From Competition to Synergy: Unlocking Reinforcement Learning for Subject-Driven Image Generation

Ziwei Huang[1], Yin Shu[2], Hao Fang[2], Quanyu Long[3], Wenya Wang[3], Qiushi Guo[2], Tiezheng Ge[2], Leilei Gan[1]*,

[1]Zhejiang University, [2]Alibaba Group, [3]Nanyang Technological University,
{ziweihuang, leileigan}@zju.edu.cn

Abstract

Subject-driven image generation models face a fundamental trade-off between identity preservation (fidelity) and prompt adherence (editability). While online reinforcement learning (RL), specifically GRPO, offers a promising solution, we find that a naive application of GRPO leads to competitive degradation, as the simple linear aggregation of rewards with static weights causes conflicting gradient signals and a misalignment with the temporal dynamics of the diffusion process. To overcome these limitations, we propose Customized-GRPO, a novel framework featuring two key innovations: (i) Synergy-Aware Reward Shaping (SARS), a non-linear mechanism that explicitly penalizes conflicted reward signals and amplifies synergistic ones, providing a sharper and more decisive gradient. (ii) Time-Aware Dynamic Weighting (TDW), which aligns the optimization pressure with the model's temporal dynamics by prioritizing prompt-following in the early, identity preservation in the later. Extensive experiments demonstrate that our method significantly outperforms naive GRPO baselines, successfully mitigating competitive degradation. Our model achieves a superior balance, generating images that both preserve key identity features and accurately adhere to complex textual prompts.[1]

1 Introduction

In recent years, subject driven image generation, which aims to create customized images that align with both a textual prompt and the specific subjects in reference images, has garnered substantial interest across both academic and industrial communities (Li et al., 2023; Ruiz et al., 2023; Gal et al., 2022; Ye et al., 2023; Zhang et al., 2024; Huang et al., 2025a; Wu et al., 2025). Unlike standard text-to-image (T2I) generation (Rombach et al., 2022; Podell et al., 2024; He et al., 2024), which focuses solely on text–image alignment, this task demands a dual capability: maintaining high-fidelity identity preservation from reference images and ensuring accurate prompt adherence in novel contexts. This dual objective introduces a critical generalization challenge: the model must learn a subject representation that is robust enough to preserve its core identity, yet adaptive enough to integrate seamlessly into new contexts as indicated in the prompt. Achieving this balance between identity fidelity and prompt adherence is central to realizing practical personalized generation.

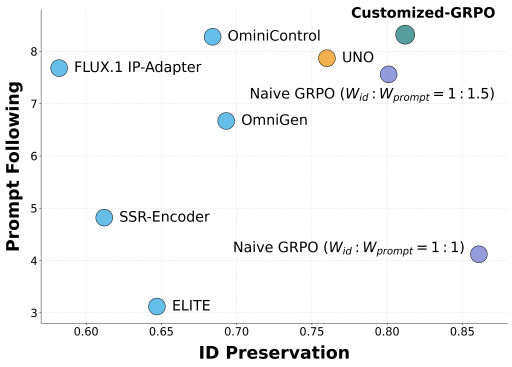

Figure 1: **Customized-GRPO** achieves a state-of-the-art balance between ID Preservation and Prompt Following on DreamBench. Our method and Naive GRPO are built upon UNO.

To address this challenge, current research has primarily evolved along two main paradigms: finetuning-based and adapter-based customization. Finetuning-based approaches, such as DreamBooth (Ruiz et al., 2023) and Textual Inversion (Gal et al., 2022), enable subject-specific customization by fine-tuning model parameters using a small set of reference images to learn the specialized representation of single subject. However, these methods are computationally intensive and requires retraining for each new subject, making it impractical for scalable customization. Instead, Adapter-based approaches (Wu et al., 2025; Ye et al., 2023;

*Corresponding author
[1] github.com/Safeoffellow/Customized-GRPO

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41117–41136

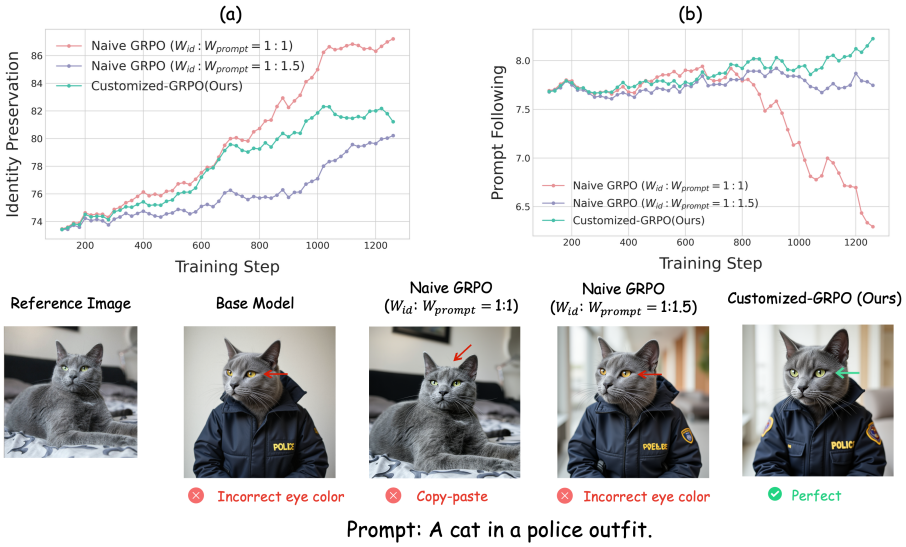

Figure 2: **Comparison of Naive GRPO with fixed linear weights against our Customized-GRPO**. The (a, b) training curves show that Naive GRPO with 1:1 weights (red) overfits to identity, causing a collapse in prompt following while 1:1.5 weights (purple) prevents this collapse at the cost of poor identity learning. Our method (green) improves both capabilities concurrently. The qualitative results (bottom) confirm these findings.

Tan et al., 2024; Xiao et al., 2025; Kumari et al., 2025) introduce lightweight cross-modal adapters to bridge text and visual representations without requiring per-subject fine-tuning. However, in practice, they often exhibit poor generalization when confronted with complex prompts or novel subjects. To mitigate this, recent methods have explored more generalized strategies, ranging from large-scaled pre-training (Kumari et al., 2025) to sophisticated attention calibration and network designs that attempt to disentangle identity and prompt representations (Kang et al., 2025; He and Yao, 2025; Huang et al., 2025a). Nonetheless, essentially, all of the above methods rely on Supervised Fine-Tuning (SFT), which operates as a form of behavior-cloning paradigm that lacks the capacity to generalize beyond the observed examples. This limitation becomes more pronounced when the model encounters novel prompts or unseen subjects, where the training distribution no longer provides reliable guidance.

To move beyond the behavior-cloning by supervised learning, we adopt the paradigm of online Reinforcement Learning (RL). Recent advances, Group Relative Policy Optimization (GRPO) (Xue et al., 2025; Liu et al., 2025; Yuan et al., 2025), have shown that directly optimizing generation policies against learned reward functions can significantly enhance text-to-image (T2I) alignment and perceptual quality. However, extending GRPO to the subject-driven domain introduces new complexity. Unlike general T2I settings that pursue a single goal such as text–image consistency, subject-driven generation task requires multi-objective alignment between identity fidelity and prompt adherence, two competing goals as illustrated in Figure 2. We find that naive GRPO with fixed linear weighting fails to reconcile these competing rewards: heavier identity weighting causes the model to overfit the subject and ignore the prompt, while increasing prompt weight restores editability at the cost of fidelity. This imbalance arises because linearly aggregated rewards produce weak and ambiguous gradients, while static weighting neglects the coarse-to-fine dynamics of diffusion. Consequently, the optimization oscillates between objectives and result in competitive degradation, where improving one capability inevitably degrades the other.

In this work, to overcome competitive degradation, we introduce **Customized-GRPO**, a novel approach featuring two synergistic innovations. First, to address the issue of reward conflict, we introduce Synergy-Aware Reward Shaping (SARS), a Pareto-inspired mechanism that provides a sharp and decisive learning signal by explicitly penalizing misaligned advantage signals and rewarding synergistic ones. Second, to tackle the problem of static optimization pressure, we develop Time-Aware Dynamic Weighting (TDW). Motivated by a Fourier analysis of the denoising process, this method aligns the optimization objective with the model's coarse-to-fine generation trajectory by dynamically allocating weights to each reward based on the current timestep.

Through extensive experiments, we demonstrate that Customized-GRPO successfully achieves a superior balance and generating images that are both faithful to the subject's identity and accurately aligned with complex textual prompts, as validated by the improvements shown in Figure 1 and 2.

2 Related Work

2.1 Subject-Driven Generation

Subject-driven image generation aims to synthesize novel images of a specific concept provided through a few reference images, while adhering to the guidance of a textual prompt. Early finetuning-based methods like DreamBooth (Ruiz et al., 2023) and Textual Inversion (Gal et al., 2022) achieve per-subject customization through fine-tuning, but this process is computationally expensive and must be repeated for each new concept.

To overcome this limitation, a significant body of work has focused on adapter-based methods. A popular approach involves training lightweight adapters to inject visual conditions into the model's attention layers (Ye et al., 2023; Tan et al., 2024; Wu et al., 2025; Huang et al., 2025a; Zhang et al., 2024). More recently, a distinct inference-time paradigm has emerged. These methods (Shin et al., 2025; Kang et al., 2025; He and Yao, 2025) require no additional training and typically operate by reframing the task as a form of guided inpainting or by directly manipulating the model's attention maps to mitigate concept confusion.

2.2 GRPO in Text-to-Image Generation

Reinforcement Learning (RL), particularly methods designed for aligning Large Language Models (LLMs) with human feedback, has recently emerged as a powerful paradigm for enhancing visual generation models. Among these, Group Relative Policy Optimization (GRPO) (Guo et al., 2025) has become prominent due to its training stability and efficiency.

GRPO in Autoregressive Models. The application of GRPO to autoregressive (AR) visual models is a natural extension of its success in the text domain. AR-GRPO (Yuan et al., 2025) adapt the framework to fine-tune AR image generators for better alignment with human perceptual preferences. Subsequent research has leveraged GRPO to unlock more complex capabilities. T2I-R1 (Jiang et al., 2025) and GoT-R1 (Duan et al., 2025) introduce Chain-of-Thought-inspired planning stages, using GRPO with sophisticated reward systems to guide the model towards discovering superior semantic-spatial reasoning strategies for complex compositional generation.

GRPO in Diffusion and Flow Matching Models. Applying online RL to non-autoregressive models like diffusion and flow matching poses a greater challenge, as their standard deterministic sampling processes lack the stochasticity required for exploration. A key breakthrough is the introduction of an **ODE-to-SDE conversion** during training, which injects controllable noise to enable policy learning. DanceGRPO (Xue et al., 2025) and Flow-GRPO (Liu et al., 2025) are foundational in this area, establishing a stable and scalable framework for applying GRPO across diverse generative paradigms and tasks. Building upon this, MixGRPO (Li et al., 2025a) further enhance training efficiency by proposing a hybrid mixed ODE-SDE sampling strategy, confining the computationally intensive SDE exploration to a small sliding window of timesteps.

3 Problem Formulation

Diffusion Model. A diffusion process gradually destroys an observed datapoint $\mathbf{x}$ over timestep t, by mixing data with noise, and the forward process of the diffusion model can be defined as (Ho et al., 2020):

$$\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \quad (1)$$

and α_t and σ_t denote the noise schedule. The noise schedule is designed in a way such that $\mathbf{z}_0$ is close to clean data and $\mathbf{z}_1$ is close to Gaussian noise. To generate a new sample, we initialize the sample $\mathbf{z}_1$ and define the sample equation of the diffusion model given the denoising model output $\hat{\boldsymbol{\epsilon}}$ at time step t:

$$\mathbf{z}_s = \alpha_s \hat{\mathbf{x}} + \sigma_s \hat{\boldsymbol{\epsilon}}, \quad (2)$$

where $\hat{\mathbf{x}}$ can be derived via Eq.(1) and then we can reach a lower noise level s.

Flow Matching. One drawback of this iterative denoising process is it can be computationally expensive and slow. To address this, Flow Matching models (Liu et al., 2022) directly learn the velocity field of the data transformation, enabling faster generation without relying on slow step-by-step denoising. In the rectified flow (Liu et al., 2022), a specific form of flow matching, the forward process is defined as a linear interpolation between the true

data sample $x_0 \sim X_0$ and $x_1 \sim X_1$ denote a noise sample:

$$\mathbf{z}_t = (1-t)\mathbf{x_0} + t\mathbf{x}_1 \tag{3}$$

Then a transformer model is trained to directly predict the velocity field $v_\theta(z_t, t)$ by minimizing the Flow Matching objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0\sim X_0,\,\mathbf{x}_1\sim X_1}\left[\|v - v_\theta(\mathbf{z}_t, t)\|^2\right] \tag{4}$$

where v is the ground-truth velocity field.

Subject-Driven Image Generation. Subject-driven image generation aims to generate images conditioned on both a textual prompt, and a reference image, which provides the visual information defining the subject's unique identity. By incorporating visual reference information, subject-driven image generation not only enhances personalization and identity coherence but also allows the model to retain key appearance details of the subject, even when the scene undergoes significant changes. Formally, we denote this image generation process as:

$$\mathbf{o} = \mathcal{G}(\mathbf{c}_{\text{prompt}}, \mathcal{I}_{\text{ref}}; \theta) \tag{5}$$

where $\mathcal{G}$ is the generative model parameterized by θ, $\mathbf{c}_{\text{prompt}}$ and $\mathcal{I}_{\text{ref}}$ represent the textual prompt and the reference image, respectively. $\mathbf{o}$ is the generated image. In existing literature, $\mathcal{G}$ is commonly implemented using diffusion models or flow matching models, both of which are based on iterative processes of noising and denoising.

Optimization with Reinforcement Learning. Recent work has framed this iterative visual generation process as a Markov Decision Process (MDP) (Black et al., 2023) and employing reinforcement learning to maximize a given reward function (Xue et al., 2025; Liu et al., 2025; Li et al., 2025a). Specifically, in the context of image generation, the MDP is a five-tuple (S, A, ρ_0, P, R), each state $s_t \in S$ at timestep t is represented as $s_t \triangleq (c, t, z_t)$ where c is the conditioning information (e.g., text prompt $\mathbf{c}_{\text{prompt}}$ and reference image $\mathcal{I}_{\text{ref}}$), t is the current diffusion timestep, and z_t denotes the corresponding noisy latent representation of the image. The action $a_t \in A$ corresponds to predicting the subsequent, less noisy latent $z_{t-\Delta t}$. The generative policy model $\mathcal{G}$ parameterized by θ serves as the the transition dynamics P which determines the transition between latent states:

$$\pi_\theta(a_t \mid s_t) = \mathcal{G}_\theta(s_t), \tag{6}$$

ρ_0 is the initial state distribution and R is the reward function to measure the image quality.

Based on the MDP formulation, reinforcement learning, specifically Group Relative Policy Optimization (GRPO) (Guo et al., 2025), is employed to optimize the policy model π_θ by maximizing the following objective function:

$$\mathcal{J}(\theta) = \mathbb{E}_{\substack{\{\mathbf{o}_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|\mathbf{c}) \\ \mathbf{a}_{t,i}\sim\pi_{\theta_{\text{old}}}(\cdot|\mathbf{s}_{t,i})}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=1}^{T}\min\Bigg(\rho_{t,i}A_i, \text{clip}\big(\rho_{t,i}, 1-\epsilon, 1+\epsilon\big)A_i\Bigg)\Bigg] \tag{7}$$

where $\rho_t = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ is the probability ratio between the new and old policies at timestep t for sample i, and ϵ is the clipping hyperparameter to stabilize training. The expectation $\mathbb{E}$ is taken over groups of samples generated by the old policy $\pi_{\theta_{\text{old}}}$. For a given condition c, generative models will sample a group of outputs $\{o_1, o_2, \ldots, o_G\}$ from the model $\pi_{\theta_{\text{old}}}$. Then, the advantage A_i for each sample o_i is then computed through intra-group normalization:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\}) + \epsilon_{\text{std}}} \tag{8}$$

where r_i is the reward score for sample o_i, ϵ_{std} is a small constant to prevent division by zero. Following prior work (Xue et al., 2025; Liu et al., 2025), we omit the KL-regularization term by default, as it yields minimal performance differences in our experiments.

4 Pilot Experiments and Analysis

However, when applying reinforcement learning to subject-driven image generation, we encounter a key challenge: optimizing the policy model simultaneously with respect to two distinct objectives: **identity preservation** and **prompt adherence**. A straightforward approach is to aggregate these independent reward signals into a single composite reward function. In this formulation, the aggregated advantage for a batch of generated outputs $\{o_1, o_2, \ldots, o_G\}$ is expressed as:

$$A_i^{\text{naive}} = w_{\text{id}} \cdot A_i^{\text{id}} + w_{\text{prompt}} \cdot A_i^{\text{prompt}}, \tag{9}$$

where A_i^{id} and A_i^{prompt} denote the advantage derived from their respective reward R_{id} and R_{prompt}. w_{id} and w_{prompt} are hyper-parameters that control the trade-off between the two objectives.

To evaluate the effectiveness of Eq. 9, we conduct a series of preliminary experiments in this section. Specifically, we adopt UNO (Wu et al., 2025) as the base model and train it using Naive GRPO on a filtered subset of Syncd dataset (Kumari et al., 2025). We investigate two weighting strategies: a balanced strategy ($w_{\text{id}} : w_{\text{prompt}} = 1 : 1$) and a prompt-biased configuration ($w_{\text{id}} : w_{\text{prompt}} = 1 : 1.5$). We use DINO and HPS-v3 (Oquab et al., 2023; Ma et al., 2025) as respective reward models for ID preservation and prompt following.

As shown in Figure 2, when using the balanced weighting strategy, the model achieves a notable improvement in ID preservation, but at the cost of significantly impaired prompt-following capability. In contrast, the imbalanced weighting strategy ($w_{\text{id}} : w_{\text{prompt}} = 1 : 1.5$) alleviates the degradation in prompt-following objective, but substantially compromises the model's ability to preserve the subject identity. Our experiments highlight a critical learning challenge: where optimizing for one objective (e.g., identity preservation) significantly conflicts with performance on the other (e.g., prompt following), leading to suboptimal overall subjective-driven image generation.

5 Method

Building on the pilot analysis presented in Section 4, we propose **Customized-GRPO**, a variant of GRPO designed to overcome the limitations of linear advantage aggregation. Specifically, we introduce a *synergy term* to mitigate reward conflicts (Section 5.1) and a *dynamic weighting strategy* to address temporal misalignment (Section 5.2).

5.1 Synergy-Aware Reward Shaping

As established in our pilot analysis, the linear aggregation of advantages is prone to suffer from signal cancellation under conflicting rewards. To address this, we introduce a non-linear term to shape the reward inspired by the principles of Pareto Optimization (Agnihotri et al., 2025; Li et al., 2025b). Instead of tolerating a trade-off, our goal is to guide the policy towards discovering solutions that the generated images excel in both ID preservation and prompt following simultaneously.

To address this issue, we introduce a *synergy term* $\mathcal{S}$ that quantifies the alignment between the two advantage signals, A_{id} and A_{prompt}. We define $\mathcal{S}$ via the hyperbolic tangent

$$\mathcal{S} = \tanh\big(A_{\text{id}} \cdot A_{\text{prompt}}\big) \tag{10}$$

which is bounded in $[-1, 1]$ and saturates for large magnitudes, thereby acting as a stabilizing mechanism that limits gradient growth. The final advantage, A^{SARS}, is then computed as a piecewise aggregation based on the polarity of the individual advantages:

$$A^{\text{SARS}} = \begin{cases} w_{\text{id}} A_{\text{id}} + w_{\text{prompt}} A_{\text{prompt}} + \alpha \cdot \mathcal{S}, & \text{if } A_{\text{id}} > 0 \text{ or } A_{\text{prompt}} > 0, \\ w_{\text{id}} A_{\text{id}} + w_{\text{prompt}} A_{\text{prompt}} - \alpha \cdot \mathcal{S}, & \text{if both } A_{\text{id}} \le 0 \text{ and } A_{\text{prompt}} \le 0, \end{cases} \tag{11}$$

where $\alpha \ge 0$ controls the influence of the synergy term. This formulation yields a more informative learning signal than simple linear aggregation.

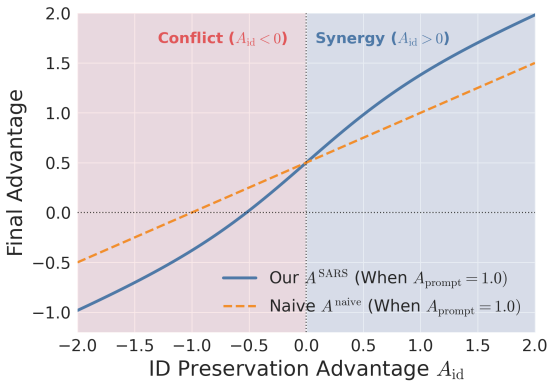

Figure 3: **Cross-section of the SARS function, illustrating how it reshapes the linear reward landscape.** We plot the final advantage as a function of A_{id} while holding A_{prompt} constant at 1.0.

As visualized in Figure 3, our SARS function (solid curve) fundamentally reshapes the naive linear baseline (dashed curve). It non-linearly amplifies the reward in the Synergy region ($A_{\text{id}} > 0$) and applies a decisive penalty in the Conflict region ($A_{\text{id}} < 0$). This non-linear transformation creates a sharper gradient around the decision boundary.

By incorporating the synergy term, our method replaces ambiguous, linearly combined feedback with sharper, decision-consistent signals. This alleviates cancellation and better aligns optimization with the goal of producing well-balanced, high-quality images.

5.2 Timestep-Aware Dynamic Weighting

Previous work has established that the diffusion denoising process is not static but exhibits distinct functional phases (Ho et al., 2020; Li et al., 2023; Salimans and Ho, 2022; Balaji et al., 2022). As confirmed by our FFT analysis in Figure 4, the early stages are dominated by the formation of low-frequency global structure, presenting an opportune window to prioritize prompt following as the model makes high-level compositional decisions. Conversely, the later stages are characterized by

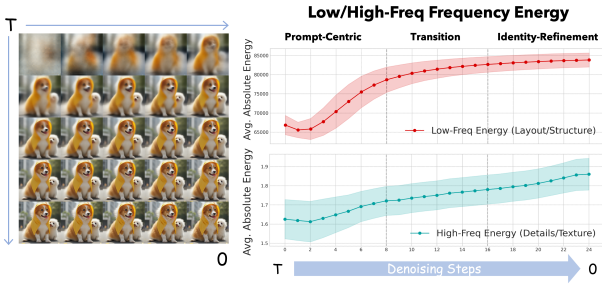

Figure 4: **Temporal Dynamics of Feature Synthesis during the Denoising Process**. (Right) The plots show that low-frequency energy (structure) converges early, whereas high-frequency energy (details) accumulates steadily over all phases. (Left) The visual progression from pure noise (t=T) to the final image (t=0) corroborates this finding.

the steady accumulation of high-frequency local details, making this phase critical for refining ID preservation. Static weighting strategy, however, suffers from Temporal Misalignment, fundamentally conflicting with the dynamic nature of the denoising process.

To address this, we propose Timestep-Aware Dynamic Weighting (TDW), which replaces static optimization with a curriculum-inspired strategy that adapts weighting over time. The core idea is to align the optimization pressure with the model's transient focus. We partition the denoising trajectory into three distinct stages: an early Prompt-Centric Phase to establish global structure, a late Identity-Refinement Phase to perfect local details, and a smooth Transition Phase bridging them.

Specifically, we design the following function for $w_{\text{prompt}}(t)$:

$$w_{\text{prompt}}(t) = \begin{cases} w_{\max} & \text{if } T \geq t > T_{\text{trans}} \\ w_{\min} + (w_{\max} - w_{\min}) \cdot & \text{if } T_{\text{trans}} \geq t > T_{\text{id}} \\ \quad (1 - \sigma(\lambda(t - t_c))) & \\ w_{\min} & \text{if } T_{\text{id}} \geq t > 0 \end{cases} \tag{12}$$

where the weight for ID preservation is $w_{\text{id}}(t) = 1 - w_{\text{prompt}}(t)$. To ensure stability, the weights are constrained by upper and lower bounds $w_{\max}$ and $w_{\min}$ and a Sigmoid function $\sigma(\cdot)$ provides a smooth interpolation during the transition phase, preventing abrupt policy shifts.

By temporally decoupling the competing objectives, TDW enables a more principled optimization. It transforms the learning problem from navigating a static, inefficient trade-off into a curriculum-like process, allowing the policy to allocate its limited gradient budget to the most critical task at each timestep.

6 Experiments

6.1 Experiment Setup

Implementation details. Our main experiments are conducted on the subject-driven generation model UNO (Wu et al., 2025), a Diffusion Transformer (DiT) based architecture built upon FLUX.1-dev (Andreas Blattmann, 2024). We freeze the pretrained DiT backbone and the image encoder, training only the injected LoRA (Hu et al., 2022) matrices. We provide training details in Appendix A.

Dataset and Reward Models. We use a high-quality, filtered subset of 10k samples from the large-scale Syncd dataset (Kumari et al., 2025) for training. To guide our policy optimization, we define two reward functions corresponding to our core objectives. For ID preservation, we use a segmentation-masked DINOv2 (Oquab et al., 2023) score, termed DINO-Seg, to measure the fidelity between the subject in the generated image and the reference. It remove the background variations when computing the image similarity to better reflect the faithfulness to the reference subject. For prompt following, we use the Human Preference Score v3 (HPS-v3) (Ma et al., 2025), a powerful reward model trained on a large dataset of human preferences to evaluate alignment with the text prompt.

Evaluation Protocol. We conduct a comprehensive evaluation on the widely-used DreamBench (Ruiz et al., 2023). Our method is compared against a suite of state-of-the-art baselines, including both fine-tuning-based methods: Textual Inversion (Gal et al., 2022), DreamBooth (Ruiz et al., 2023); finetuning-free methods BLIP-Diffusion (Li et al., 2023), ELITE (Wei et al., 2023), SSR-Encoder (Zhang et al., 2024), OminiControl (Tan et al., 2024), OmniGen (Xiao et al., 2025), FLUX IP-Adatper (team, 2025) and RL methods RPO (Miao et al., 2024). Performance is measured using established metrics: CLIP-T (Radford et al., 2021) for prompt adherence, CLIP-I and DINO Score (Hessel et al., 2021; Caron et al., 2021) for identity preservation, and the HPS-v3 (Ma et al., 2025) for text-image alignment and aesthetics.

6.2 Main Results

Table 1 reports the main quantitative results of our method and against state-of-the-art baselines

Method	DINO↑	CLIP-I↑	CLIP-T↑	HPS-v3↑
Textual Inversion	0.569	0.780	0.255	–
DreamBooth	0.668	0.803	0.305	–
BLIP-Diffusion	0.670	0.805	0.302	–
ELITE	0.647	0.772	0.296	3.12
SSR-Encoder	0.612	0.821	0.308	4.82
OmniGen	0.693	0.801	**0.315**	6.67
OminiControl	0.684	0.799	0.312	8.28
FLUX.1 IP-Adapter	0.582	0.820	0.288	7.68
UNO (Base)	0.760	0.835	0.304	7.87
UNO (RPO)	0.852	0.898	0.237	4.47
UNO (Customized-GRPO)	**0.812**	**0.862**	0.301	**8.32**

Table 1: Quantitative Results for subject-driven generation on Dreambench.

Method	DINO↑	CLIP-I↑	CLIP-T↑	HPS-v3↑
UNO (Base)	0.760	0.835	0.304	7.87
SFT	0.762	0.846	**0.307**	7.43
RPO	0.852	0.898	0.237	4.47
Naive GRPO (1:1)	0.861	0.912	0.235	4.12
Naive GRPO (1:1.5)	0.801	0.842	0.298	7.56
Customized-GRPO	**0.812**	**0.862**	0.301	**8.32**
w/o Synergy Term	0.795	0.838	0.295	7.95
w/o TDW	0.811	0.852	0.278	7.23

Table 2: Ablation Results on Dreambench. The best performance is marked in **bold**. We highlight the row in red to denote the competitive degradation.

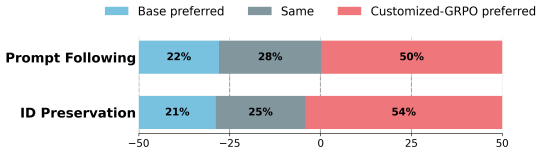

Figure 5: **Human evaluation.** Pairwise comparison between our Customized-GRPO and the base model.

on Dreambench. Our Customized-GRPO demonstrates superior overall performance. It achieves the highest scores with a DINO score of 0.812 and a CLIP-I score of 0.872. While the CLIP-T score (0.301), which measures strict text-image similarity, remains on par with top-performing baselines like OmniGen (0.315) and DreamBooth (0.305), our method attains the highest Human Preference Score (HPS-v3) of 8.32, which evaluates both text-image alignment and overall aesthetic quality. It indicates that while maintaining a strong textual alignment, our policy optimization has successfully learned to generate images with higher overall aesthetic quality and are better aligned with human perception.

Overall, Customized-GRPO improves upon our UNO base model by 6.8% in the DINO score for ID preservation while simultaneously increasing the HPS-v3 score for prompt following and human preference by 5.7%. The concurrent improvement across both ID preservation and prompt following metrics provides clear evidence that our method effectively mitigates the competitive degradation observed in naive approaches and successfully learns a more balanced policy.

We conduct a pairwise human evaluation comparing Customized-GRPO against the base UNO model. As shown in Figure 5, our method is strongly preferred by human annotators, winning in Prompt Following and ID Preservation by decisive margins of 50% vs. 22% and 54% vs. 21%, respectively. In both criteria, our method is preferred more than twice as often as the base model.

We provide qualitative comparisons in Appendix C.3, which further visualize our method's superior ability to preserve identity while adhering to complex prompts.

6.3 Ablation Study

We conduct ablation studies to validate the efficacy of our method: Synergy-Aware Reward Shaping and Time-Aware Dynamic Weighting. We compare our full Customized-GRPO method against ablated versions where each component is removed. The results are presented in Table 2.

Effect of Synergy-Aware Reward Shaping. We evaluate a model with only TDW active (w/o Synergy Term), which applies dynamic weights but relies on a simple linear advantage sum. It successfully prevents the catastrophic performance collapse seen in prompt following observed in the Naive GRPO (HPS-v3 = 7.95 vs. 4.12) but all metrics remain notably lower than our full model. This indicates that while TDW correctly allocates optimization pressure across timesteps, it is insufficient to resolve the underlying reward conflict at each individual output in the generated group. The synergy term is therefore critical for providing a sharper, more decisive learning signal that leads to a superior final policy.

Effect of Time-Aware Dynamic Weighting. We evaluate a model with only SARS active (w/o TDW), which uses our conflict-aware advantage function with fixed 1:1 weights across all timesteps. While this model also improves upon the naive baseline by penalizing conflicted outputs, its performance on prompt following metrics is substantially weaker than full method. This result confirms our hypothesis: without dynamically shifting the focus from prompt following to ID preservation, the static optimization pressure still leads to an imbalanced policy that over-emphasizes fidelity in the later, detail-focused stages of denoising.

Method	DINO↑	CLIP-I↑	CLIP-T↑	HPS-v3↑
UNO (Base)	0.760	0.835	**0.304**	7.87
Naive GRPO	0.801	0.842	0.298	7.56
Min(A_{id}, A_{prompt})	0.785	0.848	0.298	7.81
Max(A_{id}, A_{prompt})	0.884	0.899	0.276	6.26
Harmonic Mean	0.778	0.812	0.291	8.22
Tanh($A_{\text{id}} \cdot A_{\text{prompt}}$)	**0.812**	**0.862**	0.301	**8.32**

Table 3: Analysis on the choice of Synergy Term. The best performance is marked in **bold**. We highlight the row in red to denote the competitive degradation.

$w_{max} : w_{min}$	DINO↑	CLIP-I↑	CLIP-T↑	HPS-v3↑
1.0 : 0	0.832	0.874	0.272	6.39
0.9 : 0.1	0.821	0.866	0.285	7.27
0.8 : 0.2	0.814	0.864	0.296	7.68
0.7 : 0.3	**0.812**	**0.862**	**0.301**	**8.32**
0.6 : 0.4	0.795	0.848	0.292	7.88

Table 4: Analysis on the choice of Dynamic Weight. The best performance is marked in **bold**.

6.4 Analyses

Analysis on the choice of Synergy Term. We conduct a comparative analysis of linear function ($w_{\text{id}} : w_{\text{prompt}} = 1 : 1.5$) and four non-linear functions to determine the most effective function for synergy term. The results are presented in Table 3. Max function exhibits classic reward hacking, achieving high fidelity scores by sacrificing prompt following, where the model learns to maximize one objective at the severe expense of the other. Min function, conversely, proves too conservative and fails to deliver significant improvements.

While Harmonic Mean and Tanh functions both designed to encourage balance, yielding promising results by successfully avoiding reward hacking, Tanh function proves superior, attaining the highest HPS-v3 score of 8.32 while maintaining a strong, balanced performance across all metrics.

We attribute the success to its ideal mathematical properties for this task: its bounded and symmetric nature provides a stable and consistent signal. Therefore, we adopt it as the synergy function.

Comparison with RL and SFT. To further validate the efficacy of our online RL approach, we conduct a direct comparison against Supervised Fine-Tuning (SFT) and a recent offline preference optimization baseline, RPO (Miao et al., 2024). We fine-tune the UNO (base) model using SFT and RPO on the same 10k high-quality data curated for our experiments. The results, presented in Table 2, show that SFT yields only marginal or even negative changes compared to the base model. Since the base model is already well-trained on a similar data distribution, simply continuing to imitate this static dataset offers no clear path to resolving the complex trade-off. Conversely, while offline RPO significantly improves identity fidelity (DINO: 0.852), it incurs severe competitive degradation in text alignment, with the HPS-v3 score dropping drastically to 4.47. This indicates that RPO tends to overfit the reference image, failing to balance conflicting objectives due to its static harmonic mean aggregation and the inherent challenge of constructing balanced offline preference pairs (implementation details in Appendix A). In contrast, our Customized-GRPO method demonstrates significant performance gains across both objectives.

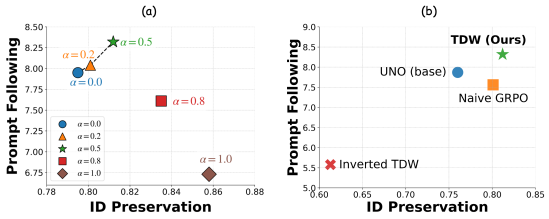

Figure 6: Analysis of the Synergy Coefficient and TDW Schedule.

Dynamic Weight Bounds Analysis. For dynamic weight bounds w_{max}, w_{min}, we evaluate the impact of different dynamic ranges on model performance while keeping other hyper-parameters fixed. The results are presented in Table 4. Our observations reveal two notable trends: **overly aggressive decay** ($w_{min} \leq 0.1$) causes *semantic collapse* in the final denoising stages, where the model tends to overfit the reference image and fails to adhere to textual constraints. Conversely, **insufficient dynamic range**($w_{max} : w_{mian} = 0.6 : 0.4$) fails to resolve competitive degradation, resulting in drops for both DINO (0.795) and HPS-v3 (7.88) compared to the optimal setting. Based on these results, we adopt $w_{max} : w_{mian} = 0.7 : 0.3$ as the default configuration in this work.

Synergy Coefficient Analysis. We analyze the impact of the synergy coefficient α, which controls the strength of the SARS adjustment. As shown in Figure 6 (a), the results reveal a clear optimal balance. When α increase from 0 to 0.5, we observe a steady improvement on both objectives, demonstrating the effectiveness of SARS in transforming conflicted signals into a productive gradient. However, further increasing α to 0.8 and 1.0 leads to a significant performance degradation. This indicates that an overly large α induces a risk-averse policy that excessively penalizes conflict, thereby failing

Synergy Coef. (α)	Standard TDW		Inverted TDW	
	DINO↑	HPS-v3↑	DINO↑	HPS-v3↑
$\alpha = 0.0$	0.795	7.95	0.605	5.65
$\alpha = 0.2$	0.801	8.04	0.609	5.62
$\boldsymbol{\alpha = 0.5}$	**0.812**	**8.32**	**0.613**	**5.58**
$\alpha = 0.8$	0.835	7.61	0.616	5.49
$\alpha = 1.0$	0.858	6.73	0.620	5.41

Table 5: Analysis of the interaction between Synergy Coefficient and TDW schedules.

to achieve a good balance. Based on this analysis , we adopt $\alpha = 0.5$ for all subsequent experiments.

Validation of TDW Weighting Schedule. We conduct an analysis comparing our standard TDW against two critical ablations: a Naive GRPO using static weights, and an Inverted TDW that reverses the curriculum. As shown in Figure 6(b), the Inverted TDW suffers a severe collapse in performance, empirically confirming that our schedule is essential and aligns with the model's coarse-to-fine generation process. Meanwhile, Naive GRPO demonstrates the inefficient trade-off inherent in static weighting. Our standard TDW is positioned distinctly in the top-right corner, demonstrating its superior ability to mitigate the trade-off and successfully improve both objectives simultaneously.

Interaction between SARS and TDW. To systematically analyse the hyperparameter interactions within our framework, we find that SARS and TDW address competitive degradation from distinct dimensions. We propose that TDW acts as the performance foundation, while SARS acts as the performance amplifier. Consequently, the efficacy of SARS depends on a reasonable TDW schedule. To validate this, we conduct comprehensive experiments on DreamBench, varying $\alpha \in 0.0, 0.2, 0.5, 0.8, 1.0$ across both the standard TDW schedule and the Inverted TDW schedule (which reverses the curriculum).

As shown in Table 5, aligning the weighting schedule with diffusion dynamics is the foundation for success. An inverted schedule leads to a performance collapse (DINO 0.6, HPS-v3 0.3) regardless of the α value, showing minimal sensitivity to parameter changes. Under the correct Standard TDW, when α increases from 0 to 0.5, we observe a steady improvement on both objectives. However, further increasing α to 0.8 and 1.0 triggers **reward hacking**, where the model over-optimizes Identity at the significant expense of Prompt adherence.

7 Conclusion

In this work, we introduce Customized-GRPO, a novel reinforcement learning framework featuring Synergy-Aware Reward Shaping (SARS) and Time-Aware Dynamic Weighting (TDW) to resolve the critical fidelity-editability trade-off in subject-driven generation. Experiments show that naive GRPO approaches fail due to reward conflict and temporal misalignment, while our method successfully generates images that are simultaneously faithful to the subject's identity and accurately adhere to complex prompts.

Limitation

In this section, we discuss the limitation of our work: (1) Due to computational constraints, our experiments are conducted exclusively on a Diffusion Transformer (DiT) architecture. While our framework is theoretically model-agnostic, future work should involve applying Customized-GRPO to other generative architectures, particularly autoregressive models, to fully validate its generalization across different modeling paradigms. (2) Our current framework focuses on optimizing the generation of a single subject per image. However, a significant emerging challenge in personalized generation is multi-concept composition, which involves generating an image containing multiple, distinct subjects. We will extend our reinforcement learning framwork to manage and balance the rewards for multiple, interacting concepts in the future work.

Ethics Statement

The training data used in our experiments is a filtered subset of the publicly available dataset. During the curation process, we applied automated filtering mechanisms to detect and remove any images containing material, or personally identifiable information. For the textual prompts, we conduct a comprehensive manual review to further ensure the appropriateness of the content.

Acknowledgment

This work was supported in part by the Ningbo Youth Science and Technology Innovation Leading Talent Program (No. 2025QL059), and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. 2025. Multi-objective preference optimization: Improving human alignment of generative models. *arXiv preprint arXiv:2505.10892*.

Dominik Lorenz Andreas Blattmann, Axel Sauer. 2024. Flux.1.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, and 1 others. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Jinhe Bi, Yifan Wang, Danqi Yan, Aniri, Wenke Huang, Zengjie Jin, Xiaowen Ma, Artur Hecker, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. 2025. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *Preprint*, arXiv:2502.12119.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. 2025. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Qiyuan He and Angela Yao. 2025. Conceptrol: Concept control of zero-shot personalized image generation. *arXiv preprint arXiv:2503.06568*.

Wanggui He, Siming Fu, Mushui Liu, Xierui Wang, Wenyi Xiao, Fangxun Shu, Yi Wang, Lei Zhang, Zhelun Yu, Haoyuan Li, and 1 others. 2024. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. *arXiv preprint arXiv:2407.07614*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Linyan Huang, Haonan Lin, Yanning Zhou, and Kaiwen Xiao. 2025a. Flexip: Dynamic control of preservation and personality for customized image generation. *arXiv preprint arXiv:2504.07405*.

Ziwei Huang, Wanggui He, Quanyu Long, Yandi Wang, Haoyuan Li, Zhelun Yu, Fangxun Shu, Weilong Dai, Hao Jiang, Fei Wu, and Leilei Gan. 2025b. T2I-FactualBench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27501–27524, Vienna, Austria. Association for Computational Linguistics.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. 2025. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*.

Hao Kang, Stathi Fotiadis, Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Min Jin Chong, and Xin Lu. 2025. Flux already knows–activating subject-driven image generation without training. *arXiv preprint arXiv:2504.11478*.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

Nupur Kumari, Xi Yin, Jun-Yan Zhu, Ishan Misra, and Samaneh Azadi. 2025. Generating multi-image synthetic data for text-to-image customization. *arXiv preprint arXiv:2502.01720*.

Dongxu Li, Junnan Li, and Steven Hoi. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. 2025a. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*.

Moxin Li, Yuantao Zhang, Wenjie Wang, Wentao Shi, Zhuo Liu, Fuli Feng, and Tat-Seng Chua. 2025b. Self-improvement towards pareto optimality: Mitigating preference conflicts in multi-objective alignment. *arXiv preprint arXiv:2502.14354*.

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

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

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. 2025. Hpsv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*.

Yanting Miao, William Loh, Suraj Kothawade, Pascal Poupart, Abdullah Rashwan, and Yeqing Li. 2024. Subject-driven text-to-image generation via preference-based reinforcement learning. *Advances in Neural Information Processing Systems*, 37:123563–123591.

OpenAI. 2024. Chatgpt. https://openai.com/index/gpt-4o-system-card/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2024. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510.

Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. 2025. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7986–7996.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

XLabs AI team. 2025. x-flux. Accessed: 2025-02-07.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. 2025. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. 2025. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and 1 others. 2025. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Shihao Yuan, Yahui Liu, Yang Yue, Jingyuan Zhang, Wangmeng Zuo, Qi Wang, Fuzheng Zhang, and Guorui Zhou. 2025. Ar-grpo: Training autoregressive image generation models via reinforcement learning. *arXiv preprint arXiv:2508.06924*.

Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, and 1 others. 2024. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078.

A Details on Experimental Setup

A.1 Implementation Details

Our main experiments are conducted on the subject-driven generation model UNO (Wu et al., 2025), a Diffusion Transformer (DiT) based architecture built upon FLUX.1-dev (Andreas Blattmann, 2024). We freeze the pretrained DiT backbone and the image encoder, training only the injected LoRA (Hu et al., 2022) matrices. For training-time sampling, we set T = 25 as the total sampling steps. For GRPO, the model enerates 12 images for each prompt and clips the advantage to the range [1e-5, 1e-5]. We use AdamW as the optimizer with a learning rate of 1e-5 and a weight decay coefficient of 0.0001. All experiments are conducted on 8 NVIDIA A100 GPUs with a batch size of 1, for a maximum of 1250 training steps, resulting in a total of approximately 80 wall-clock hours (i.e., 640 GPU hours).

Hyperparameter Configuration. For Synergy-Aware Reward Shaping (SARS), we set the synergy coefficient α to 0.5 based on the ablation study in Section 6.4. For Time-Aware Dynamic Weighting (TDW), the weights for prompt following, $w_{\text{prompt}}(t)$, are bounded by $w_{\max} = 0.7$ and $w_{\min} = 0.3$. The three optimization phases are partitioned as follows: Prompt-Centric Phase for $t < 6$, Transition Phase for $6 \leq t < 22$, and Identity-Refinement Phase for $t \geq 22$.

Training Dataset. We utilize the Syncd dataset (Kumari et al., 2025), a large-scale dataset designed for subject-driven generation, including rigid and deformable categories. To create a more focused and high-quality subset for our experiments, we filter 10,000 datasets with the highest average DINOv2 similarity and aesthetic scores.

Reward Model. For ID Preservation, we use DINOv2 (ViT-L/14) (Oquab et al., 2023) as the foundation for our fidelity reward. To ensure that the reward signal is focused specifically on the subject and not influenced by background, we employ a object detection and segmentation approach. This helps remove the background variations when computing the image similarity scores to better reflect the faithfulness to the reference subject. We call it DINO-Seg. For Prompt Following, we use the Human Preference Score v3 (HPSv3) (Ma et al., 2025) as our reward model. HPSv3 is a powerful reward model trained on a large dataset of human preferences, and it provides a reliable score that reflects how well a generated image aligns with the semantics of a given text prompt.

A.2 Evaluation Details

Benchmark. To ensure a robust and comprehensive assessment, we evaluate all methods on the widely-used DreamBench (Ruiz et al., 2023). For each of the 30 subjects in the benchmark, we use all associated prompts. To ensure evaluation stability and account for stochasticity in the generation process, we generate 4 images per prompt and report the average score across all generated images.

Baselines. We compare our method with SOTA methods including both finetuning-based methods: Textual Inversion (Gal et al., 2022), DreamBooth (Ruiz et al., 2023) and finetuning-free methods: BLIP-Diffusion (Li et al., 2023), ELITE (Wei et al., 2023), SSR-Encoder (Zhang et al., 2024), OminiControl (Tan et al., 2024), OmniGen (Xiao et al., 2025) and FLUX IP-Adatper (team, 2025).

Regarding the comparison with RPO (Miao et al., 2024), given that their original data is not publicly available, we strictly replicate their data construction pipeline to curate a paired preference dataset using our own training data (Kumari et al., 2025). Specifically, we employ the frozen UNO base model to generate candidate images conditioned solely on text prompts. Subsequently, we adopt RPO's core Harmonic Mean reward aggregation mechanism to distinguish between winning and losing samples for preference pair construction. Following RPO protocol, we fine-tune UNO using the composite objective proposed, which combines the Offline DPO (Rafailov et al., 2023) loss for preference optimization with a Similarity Loss to ensure identity consistency.

We adopted the default hyperparameters specified for each model by their respective authors.

Evaluation Metrics. We follow previous methods to adopt three metrics (CLIP-T, CLIP-I, and DINO) for evaluation. Specifically, CLIP-T evaluates the similarity between the generated images and given text prompts; CLIP-I and DINO evaluate the similarity between the generated images and the reference images. To better capture overall quality and alignment with human perception, we additionally report the HPS-v3 (Ma et al., 2025). This metric evaluates not only text-image alignment but also broader aesthetic qualities.

A.3 Human Evaluation Details

To complement our quantitative metrics, we conduct a comprehensive human evaluation comparing our Customized-GRPO against base UNO model. We randomly select 200 diverse subject-prompt pairs from DreamBench, generating one image per model (Customized-GRPO and UNO base) for each pair. We engage eight experienced, English-fluent annotators on iTAG[2] platform for the task. Before evaluation, all annotators completed a detailed tutorial to align their understanding of the criteria. As illustrated in Figure 8, the evaluation interface present annotators with the text prompt and two sets of images. For each model being compared (Model 1 and Model 2), we display its generated image alongside the corresponding reference image. The positions of Model 1 and Model 2 were randomized to prevent positional bias. Annotators perform a pairwise comparison, answering two question: "which model performs better based on the visual similarity between the generated image and the reference image of the object/animal?" and "which model performs better by evaluating both (1) how well the generated image align with the provided prompt, and (2) the overall visual quality (aesthetics) of the generated image?" For each question, annotators could choose one of the two models or select "Tie" if they are of similar quality.

B Additional Methodological Details

B.1 ODE-to-SDE for Exploration

A fundamental challenge in applying GRPO to rectified flow models is that their standard sampling process is deterministic Ordinary Differential Equation (ODE) (Song et al., 2020): $\mathrm{d}\mathbf{z}_t = \mathbf{u}_t \mathrm{d}t$. This deterministic nature prevents the stochastic exploration required for the policy to discover multiple trajectory. The key insight from prior work(Xue et al., 2025; Liu et al., 2025) is to convert this deterministic ODE into a Stochastic Differential Equation (SDE) during training. This is achieved by introducing a controlled noise term into the reverse-time generative process:

$$d\mathbf{z}_t = \left(u_\theta(\mathbf{z}_t, t) - \frac{1}{2}\varepsilon_t^2 \nabla \log p_t(\mathbf{z}_t) \right) dt + \varepsilon_t d\mathbf{w} \tag{13}$$

where $\mathrm{d}\mathbf{w}$ is a Brownian motion, and ε_t introduces the stochasticity during sampling. The SDE specifically designed to preserve the marginal distributions $p_t(\mathbf{z}_t)$ of the original ODE, enabling valid stochastic exploration for GRPO.

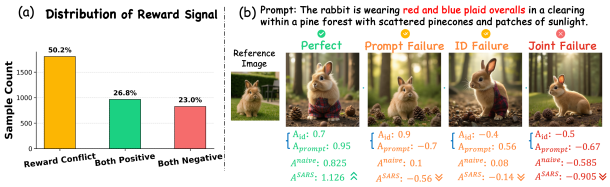

Figure 7: **The Impact of Reward Conflict and Our Synergy-Aware Reward Shaping**. (a) Distribution of advantage alignment, showing that reward conflict is the most common outcome, affecting 50.2% of samples. (b) Comparison of final advantage calculation. Our method amplifies the signal for perfect outputs while correctly penalizing conflicted cases (Poor Prompt/ID) where Naive GRPO provides a weak, misleading signal.

B.2 Detailed Analysis of Reward Conflict and SARS

This section provides a detailed empirical analysis that motivates our Synergy-Aware Reward Shaping (SARS). As discussed in the main paper, a naive application of GRPO leads to Competitive Degradation. Our central hypothesis is that this macroscopic, policy-level failure stems from a microscopic, sample-level issue: a fundamental conflict in the reward signals themselves.

To validate this, we conducted a rigorous analysis of 300 training groups, each comprising 12 images, reveals a pervasive and fundamental issue we term **Reward Conflict**. As quantified in Figure 7, we find that within a typical group, approximately 50.2% of the samples exhibit conflicting reward signals.

This high incidence of conflict leads to two critical failures in the learning process. The linear aggregation causes the positive and negative advantages from these conflicted samples to neutralize each other. This consistently pushes the final advantage A_i^{naive} towards zero, providing a weak and, more critically, misleading learning signal.

This misalignment with non-compensatory human preferences–where a single flaw is unacceptable–is evident in Figure 7(b). Naive GRPO assigns similarly low, positive advantages to a "Prompt Failure" ($A^{\text{naive}} = 0.1$) and "ID Failure" ($A^{\text{naive}} = 0.08$), failing to distinguish between these critically flawed outputs. Consequently, the model receives an ambiguous gradient that incorrectly signals these "specialist" failures are acceptable, impeding effective policy optimization.

[2]https://www.alibabacloud.com/help/en/pai/user-guide/itag/

Algorithm 1 Customized-GRPO Training for Subject-Driven Generation

Require: Initial policy model π_θ; ID reward model R_{id}; Prompt reward model R_{prompt};
Require: Paired dataset $\mathcal{D} = \{(\mathcal{I}_j, \mathbf{c}_j)\}_{j=1}^{N}$ of reference images and prompts;
Require: Total sampling steps T; Synergy hyperparameter α;
Require: Dynamic weighting function $f(t) \to (w_{\text{id}}(t), w_{\text{prompt}}(t))$
Ensure: Optimized policy model π_θ
1: **for** training iteration $= 1$ **to** M **do**
2: Sample a batch of (reference image, prompt) pairs $\{(\mathcal{I}_j, \mathbf{c}_j)\}_{j=1}^{B} \sim \mathcal{D}$
3: Update old policy: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
4: **for** each pair $(\mathcal{I}, \mathbf{c})$ in the batch **do**
5: Generate a group of G outputs: $\{\mathbf{o}_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathcal{I}, \mathbf{c})$
6: Compute advantages A^{id}, A^{prompt} for each sample i: $A_i \leftarrow \sum_{k=1}^{K} \frac{r_i^k - \mu^k}{\sigma^k}$
7: **for** each output $i \in \{1, \ldots, G\}$ **do**
8: $\mathcal{S}_i \leftarrow \tanh(A_i^{\text{id}} \cdot A_i^{\text{prompt}})$ ▷ Synergy-Aware Shaping
9: **end for**
10: Subsample a subset of training timesteps $\mathcal{T}_{\text{sub}} \subset \{1, \ldots, T\}$
11: **for** $t \in \mathcal{T}_{\text{sub}}$ **do**
12: $(w_{\text{id}}(t), w_{\text{prompt}}(t)) \leftarrow f(t)$ ▷ Timestep-Aware Weighting
13: **for** each output $i \in \{1, \ldots, G\}$ **do**
14: **if** $A_i^{\text{id}} > 0$ **or** $A_i^{\text{prompt}} > 0$ **then**
15: $A_i^{\text{final}}(t) \leftarrow w_{\text{id}}(t) A_i^{\text{id}} + w_{\text{prompt}}(t) A_i^{\text{prompt}} + \alpha \cdot \mathcal{S}_i$
16: **else**
17: $A_i^{\text{final}}(t) \leftarrow w_{\text{id}}(t) A_i^{\text{id}} + w_{\text{prompt}}(t) A_i^{\text{prompt}} - \alpha \cdot \mathcal{S}_i$
18: **end if** ▷ Combine synergy and dynamic weights
19: **end for**
20: Compute GRPO objective $\mathcal{J}$ using time-dependent advantages $\{A_i^{\text{final}}(t)\}_{i=1}^{G}$
21: Update policy parameters via gradient ascent: $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{J}$
22: **end for**
23: **end for**
24: **end for**
25: **return** π_θ

Our Synergy-Aware Reward Shaping (SARS) is explicitly designed to resolve this failure mode. As illustrated in the same case study in Figure 7(b). For a synergistic ("Perfect") output, SARS significantly amplifies the learning signal. More critically, it resolves the signal cancellation issue for conflicted samples. While Naive GRPO provides a weak and misleadingly positive signal for both "Prompt Failure" and "ID Failure" cases, SARS correctly identifies these as undesirable and transforms their advantage into a decisive negative penalty.

B.3 Customized-GRPO Algorithm

We provide a detailed training procedure for our Customized-GRPO framework in Algorithm 25. The overall process can be summarized as follows: In each iteration, we first sample a batch of reference image and text prompt pairs from our training dataset. For each pair, we use the current policy π_θ to generate a group of G candidate images. Next, we enter the core reward calculation phase. We compute the initial advantages for ID Preservation (A_{id}) and Prompt Following (A_{prompt}) for every generated image in the group. Following this, our two main contributions are applied:

Synergy-Aware Shaping: We calculate the synergy term $\mathcal{S}_\backslash$ for each sample based on the product of its advantages (Line 8).

Timestep-Aware Weighting: We iterate through a subset of training timesteps. For each timestep t, we first retrieve the dynamic weights $(w_{\text{id}}(t), w_{\text{prompt}}(t))$ from our predefined function $f(t)$ (Line 12). These weights are then combined with the synergy term $\mathcal{S}_\backslash$ to compute the final, time-dependent advantage $A_i^{\text{final}}(t)$ for each sample (Lines 14-18).

Finally, these time-dependent advantages are used to compute the GRPO objective $\mathcal{J}$.We then update the policy parameters θ via gradient ascent to maximize this objective (Line 21). This entire process is repeated until convergence.

C Additional Results

C.1 Impact of Reward Model Choice

To further validate the robustness and generalizability of our Customized-GRPO framework, we investigate its performance with a diverse suite of reward models. Beyond our primary DINO-Seg and HPSv3 combination, we also experiment

Reward Model	ID Preservation		Prompt Following	
	Pers.	Spear.	Pers.	Spear.
CLIP-I	0.37	0.48	-	-
DINO	0.42	0.65	-	-
CLIP-T	-	-	0.38	0.51
Qwen-VL-Max	0.27	0.42	0.41	0.57
GPT-4o-0816	0.38	0.63	0.52	0.69
Gemini-2.5-Pro	0.49	0.68	0.57	0.72
DINO-Seg	**0.60**	**0.75**	-	-
HPS-V3	-	-	**0.62**	**0.76**

Table 6: Comparisons of different reward model. Pers. and Spear. represents Person and Spearman correlations, respectively.

Reward Model	DINO↑	CLIP-I↑	CLIP-T↑	HPS-V3↑
UNO (Base)	0.760	0.835	0.304	7.87
DINO, CLIP-T	0.731	0.782	0.287	7.12
CLIP-I, CLIP-T	0.832	0.849	0.213	5.24
Qwen-VL-Max	0.678	0.732	0.281	6.92
GPT-4o	0.752	0.802	0.307	7.44
Gemini-2.5-Pro	0.803	0.842	0.315	8.17
DINO-Seg, HPS-v3	0.812	0.862	0.301	8.32

Table 7: Performance on DreamBench after training with different reward model combinations.

with CLIP-I and CLIP-T as alternative reward signals. Furthermore, inspired by recent work (Huang et al., 2025b; Peng et al., 2024; Ku et al., 2023; Hu et al., 2023; Bi et al., 2025), we evaluate the use of a powerful Vision-Language Model (VLM) including Gemini-2.5-Pro, Qwen-VL-Max and GPT-4o (Team et al., 2023; Bai et al., 2023; OpenAI, 2024) as an automated evaluator for both ID preservation and prompt following.

Human Alignment of Different Reward Model. We first assess how well various automated reward models align with human annotators. To do this, we created a validation set of 200 generated images, covering a diverse range of subjects and prompts. For each image, three expert annotators rated both ID preservation and prompt following on a scale of 0 to 5. The final human score for each criterion is the average of these three ratings, ensuring robustness. For the same set of images, we calculated scores from each automated reward model, using carefully designed prompts for the Vision-Language Models (VLMs). We then measured the agreement between the models' scores and the averaged human scores using both Pearson and Spearman correlation coefficients.

The results present in Table 6. For ID preservation, our proposed DINO-Seg achieves the highest correlation with human annotators (Speraman = 0.75), significantly outperforming both standard vision models like DINO and the best-performing VLM, Gemini-2.5-Pro. For prompt following, HPS-V3 demonstrates the strongest human alignment (Spearman = 0.76). Therefore, we choose DINO-Seg and HPS-V3 as reward models for Customized-GRPO training.

Performance with Different Reward Model Combinations. Following the alignment analysis, we trained Customized-GRPO using various combinations of reward models and evaluated their final performance on DreamBench. The results, detailed in Table 7, reveal several key insights.

First, the choice of reward model is critical to the final generation quality. Using reward models with lower human alignment, such as the CLIP-I, CLIP-T combination, leads to a significant degradation in performance, particularly in Prompt Following (HPS-V3 score drops to 5.24). This highlights that a reward signal misaligned with human preference can actively harm the policy optimization process.

Second, while powerful VLMs like Gemini-2.5-Pro can serve as effective reward models and yield strong results (HPS-V3 of 8.17), they do not surpass our specialized combination. We hypothesize this is because VLMs, while excellent at understanding high-level semantics, may be less sensitive to the fine-grained details crucial for ID Preservation. For instance, they might overlook subtle differences in texture, color shades, or small patches that are easily discernible to humans.

Ultimately, our primary combination of DINO-Seg and HPS-v3, which exhibited the highest human alignment in the previous analysis, also yields the best-balanced performance on the final benchmark.

C.2 Computational Efficiency Analysis

To comprehensively evaluate the computational efficiency of our proposed method, we conduct a wall-clock time analysis using 8 NVIDIA A100 GPUs on the 10k subset of the Syncd dataset. The comparison setup is configured as follows:

- **SFT**: Batch size = 8, trained for 2 epochs.
- **GRPO (Naive & Ours)**: Batch size = 1, Group size = 12, trained for 1 epoch.

The quantitative results regarding training and inference efficiency are summarized in Table 8. Based on these results, we derive two main observations:

Method	Training Paradigm	Iteration Time (s)	Training Time (h)	Sampling Time (s)	Performance (DINO / HPS-v3)
UNO (Base)	SFT	2.32	3	8.1	0.762 / 7.43
Naive-GRPO	Online RL	291.28	80	8.1	0.801 / 7.56
Customized-GRPO	Online RL + SARS + TDW	311.82	80	8.1	**0.812 / 8.32**

Table 8: Wall-clock time analysis and computational efficiency comparison evaluated on 8 NVIDIA A100 GPUs.

Training Efficiency: While RL-based training ($\sim$80h) involves higher computational costs than standard SFT (3h) due to the inherent nature of online exploration and reward evaluation, it yields substantial performance gains across both objectives (DINO score increases from 0.762 to 0.812; HPS-v3 increases from 7.43 to 8.32). Crucially, compared to the Naive-GRPO baseline, our Customized-GRPO incurs **negligible additional overhead**. This demonstrates that our core contributions—Synergy-Aware Reward Shaping (SARS) and Time-Aware Dynamic Weighting (TDW)—are computationally lightweight and do not burden the overall RL training process.

Sampling Efficiency: Our method maintains a sampling latency ($8.1s$) identical to the base UNO model and Naive-GRPO. Since our optimization strictly updates model weights without altering the architecture, it introduces **zero additional cost during inference**.

C.3 Qualitative Results

Figure 9 presents a qualitative comparison of our Customized-GRPO against several state-of-the-art methods to visually substantiate the effectiveness of our approach. The examples highlight our model's superior ability to improve both demands of ID Preservation and Prompt Following. In the first two rows, our method demonstrates the best ID Preservation. For the "dog" subject, our method is the only one that accurately captures the specific facial structure and dense fur texture of the reference dog. Similarly, for the "bowl of blueberries," our model faithfully reproduces the text and the shape.

The following three rows showcase performance on more complex compositional prompts. Our method consistently excels at both ID Preservation and Prompt Following. For instance, in the "vase with a tree" example, our model correctly places the vase within an autumn scene, avoiding the common hallucination of generating a tree growing out of the vase itself. In the final row, it successfully renders both the subject backpack and the specified "blue house" in the background, instead of a blue backpack. Furthermore, a consistent qualitative improvement is observed in the overall aesthetic quality of our generations. Beyond simply satisfying the core objectives, the outputs from Customized-GRPO tend to be more coherent, with better lighting and composition.

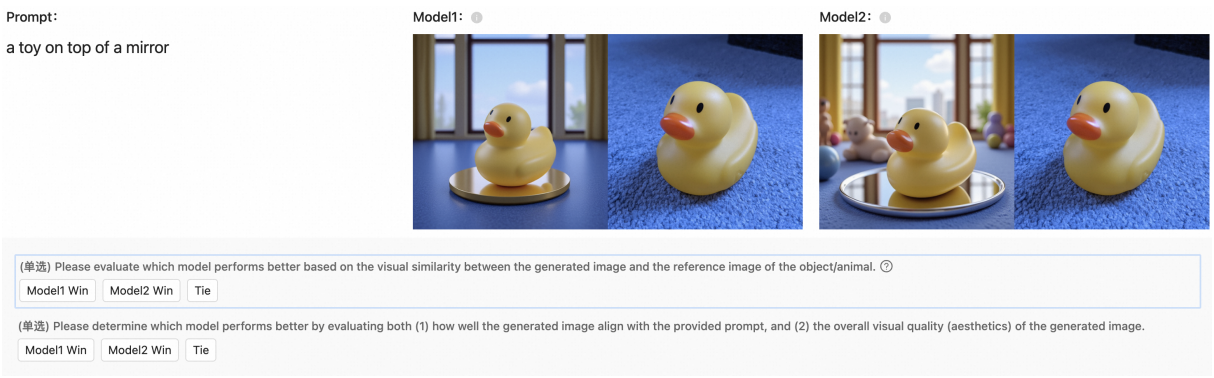

Figure 8: iTAG Interface for human evaluation.

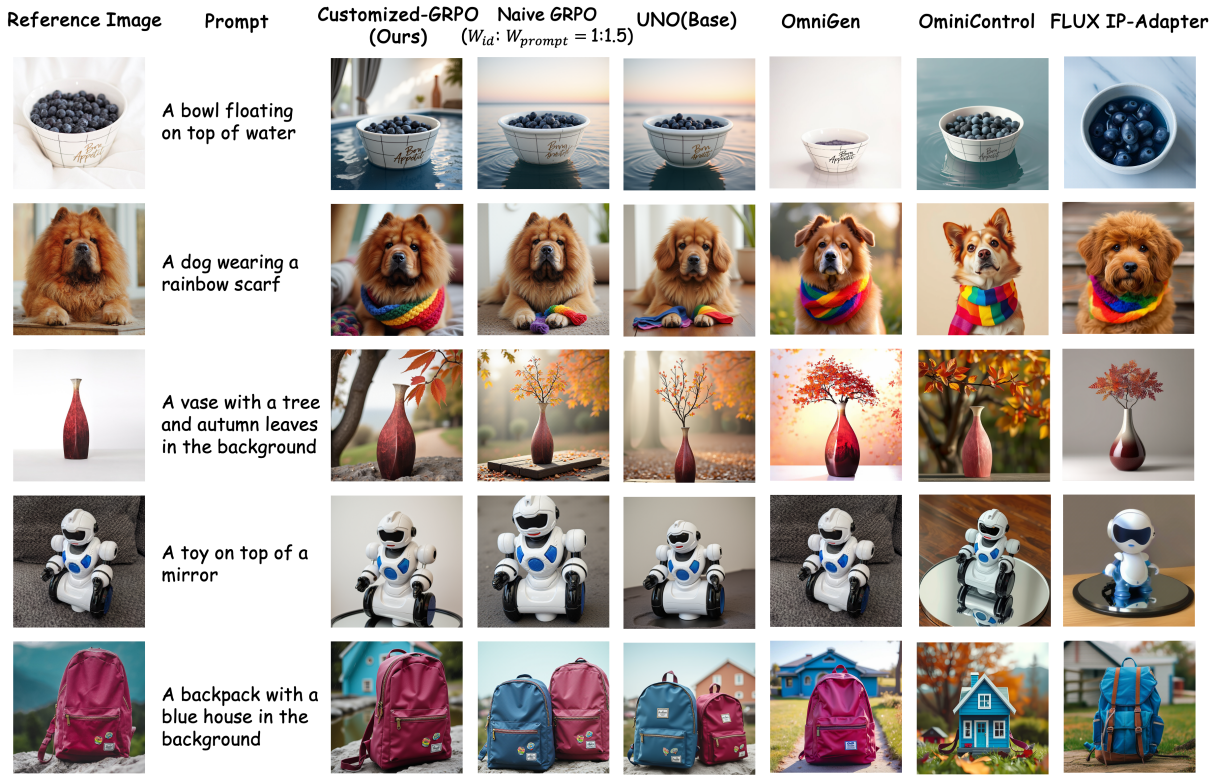

Figure 9: **Qualitative comparison of Customized-GRPO with state-of-the-art baselines.** Our method consistently generates higher-quality images that better balance identity preservation and prompt following.

Figure 10: Comparison of the visualization results of UNO and Customized-GRPO

Figure 11: Comparison of the visualization results of UNO and Customized-GRPO