal that different chunk settings (e.g. how many consecutive timesteps for a chunk) have a substantial impact on performance.

We argue that this is due to the overlooking of temporal dynamics of flow matching generation, which we proposed before. Different from LLMs, flow matching exhibits distinct temporal dynamics: each timestep operates under different noise conditions and contributes differently to the final image. Specifically, following (Wimbauer et al., 2024; Liu et al., 2025a) , we analyze the relative  $L1$  distance of intermediate latents. As shown in Figure 3, the results reveal clear, prompt-invariant dynamic patterns that naturally segment the trajectory into meaningful chunks. These observations suggest that chunks should not be arbitrary but guided by the inherent temporal dynamics, with timesteps that are dynamically correlated optimized together.

Based on these, we propose Chunk-GRPO, a novel chunk-level RL approach for flow-matching-based T2I generation. As demonstrated in Figure 4, our key innovation is grouping timesteps into chunks that reflect temporal dynamics, and optimizing them as units with a principled chunk-level importance ratio. Furthermore, motivated by the varying contributions of different chunks, we design an optional weighted sampling strategy to further boost Chunk-GRPO’s performance.

Our contributions can be summarized as follows:

- We are the first to introduce the chunk-level RL optimization for T2I generation. We pinpoint that chunk-level optimization alleviates the inaccurate advantage attribution and mitigates the neglect of temporal dynamics from GRPO-based approaches.

- We propose Chunk-GRPO, a novel chunk-level approach for flow-matching-based T2I generation, which integrates chunk-level optimization with temporal-dynamic-guided chunking. An optional weighted sampling strategy is introduced to push Chunk-GRPO further.
- Extensive experiments demonstrate that Chunk-GRPO achieves superior performance on preference alignment and standard T2I benchmarks, highlighting the effectiveness of chunk-level optimization.

## 2 RELATED WORK

### 2.1 ACTION CHUNK

Action chunk (Zhao et al., 2023; Lai et al., 2022) has been widely applied to robotics Chi et al. (2023). This approach mitigates compounding error and non-Markovian noise in human demonstrations by jointly predicting a sequence of future actions rather than a single step. By shortening the effective control horizon, action chunking enables smoother and more stable rollouts. Recently, it has also proven effective in vision-language-action (VLA) models (Black et al., 2024a; Intelligence et al.) and in RL (Li et al., 2025b). These successes suggest that chunking stabilizes long-horizon prediction, accelerates value propagation, and more effectively leverages non-Markovian behavior.

### 2.2 REINFORCEMENT LEARNING FOR DIFFUSION-BASED IMAGE GENERATION

Diffusion models (Ho et al., 2020; Rombach et al., 2022; Podell et al., 2023; Labs et al., 2025; Wu et al., 2025) have become one of the dominant paradigms for T2I generation. Early works (Xu et al., 2023; Black et al., 2024b; Fan et al., 2023) introduced RL into diffusion models through policy gradient optimization. Preference-based methods (Wallace et al., 2024; Sun et al., 2025a;c;d;e) were later developed, achieving competitive alignment without explicit reward modeling.

More recently, GRPO (Shao et al., 2024; Sun et al., 2025b) has attracted attention as an efficient alternative. Dance-GRPO (Xue et al., 2025) and Flow-GRPO (Liu et al., 2025b) pioneered the use of GRPO for T2I generation, unifying diffusion and flow matching through an SDE-based reformulation. MixGRPO (Li et al., 2025a) further improved efficiency via a mixed ODE–SDE paradigm. TempFlow-GRPO (He et al., 2025) introduced temporal-aware weighting across denoising steps. Pref-GRPO (Wang et al., 2025) identified the issue of illusory advantage and reformulated the optimization objective as pairwise preference fitting. BranchGRPO (Li et al., 2025c) restructured the rollout process into a branching tree, amortizing computation across shared prefixes.

In contrast to these works, our approach explicitly addresses two key issues in GRPO-based T2I generation: (1) inaccurate advantage attribution, and (2) neglect of temporal dynamics. By introducing chunk-level optimization guided by the inherent temporal structure of flow matching, we enhance GRPO from the perspective of optimization granularity.

## 3 PRELIMINARY

### 3.1 FLOW MATCHING

Suppose that  $x_0 \sim \mathbb{X}_0$  is a data sample from the true distribution, and  $x_1 \sim \mathbb{X}_1$  is a noise sample. Following (Liu et al., 2023), the intermediate noised samples  $x_t$  can be expressed as:

$$x_t = (1 - t)x_0 + tx_1, \quad (1)$$

where  $t \in [0, 1]$  denotes the noise level. Then, flow matching aims to directly regress the estimated velocity field  $\hat{v}_\theta(x_t, t)$  by minimizing the objective function (Lipman et al., 2022):

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t, x_0 \sim \mathbb{X}_0, x_1 \sim \mathbb{X}_1} [\|v - \hat{v}_\theta(x_t, t)\|_2^2], \quad (2)$$

where  $v = x_1 - x_0$  represents the target velocity field. Furthermore, a deterministic Ordinary Differential Equation (ODE) is utilized to model the forward process of flow matching:

$$dx_t = \hat{v}_\theta(x_t, t)dt. \quad (3)$$

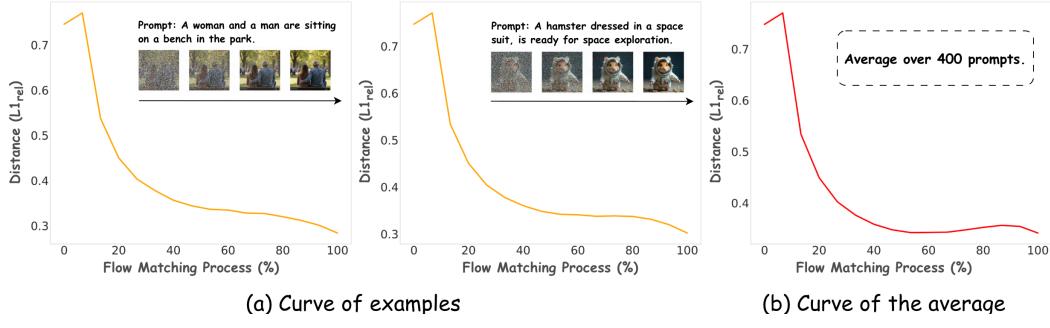


Figure 3: The prompt-invariant temporal dynamics of flow matching.

### 3.2 GRPO ON FLOW MATCHING

As an RL algorithm, GRPO (Guo et al., 2025; Shao et al., 2024) effectively eliminates the need for an additional critic model by estimating the baseline through group-wise relative rewards. In line with the settings of DDPG (Black et al., 2024b), GRPO is also applied in flow matching. Given a group of  $G$  images  $\{x_0^i\}_{i=1}^G$  generated from the same prompt  $c$ , the advantage corresponding to the  $i$ -th sample is formulated as:

$$A_t^i = \frac{r(x_0^i, c) - \text{mean}(\{r(x_0^j, c)\}_{j=1}^G)}{\text{std}(\{r(x_0^j, c)\}_{j=1}^G)}. \quad (4)$$

Notice that  $A_t^i$  always keeps the same for any timestep  $t$ . For simplicity, we neglect the subscript and denote it as  $A^i$ . The policy is updated by maximizing the following GRPO objective:

$$J(\theta) = E_{c, \{x^i\}_{i=1}^G} \left[ \frac{1}{G} \frac{1}{T} \sum_{i=1}^G \sum_{t=1}^T \left( \min(r_t^i(\theta) A^i, \text{clip}(r_t^i(\theta), 1-\epsilon, 1+\epsilon) A^i) - \beta D_{KL}(\pi_\theta || \pi_{ref}) \right) \right], \quad (5)$$

Where  $r_t^i$  denotes the importance ratio:

$$r_t^i(\theta) = \frac{p_\theta(x_{t-1}^i | x_t^i, c)}{p_{\text{old}}(x_{t-1}^i | x_t^i, c)}. \quad (6)$$

Furthermore, to meet the exploration requirement of RL, Flow-GRPO (Liu et al., 2025b) and Dance-GRPO (Xue et al., 2025) introduce stochasticity into flow matching by transforming the deterministic ODE into an equivalent Stochastic Differential Equation (SDE):

$$dx_t = (v_\theta(x_t, t) + \frac{\sigma_t^2}{2t}(x_t + (1-t)v_\theta(x_t, t)))dt + \sigma_t dw_t, \quad (7)$$

where  $dw_t$  represents the increments of the Wiener process and  $\sigma_t$  controls the stochasticity.

## 4 METHOD

In this section, we begin by introducing chunk-level optimization for GRPO and show why it improves upon standard step-level GRPO in Section 4.1. Next, we describe how to set chunks using the inherent temporal dynamics of flow matching in Section 4.2. Finally, we present our proposed Chunk-GRPO along with an optional weighted sampling strategy in Section 4.3.

### 4.1 CHUNK-LEVEL OPTIMIZATION FOR GRPO

Recall the example in Figure 2. With standard step-level GRPO loss in Equation (5), the optimization object for timesteps  $t = 1$  and  $t = 2$  in the two trajectories is:

$$J(\theta) = \frac{1}{G} \frac{1}{T} \sum_{i=1}^G \sum_{t=1}^2 \left( \min(r_t^i(\theta) A^i, \text{clip}(r_t^i(\theta), 1-\epsilon, 1+\epsilon) A^i) - \beta D_{KL}(\pi_\theta || \pi_{ref}) \right). \quad (8)$$

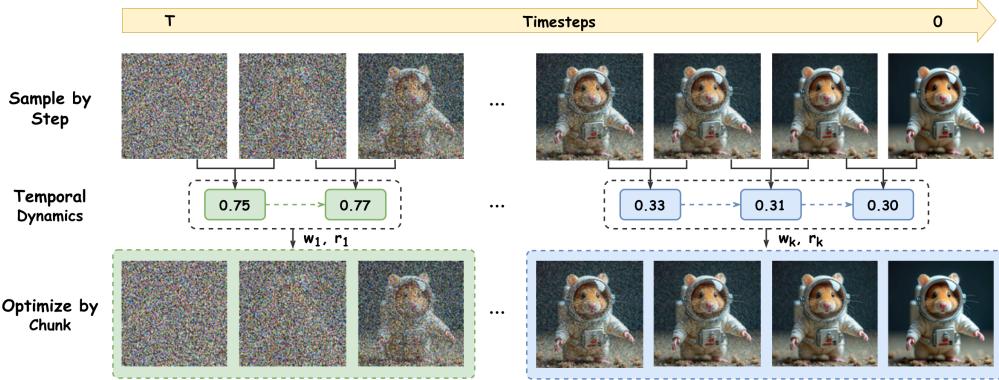


Figure 4: The overall framework of Chunk-GRPO. Chunk-GRPO integrates chunk-level optimization with temporal-dynamic-guided chunking, based on the grounded defined chunk-level importance ratio  $r$ . It also introduces an optional weighted sampling strategy, assigning sampling weight  $w$  for each chunk.

As discussed in Section 1, this uniform stepwise assignment introduces inaccurate advantage attribution. To alleviate this, the first principle of chunk-level optimization is to group consecutive timesteps into chunks and optimize each chunk as a unit. In the example case, the optimization then becomes:

$$J(\theta) = \frac{1}{G} \sum_{i=1}^2 \left( \min(r^i(\theta) A^i, \text{clip}(r^i(\theta), 1 - \epsilon, 1 + \epsilon) A^i) - \beta D_{KL}(\pi_\theta || \pi_{ref}) \right), \quad (9)$$

where the importance ratio is redefined over the chunk likelihood:

$$r^i(\theta) = \left( \prod_{t=1}^2 \frac{p_\theta(x_{t-1}^i | x_t^i, c)}{p_{\text{old}}(x_{t-1}^i | x_t^i, c)} \right)^{\frac{1}{2}}. \quad (10)$$

The key underlying proposition is as follows:

**Proposition 1.** *When advantage attribution is inaccurate at individual timesteps, optimizing them jointly within chunk yields better performance than optimizing them independently as steps, especially when the chunk size is small(e.g. a chunk size of 5).*

A mathematical analysis is provided in Section A. With this insight, we formally define chunk-level optimization for GRPO as follows: Given an image generation trajectory:

$$(x_T, x_{T-1}, \dots, x_2, x_1, x_0)^i, \quad (11)$$

we split it into  $K$  different chunks :

$$\begin{aligned} \{ch_1, ch_2, \dots, ch_K\}^i &= \{(x_T, \dots, x_{T-cs_1+1}), (x_{T-cs_1}, \dots, x_{T-cs_1+cs_2+1}), \dots, (\dots, x_1)\}^i, \\ \sum_{j=1}^k cs_j^i &= T, \end{aligned} \quad (12)$$

where  $cs_j$  denotes the chunk size of the  $j$ -th chunk  $ch_j$ . The chunk-level optimization objective is then:

$$\begin{aligned} J(\theta) &= E_{c, \{x^i\}_{i=1}^G} \\ &\left[ \frac{1}{G} \frac{1}{K} \sum_{i=1}^G \sum_{j=1}^K \left( \min(r_j^i(\theta) A^i, \text{clip}(r_j^i(\theta), 1 - \epsilon, 1 + \epsilon) A^i) - \beta D_{KL}(\pi_\theta || \pi_{ref}) \right) \right], \end{aligned} \quad (13)$$

we neglect  $x_0^i$  because there is no more transition into  $x_{-1}^i$

where we redefine the importance ratio  $r_j^i(\theta)$  based on chunk likelihood:

$$r_j^i(\theta) = \left( \prod_{t \in ch_j} \frac{p_\theta(x_{t-1}^i | x_t^i, c)}{p_{\text{old}}(x_{t-1}^i | x_t^i, c)} \right)^{\frac{1}{cs_j}}. \quad (14)$$

Thus, optimization shifts from step-level to chunk-level, alleviating the issue of inaccurate advantage attribution. Notably, setting  $K = 1$  will group the whole trajectory into a single chunk, and the optimization further shifts to sequence-level similar to GSPO (Zheng et al., 2025). Conversely, setting  $K = T$  will force  $cs_j = 1$ , and the optimization reverts to standard step-level GRPO.

The central question then becomes: given the many possible chunk configurations ( $ch_j$  and  $cs_j$ ), how should chunks be determined?

#### 4.2 CHUNK WITH TEMPORAL DYNAMICS

Before diving into the deeper analysis, we first designed a toy experiment, where all chunks are fixed with an equal chunk size  $cs_1 = cs_2 \dots = cs_k$ . As shown in Figure 5, performance varies with chunk size, indicating that chunk design is non-trivial.

We attribute this to the inherent temporal dynamics of flow matching. Unlike LLMs, flow matching consists of time-dependent dynamics in the generation process, where different timesteps contribute unequally to image quality. To better understand this, following (Wimbauer et al., 2024; Liu et al., 2025a), we illustrate the relative  $L1$  distance  $L1_{\text{rel}}(x, t)$  throughout the generation process:

$$L1_{\text{rel}}(x, t) = \frac{\|x_t - x_{t-1}\|_1}{\|x_t\|_1}. \quad (15)$$

As shown in Figure 3,  $L1_{\text{rel}}(x, t)$  exhibits prompt-invariant yet timestep-dependent patterns. A large  $L1_{\text{rel}}(x, t)$  indicates rapid latent changes, while a small value indicates that adjacent latents are similar to each other. From this observation, we argue that: Timesteps with similar dynamics should be grouped into the same chunk, while timesteps with different dynamics should be separated into different chunks.

Fortunately, the prompt-invariant dynamic patterns of  $L1_{\text{rel}}(x, t)$  naturally segment the trajectory into meaningful chunks, yielding temporal-dynamic-guided chunks. Thus, we can set chunks based on the relative  $L1$  distance, aligning the optimization process with the intrinsic temporal structure of flow matching.

#### 4.3 CHUNK-GRPO

We now present Chunk-GRPO, which integrates chunk-level optimization with temporal-dynamic-guided chunking.

Specifically, given an image generation trajectory, we first compute the relative  $L1$  distance and set chunks like Equation (12) according to the dynamic profile. This yields the chunk numbers  $K$  and chunk sizes  $cs_j$ . The optimization then follows the chunk-level object in Equation (13). The whole framework is shown in Figure 4.

In practice, we observe that the choice of  $K$  and  $cs_j$  is closely tied to the total number of sampling steps  $T$ . A practical strategy, which we adopt in our experiment, is to precompute chunk boundaries based on observed dynamics and keep them fixed throughout training. A detailed discussion is provided in Section 5.1 and Section B.1.

Furthermore, we propose an optional weighted sampling strategy to further enhance Chunk-GRPO. Following Dance-GRPO (Xue et al., 2025), we select only a fraction of chunks (e.g. with fraction 0.5) per update; but instead of uniform sampling, we assign sampling weight  $w$  for each chunk:

$$w(ch_j) = \frac{\frac{1}{cs_j} \sum_{t \in ch_j} L1_{\text{rel}}(x, t)}{\frac{1}{T} \sum_{t=1}^T L1_{\text{rel}}(x, t)}. \quad (16)$$

From Figure 3, this strategy biases the sampling process toward high-noise regions, and the motivation primarily stems from our ablation studies on training specific chunks. However, although this strategy improves certain aspects of Chunk-GRPO, its overall effects on image quality remain nuanced, as discussed in Section 5.3.

## 5 EXPERIMENTS

### 5.1 EXPERIMENT SETUP

**Training Settings** We adopt Dance-GRPO (Xue et al., 2025) as the baseline and conduct experiments with FLUX.1 Dev (Labs, 2024) as our base model. HPDv2.1 (Wu et al., 2023) serves as the dataset, while HPSv3 (Ma et al., 2025) is used as the primary reward model. In ablation studies Section 5.3, we additionally validate our approach with Pick Score (Kirstain et al., 2023) and Clip (Radford et al., 2021) as the reward model. For the chunk setting, we use  $\{cs_j\}_{j=1}^4 = [2, 3, 4, 7]$  with total sampling steps  $T = 17$ , fixed throughout training. Further explanation of chunk configuration and additional training details are provided in Section B.

**Evaluation Details** We evaluate both preference alignment and standard T2I benchmarks. For preference alignment, we use HPSv3 (Ma et al., 2025) and ImageReward (Xu et al., 2023) as in-domain and out-of-domain evaluation metrics, respectively, on the HPDv2.1 (Wu et al., 2023) test set. For the standard T2I benchmark, we report results on WISE (Niu et al., 2025), using its rewritten version due to its improved generalization. We also report results on GenEval (Ghosh et al., 2023) in ablation studies Section 5.3. All evaluations adopt hybrid inference from (Li et al., 2025a), which has proven effective in mitigating reward hacking. More details are provided in Section B.3.

### 5.2 MAIN RESULTS

Table 1 presents results on preference alignment, and Table 2 shows WISE benchmark results. Chunk-GRPO consistently outperforms both the base model and Dance-GRPO, confirming the effectiveness of chunk-level optimization. Qualitative comparisons in Figure 1, Figure 7, and Figure 8 further highlight Chunk-GRPO’s improvements in image quality. Chunk-GRPO generates outputs that align more closely with human aesthetic preferences, exhibiting stronger lighting contrast, more vivid colors, and finer details.

For preference alignment, our approach achieves additional gains of up to 23% over the baseline across both in-domain and out-of-domain evaluations. On WISE, our approach achieves the strongest overall performance. Notably, the weighted sampling strategy enhances preference alignment but has mixed effects on WISE, a phenomenon we further analyze in Section 5.3.

### 5.3 ABLATION STUDY

**Chunk Setting.** We first vary chunk settings under different total sampling steps, excluding the weighted sampling strategy to isolate chunk setting effects. Results in Table 3 show that chunk-level optimization consistently outperforms standard step-level GRPO.

We neglect the last timestep following Dance-GRPO, as the last step does not introduce stochasticity.

Table 2: Results on WISE

Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
Flux	0.75	0.70	0.76	0.69	0.71	0.68	0.73
Dance-GRPO	0.82	0.75	0.78	0.66	0.69	0.64	0.75
Chunk-GRPO w/o ws	0.82	0.76	0.77	0.68	0.69	0.68	0.76
Chunk-GRPO w/ ws	0.80	0.73	0.76	0.64	0.65	0.62	0.73

<sup>1</sup> The ‘ws’ refers to the weighted sampling strategy.

<sup>2</sup> We use the rewritten version of WISE.

Table 3: Ablation Results of Chunk Setting

Model	Sampling Timesteps	Chunk Setting	HPSv3
Flux	-	-	13.804
DanceGRPO	17	-	15.080
	25	-	15.015
Chunk-GRPO w/o td	17	[2, 2, ..., 2]	15.115
	17	[4, 4, 4, 4]	15.078
	17	[8, 8]	15.173
	17	[16]	15.142
	25	[2, 2, ..., 2]	15.057
	25	[4, 4, ..., 4]	15.136
	25	[12, 12]	15.111
	25	[24]	15.100
	17	[2, 3, 4, 7]	15.236
Chunk-GRPO w/ td	17	[2, 3, 4, 7]	15.236

<sup>1</sup> The ‘td’ refers to the temporal dynamics.

Moreover, temporal-dynamics-guided chunking outperforms that of fixed chunk size, underscoring the importance of aligning the optimization process with the intrinsic temporal structure of flow matching.

**Training on Specific Chunks.** We next train Chunk-GRPO on individual chunks only. Note that we also remove the weighted sampling strategy here. Results in Figure 6 show that high-noise chunks (e.g.,  $ch_1$ ) yield larger improvements than low-noise chunks (e.g.,  $ch_4$ ), but also suffer from training instability (e.g. after 60 steps). This observation motivated our weighted sampling strategy in Equation (16), which adaptively emphasizes high-noise chunks.

**Weighted Sampling Strategy.** As shown in Table 1 and Table 2, the optional weighted sampling strategy improves preference alignment but slightly reduces WISE performance. Careful qualitative analysis reveals a trade-off: while the strategy accelerates preference optimization, it can destabilize image structure in high-noise regions, occasionally leading to semantic collapse. A failure example is shown in Figure 9. Although all methods struggle with this challenging prompt (e.g. Dance-GRPO misses the attribute ‘sleeveless’), the weighted sampling strategy further alters the overall image structure, producing the worst case by omitting the entire item ‘black loafers’ and only partially showing ‘capris’). This demonstrates the complex effects of the strategy.

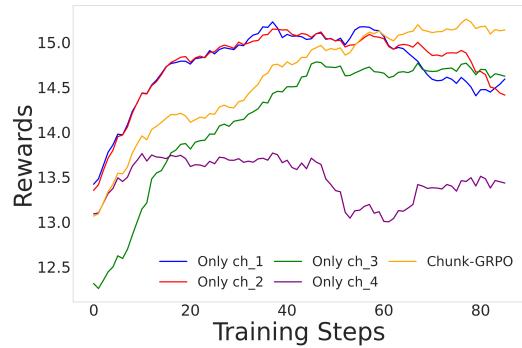


Figure 6: The results of training specific chunks.

2B from NieR Automata eating a bagel.



A girl with pink pigtails and face tattoos.



16-year-old teenager wearing a white bear-ear hat with a smirk on their face.



"The image is a Roy Lichtenstein emoji portraying a woman with dark brown pixie hair, entirely black eyes, wearing a black tank top, leather jacket, skirt, choker, and boots."



Figure 7: Additional visualization comparison between the FLUX, DanceGRPO, Chunk-GRPO w/o temporal dynamics, Chunk-GRPO w/ temporal dynamics and Chunk-GRPO w/ weighted sampling.

A corgi puppy with many eyes depicted in a horror manga drawn by Junji Ito.



A young woman witch cosplaying with a magic wand and broom, wearing boots, and posing in a full body shot with a detailed face.



The image depicts a stunning supernova within a fantasy artwork on Artstation.



A train is moving along the track in the countryside.



Figure 8: Additional visualization comparison between the FLUX, DanceGRPO, Chunk-GRPO w/o temporal dynamics, Chunk-GRPO w/ temporal dynamics and Chunk-GRPO w/ weighted sampling.

A painting of a woman by Zinaida Serebriakova wearing a T-shirt with the Supreme brand logo, a sleeveless white blouse, dark brown capris, and black loafers.



Figure 9: A failure case of the weighted sampling strategy. The strategy wrongly changes the image structure in the high-noise region, leading to the worst variant.

**Reward Models.** Finally, we test Chunk-GRPO’s robustness under different reward models. We first replace Hpsv3 with Pick Score (Shukor et al., 2025) as our reward model. Results in Table 4 confirm that Chunk-GRPO consistently outperforms standard step-level GRPO regardless of the reward model, validating its generality.

Since both HPSv3 and PickScore are reward models primarily designed for preference alignment, we further validate our approach using Clip (Radford et al., 2021), which, while not a preference alignment model, is well recognized for its ability to capture high-level semantics. We evaluate this on GenEval (Ghosh et al., 2023), a benchmark that mainly tests instruction-following capability. Results in Table 5 demonstrate that Chunk-GRPO also outperforms standard step-level GRPO, demonstrating its broader generalization and robustness beyond preference alignment tasks. It is worth noting that the weighted sampling strategy results in a decline in GenEval’s semantic performance, which further corroborates our previous analysis.

Table 4: Ablation on Different Reward Models

Model	Pick Score	HPSv3	Image Reward
Flux	22.643	13.804	1.086
DanceGRPO	23.427	14.612	1.208
Chunk-GRPO w/o ws	23.442	14.810	1.222
Chunk-GRPO w/ ws	23.476	14.913	1.233

<sup>1</sup> The ‘ws’ refers to the weighted sampling strategy.

Table 5: Results on GenEval

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
Flux	0.99	0.83	0.71	0.75	0.24	0.44	0.66
Dance-GRPO	1.00	0.86	0.71	0.78	0.22	0.46	0.67
Chunk-GRPO w/o ws	0.99	0.85	0.75	0.81	0.21	0.51	0.69
Chunk-GRPO w/ ws	0.98	0.82	0.73	0.76	0.27	0.48	0.67

<sup>1</sup> The ‘ws’ refers to the weighted sampling strategy.

## 6 CONCLUSION

In this paper, we propose Chunk-GRPO, the first chunk-level GRPO-based approach for flow-matching-based T2I generation. By leveraging the temporal dynamics of flow matching, Chunk-GRPO groups consecutive timesteps into chunks and optimizes at the chunk level, achieving consistent improvements over standard step-level GRPO. We further introduce an optional weighted sampling strategy to push Chunk-GRPO further.

Despite its strong performance, several limitations remain. First, exploring how to combine heterogeneous rewards across different chunks (e.g., employing different reward models for high- vs. low-noise regions) could unlock further improvements. Second, our chunk segmentation is fixed throughout training. Developing self-adaptive or dynamic chunking strategies that adjust to training signals would be an important next step.

## ACKNOWLEDGMENTS

This work was supported in part by the Natural Science Foundation of Shenzhen (No. JCYJ20230807111604008, No. JCYJ20240813112007010), the Natural Science Foundation of Guangdong Province (No. 2024A1515010003), National Key Research and Development Program of China (No. 2022YFB4701400) and Cross-disciplinary Fund for Research and Innovation of Tsinghua SIGS (No. JC2024002).

## REFERENCES

- Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. *pi\_0: A vision-language-action flow model for general robot control*. *arXiv preprint arXiv:2410.24164*, 2024a.
- Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024b.
- Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al.  $\pi_0.5$ : a vision-language-action model with open-world generalization, 2025. URL <https://arxiv.org/abs/2504.16054>, 1(2):3.
- 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.
- 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.
- Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.

- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL <https://arxiv.org/abs/2506.15742>.
- Lucy Lai, Ann ZX Huang, and Samuel J Gershman. Action chunking as conditional policy compression. 2022.
- Junzhe Li, Yutao Cui, Tao Huang, Yiping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025a.
- Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking. *arXiv preprint arXiv:2507.07969*, 2025b.
- Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*, 2025c.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7353–7363, 2025a.
- Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025b.
- Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.
- Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025.
- Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.
- Haoyuan Sun, Bin Liang, Bo Xia, Jiaqi Wu, Yifei Zhao, Kai Qin, Yongzhe Chang, and Xueqian Wang. Diffusion-rainbowpa: Improvements integrated preference alignment for diffusion-based text-to-image generation. *Transactions on Machine Learning Research*, 2025a.
- Haoyuan Sun, Jiaqi Wu, Bo Xia, Yifei Luo, Yifei Zhao, Kai Qin, Xufei Lv, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement fine-tuning powers reasoning capability of multimodal large language models. *arXiv preprint arXiv:2505.18536*, 2025b.
- Haoyuan Sun, Bo Xia, Yongzhe Chang, and Xueqian Wang. Generalizing alignment paradigm of text-to-image generation with preferences through f-divergence minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27644–27652, 2025c.
- Haoyuan Sun, Bo Xia, Yifei Zhao, Yongzhe Chang, and Xueqian Wang. Identical human preference alignment paradigm for text-to-image models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025d.
- Haoyuan Sun, Bo Xia, Yifei Zhao, Yongzhe Chang, and Xueqian Wang. Positive enhanced preference alignment for text-to-image models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025e.
- Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.
- 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*, pp. 8228–8238, 2024.
- Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. *arXiv preprint arXiv:2509.05952*, 2025.
- Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025.
- Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, et al. Cache me if you can: Accelerating diffusion models through block caching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6211–6220, 2024.
- Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.
- 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.
- 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.
- Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

## A MATHEMATICAL ANALYSIS

Here we provide a mathematical analysis for Proposition 1. For simplicity, we assume that there are  $m$  timesteps with inaccurate advantage attribution between two trajectory segments:

$$(x_T, x_{T-1}, \dots, x_2, x_1, x_0)^1, \\ (x_T, x_{T-1}, \dots, x_2, x_1, x_0)^2, \quad (17)$$

where  $1 \leq m \leq T$ . We denote  $T_a$  and  $T_{ia}$  as the sets of timesteps with accurate and inaccurate advantage attribution, respectively, and:

$$T_a \cap T_{ia} = \emptyset, \quad T_a \cup T_{ia} = \{1, 2, \dots, T\}. \quad (18)$$

Let  $A^i$  and  $A^j$  as the advantage of the two trajectories. Without loss of generality, we assume:

$$A^1 = 1, \quad A^2 = -1. \quad (19)$$

We denote  $\hat{A}_t^i$  as the ground-truth advantage. Then for each timestep  $t$ :

$$\begin{aligned} \hat{A}_t^1 &= A_t^1 = 1, & \hat{A}_t^2 &= A_t^2 = -1, & t \in T_a, \\ \hat{A}_t^1 &= -A_t^1 = -1, & \hat{A}_t^2 &= -A_t^2 = 1, & t \in T_{ia}. \end{aligned} \quad (20)$$

The expected ground-truth loss object can thus be expressed as:

$$J(\hat{\theta}) = \sum_{i=1}^2 \sum_{t=1}^T \min \left( r_t^i(\theta) \hat{A}^i, \text{clip} (r_t^i(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}^i \right). \quad (21)$$

Here we omit constant factors such as  $\frac{1}{2}$ ,  $\frac{1}{T}$ , and KL regularization. The step-level importance ratio  $r_t^i$  is defined in Equation (6), reproduced here for clarity::

$$r_t^i(\theta) = \frac{p_\theta(x_{t-1}^i | x_t^i, c)}{p_{old}(x_{t-1}^i | x_t^i, c)}. \quad (22)$$

Substituting Equation (20) into eq. (21), we obtain:

$$\begin{aligned} J(\hat{\theta}) &= \sum_{t \in T_a} [\min (r_t^1(\theta), \text{clip} (r_t^1(\theta), 1 - \epsilon, 1 + \epsilon)) + \min (-r_t^2(\theta), -\text{clip} (r_t^2(\theta), 1 - \epsilon, 1 + \epsilon))] \\ &\quad + \sum_{t \in T_{ia}} [\min (r_t^2(\theta), \text{clip} (r_t^2(\theta), 1 - \epsilon, 1 + \epsilon)) + \min (-r_t^1(\theta), -\text{clip} (r_t^1(\theta), 1 - \epsilon, 1 + \epsilon))]. \end{aligned} \quad (23)$$

Since the clipping operation only affects timesteps where the importance ratio lies outside the trust region (Schulman et al., 2015), and such cases are rare under small policy updates, we approximate the gradient of Equation (23) by the gradient of following expression:

$$J(\hat{\theta}) = \sum_{t \in T_a} (r_t^1(\theta) - r_t^2(\theta)) + \sum_{t \in T_{ia}} (r_t^2(\theta) - r_t^1(\theta)). \quad (24)$$

Similarly, the step-level GRPO loss has gradient approximated to the gradient of following:

---

we neglect  $x_0$  because there is no more transition into  $x_{-1}$

$$J(\theta)_{GRPO} = \sum_{t \in T_a} (r_t^1(\theta) - r_t^2(\theta)) + \sum_{t \in T_{ia}} (r_t^1(\theta) - r_t^2(\theta)). \quad (25)$$

We now analyze chunk-level optimization. For simplicity, we treat each trajectory in Equation (17) as a single chunk. Following Equation (12), we have:

$$\begin{aligned} \{ch_1\}^i &= \{(x_T, \dots, x_1)\}^i, \quad i = 1, 2, \\ cs_1^1 &= cs_1^2 = T, \end{aligned} \quad (26)$$

The reason is that if trajectories are split into smaller chunks, each chunk can be viewed as a complete trajectory as in Equation (17). For convenience, we rewrite the chunk-level importance ratio from Equation (14) as:

$$s_j^i(\theta) = \left( \prod_{t \in ch_j} \frac{p_\theta(x_{t-1}^i | x_t^i, c)}{p_{\theta_{old}}(x_{t-1}^i | x_t^i, c)} \right)^{\frac{1}{cs_j}}. \quad (27)$$

The chunk-level objective then becomes:

$$J(\theta)_{chunk} = \sum_{i=1}^2 \min(s_1^i(\theta) A^i, \text{clip}(s_1^i(\theta), 1 - \epsilon, 1 + \epsilon) A^i). \quad (28)$$

Similarly, the gradient of  $J(\theta)_{chunk}$  can be approximated by the gradient of following expression::

$$J(\theta)_{chunk} = s_1^1 - s_1^2, \quad (29)$$

where

$$\begin{aligned} s_1^i(\theta) &= \left( \prod_{t \in ch_1} \frac{p_\theta(x_{t-1}^i | x_t^i, c)}{p_{\theta_{old}}(x_{t-1}^i | x_t^i, c)} \right)^{\frac{1}{cs_1}} \\ &= \left( \prod_{t=1}^T \frac{p_\theta(x_{t-1}^i | x_t^i, c)}{p_{\theta_{old}}(x_{t-1}^i | x_t^i, c)} \right)^{\frac{1}{T}} \\ &= \left( \prod_{t=1}^T r_t^i(\theta) \right)^{\frac{1}{T}}, \quad i = 1, 2. \end{aligned} \quad (30)$$

In Proximal Policy Optimization (PPO) (Schulman et al., 2017) and GRPO-based methods, the importance ratio  $r_t^i(\theta)$  remains close to 1 due to trust-region constraints Schulman et al. (2015; 2017). We therefore set:

$$r_t^i(\theta) = 1 + \epsilon_t^i, \quad (31)$$

where  $\epsilon_t^i$  is a minimal term. Substituting into Equation (24) and Equation (25):

$$\hat{J}(\theta) = \sum_{t \in T_a} (\epsilon_t^1 - \epsilon_t^2) + \sum_{t \in T_{ia}} (\epsilon_t^2 - \epsilon_t^1) \quad (32)$$

$$\begin{aligned} J(\theta)_{GRPO} &= \sum_{t \in T_a} (\epsilon_t^1 - \epsilon_t^2) + \sum_{t \in T_{ia}} (\epsilon_t^1 - \epsilon_t^2) \\ &= \sum_{t=1}^T (\epsilon_t^1 - \epsilon_t^2). \end{aligned} \quad (33)$$

For the chunk-level ratio in Equation (30), applying the logarithm and Taylor expansion gives:

$$\begin{aligned}
s_1^i(\theta) &= \left( \prod_{t=1}^T r_t^i(\theta) \right)^{\frac{1}{T}} \\
&= \left( \prod_{t=1}^T (1 + \epsilon_t^i) \right)^{\frac{1}{T}} \\
&= 1 + \frac{1}{T} \sum_1^T \epsilon_t^i.
\end{aligned} \tag{34}$$

Thus the chunk-level objective reduces to:

$$\begin{aligned}
J(\theta)_{chunk} &= s_1^1 - s_1^2 \\
&= \left( 1 + \frac{1}{T} \sum_1^T \epsilon_t^1 \right) - \left( 1 + \frac{1}{T} \sum_1^T \epsilon_t^2 \right) \\
&= \frac{1}{T} \sum_{t=1}^T (\epsilon_t^1 - \epsilon_t^2) \\
&= \frac{1}{T} J(\theta)_{GRPO}.
\end{aligned} \tag{35}$$

This shows that chunk-level optimization yields a smoothed version of the step-level GRPO objective. More formally, by comparing the squared distances between coefficient vector of  $J(\hat{\theta})$ ,  $J(\theta)_{GRPO}$ , and  $J(\theta)_{chunk}$ , we find:

$$\begin{aligned}
\|J(\hat{\theta}) - J(\theta)_{GRPO}\|_2^2 &= 2m \times (1 - (-1))^2 \\
&= 8m.
\end{aligned} \tag{36}$$

$$\begin{aligned}
\|J(\hat{\theta}) - J(\theta)_{chunk}\|_2^2 &= \|J(\hat{\theta}) - \frac{1}{T} J(\theta)_{GRPO}\|_2^2 \\
&= \|J(\hat{\theta})\|^2 + \frac{1}{T^2} \|J(\theta)_{GRPO}\|^2 - \frac{2}{T} J(\hat{\theta}) \cdot J(\theta)_{GRPO} \\
&= 2T + \frac{2T}{T^2} - \frac{2}{T} \cdot 2(T - 2m) \\
&= 2T - 4 + \frac{8m + 2}{T},
\end{aligned} \tag{37}$$

Where  $m$  denotes the number of inaccurately attributed timesteps, which we mentioned in the beginning of this section. We want Equation (37) to be smaller than Equation (36), i.e.,

$$\|J(\hat{\theta}) - J(\theta)_{GRPO}\|_2^2 - \|J(\hat{\theta}) - J(\theta)_{chunk}\|_2^2 \geq 0 \tag{38}$$

Solving yields:

$$\begin{aligned}
&\|J(\hat{\theta}) - J(\theta)_{GRPO}\|_2^2 - \|J(\hat{\theta}) - J(\theta)_{chunk}\|_2^2 \geq 0 \\
\Leftrightarrow &8m - 2T + 4 - \frac{8m + 2}{T} \geq 0 \\
\Leftrightarrow &2T^2 - (4m + 8)T + (8m + 2) \leq 0 \\
\Leftrightarrow &T^2 - (2m + 4)T + (4m + 1) \leq 0 \\
\Leftrightarrow &m - \sqrt{m^2 + 3} + 2 \leq T \leq m + \sqrt{m^2 + 3} + 2
\end{aligned} \tag{39}$$

Since  $1 \leq m \leq T$ , the first inequality always holds. As both  $T$  and  $m$  are positive integers, we obtain:

$$T \begin{cases} \leq 5, & \text{if } m = 1 \\ \leq 2m + 2, & \text{if } m \geq 2. \end{cases} \quad (40)$$

Note that here  $cs_1 = T$ , and the whole trajectory is treated as a single chunk. When the chunk size  $cs \leq 5$ , Equation (38) always holds, meaning that the chunk-level objective  $J(\theta)_{chunk}$  is closer to the ground-truth object  $J(\hat{\theta})$  than  $J(\theta)_{GRPO}$ . For larger chunks, Equation (38) still holds when  $m \leq \frac{T-2}{2}$ .

The insights of this solution are:

- For small chunks (e.g.  $cs_j = 5$ ), chunk-level optimization always outperforms step-level GRPO.
- For large chunk sizes, it also holds when roughly half of the timesteps suffer from inaccurate advantage attribution.
- From Equation (35), chunk-level optimization consistently provides smoother gradients than step-level GRPO..

## B EXPERIMENT DETAILS

### B.1 CHUNK CONFIGURATION

In practice, the default Chunk-GRPO segments the image generation trajectory into  $K = 4$  chunks with  $cs_{j=1}^4 = 2, 3, 4, 7$  under  $T = 17$  timesteps. The rationale is as follows:

- Following Figure 3, We set the first chunk as  $cs_1 = 2$ .
- For the last chunk, we first conduct a pre-observation: we compute the relative  $L1$  distance in Equation (15) again, but with a Dance-GRPO-trained model instead of the base model. As shown in Figure 10, RL alters the relative  $L1$  distance primarily in the latter half of timesteps. Based on this, we set  $ch_4 = 7$ .
- For  $ch_2 = 3$  and  $ch_3 = 4$ , we base the segmentation on the second derivative of the  $L1$  curve.
- This configuration also satisfies the requirement in Proposition 1, which recommends keeping chunk size small (e.g. 5).

We emphasize that this segmentation is not guaranteed to be the only optimal choice. Exploring adaptive chunk configurations under different  $T$  is an interesting direction for future work.

### B.2 TRAINING DETAILS

All experiments were conducted on 8 Nvidia H800 GPUs. The hyperparameters are summarized in table 6.

### B.3 EVALUATION DETAILS

We set  $T = 50$  during evaluation. Following (Li et al., 2025a), the first 30 steps are sampled with the trained model, while the remaining 20 steps are sampled with the base model. This hybrid inference strategy and corresponding settings, also used in (Li et al., 2025a), have proven effective in mitigating reward hacking.

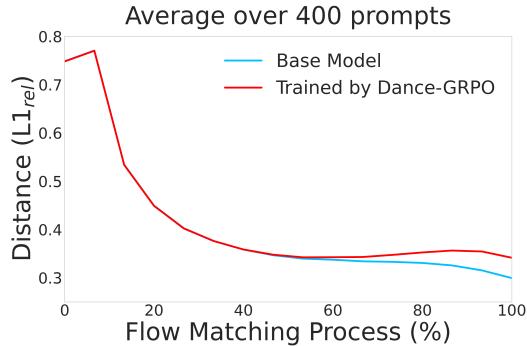


Figure 10: The relative  $L1$  distance comparison, before and after the training of Dance-GRPO.

We neglect the last timestep following Dance-GRPO, as the last step does not introduce stochasticity.

Table 6: Hyperparameter Settings

Parameter	Value	Parameter	Value
Learning rate	$1 \times 10^{-5}$	Weight decay	$1 \times 10^{-4}$
Train batch size	2	SP size	1
SP batch size	2	Max grad norm	0.01
Resolution	$720 \times 720$	Sampling steps	17
Eta	0.7	Num. generations	12
Grad. accum. steps	12	Shift (branch offset)	3
Clip range	$5 \times 10^{-5}$	Training steps	150
Timestep fraction	0.5		