Do Less, Achieve More: Do We Need Every-Step Optimization for RL Fine-tuning of Diffusion Models?

Supplementary Material

1. Supplementary Related Works

1.1. RL Fine-Tuning in Diffusion Models

Existing diffusion models [4, 5, 24, 58, 61] primarily approximate the data distribution through denoising reconstruction loss. However, this training approach struggles to capture high-level metrics such as semantic consistency, aesthetic preferences, and user subjective judgments [2, 20, 22, 56]. To enhance preference alignment, recent studies have introduced RL fine-tuning, leveraging explicit reward signals from human feedback, reward models, or preference predictors. This design shifts the optimization of generation from a reconstruction-based to a reward-based paradigm. Such RL fine-tuning methods typically treat the diffusion denoising process as a sequential decision-making process and optimize the rewards obtained by the final generated image using policy gradient techniques.

1.2. Sparse Rewards and the Reward Hacking

RL-based fine-tuning of diffusion models has garnered increasing attention in recent years. However, existing studies consistently point to two fundamental challenges: sparse rewards and reward hacking.

The issue of sparse rewards arises because the evaluation signals are computed on the fully denoised final image. These signals include CLIP-based semantic consistency scores, aesthetic-quality predictors, and human-preference models. As a result, intermediate diffusion states cannot receive effective feedback, forcing the policy to explore a vast action space without guidance. Classical RL methods for generative models, such as DDPO and DPOK, explicitly highlight that sparse and delayed rewards lead to unstable policy learning and high gradient variance.

At the same time, reward hacking has been repeatedly observed when using proxy rewards to finetune generative models. Similar to phenomena in RLHF for language models, reward model-based RL for diffusion models often tends to 'exploit loopholes in the reward model' rather than genuinely improving image quality. Existing research shows that policies may overfit specific objectives to achieve higher reward scores, such as generating over-saturated colors, exaggerated object features, or unnatural layouts. Similar mis-optimization behaviors have been explicitly discussed in DDPO, Diffusion-PPO, and subsequent reward-guided sampling methods. Although the model achieves higher numerical rewards, the generated images deviate from human preferences and even deteriorate in terms of semantics and aesthetics. One of the most significant drawbacks of reward hacking is its severe impact on output diversity. Overall, existing research indicates that sparse rewards limit the stability of RL fine-tuning for diffusion models, while reward hacking undermines the correctness of model optimization.

Table 3. Quantitative comparisons with SoTA. All metrics are obtained with SDv15 as backbone, Pick-a-pic as prompt set, and PickScore as training reward.

Method	Preference				Fidelity			Diversity			Richness			#Top2
	PS↑	AES↑	IR↑	HPS↑	FID↓	CLIP↑	iFS↑	LPIPS↑	IS↑	TCE↑	BRI↓	NIQE↓	SE↑	
SDv15 [40]	20.48	5.412	0.181	0.262	-	0.243	-	**0.654**	**23.79**	38.05	18.66	5.401	11.26	3
Diff-DPO [53]	20.87	5.551	0.443	0.271	109.1	**0.244**	0.795	0.639	22.77	39.20	15.12	4.463	11.02	1
Diff-KTO [27]	20.83	5.585	**0.599**	0.272	101.3	0.240	0.801	0.634	22.70	39.12	26.25	4.361	11.22	1
SPO [28]	20.76	5.613	0.282	0.218	**78.76**	0.241	**0.855**	0.649	23.18	39.32	25.67	4.103	11.09	3
DDPO [1]	21.79	5.704	0.196	0.212	147.5	0.242	0.539	0.629	20.01	39.18	12.46	4.328	11.74	1
DPOK [7]	20.97	5.661	0.582	0.272	99.30	0.242	0.820	0.641	22.52	39.35	12.28	4.608	11.23	3
TDPO [65]	22.94	5.991	0.504	0.273	128.66	0.265	0.820	0.635	21.99	39.09	12.96	4.077	11.59	1
Ours	**23.01**	**6.071**	0.542	**0.278**	85.37	0.243	0.846	0.652	23.73	**39.37**	**12.20**	**4.019**	**11.91**	**12**

2. Experiment List in Our Paper

To help readers quickly grasp the extensive experiments conducted in this work, we summarize the full list of experiments below.

- **(1) Visualization Experiments.** See Fig. 1. This experiment provides a solid justification for the motivation of this work.
- **(2) Reward Backfilling Validation.** See Fig. 3. This experiment demonstrates the effectiveness of the reward backfilling mechanism. Prepares the ground for the subsequent discussion that, despite stabilizing training, it can lead to attribution mismatch.
- **(3.1) Quality–Cost** Dual Optimization. See Tab. 1. By comparing the **runtime** and **performance** with state-of-the-art methods in the field. We validate our claim that the proposed approach can improve generation quality while reducing computational cost.
- **(3.2) Ensemble Experiments.** See Tab. 1. This table further shows the performance gains obtained by incorporating our RL-based enhancement plug-in into existing approaches.
- **(4) Dataset Switching Experiment.** See Fig. 4. To verify that the superiority of our method is not affected by differences in datasets or task difficulty, we conduct experiments across multiple datasets.
- **(5) Backbone Switching** Experiment. See Fig. 5. We provide this experiment to verify that the advantages of our method are not affected by replacing the backbone to be fine-tuned.
- **(6) Reward Switching** Experiment. See Fig. 6 and 4. We verify that the advantages of our method are not affected by replacing the reward model.
- **(7) Ablation Experiment.** See Fig. 7. We present the impact of each module in our method on both performance and computational cost.
- **(8) Generation Distribution Visualization.** See Fig. 8. In this experiment, we demonstrate the effectiveness of our method in mitigating reward hacking and preventing excessive diversity loss during the fine-tuning of diffusion models.
- **(9) Semantic Alignment Visual Experiment.** See Fig. 9. This experiment verifies the semantic alignment capability of our method.
- **(10) Complex-Prompt Generalization.** See Fig. 10. This experiment verifies our method's ability to handle complex prompts.
- **(11) Comprehensive Generation Comparison.** See Tab.3. We conduct a comprehensive evaluation across several key aspects of generative performance:**Preference**, **Fidelity**, **Diversity**, and **Richness** each assessed using multiple metrics, demonstrating the broad effectiveness of our method.
- **(12) Diversity Visual Experiment.** See Fig. 13. This experiment compares our method with state-of-the-art approaches in terms of diversity. Validating our claim that it mitigates reward hacking and prevents excessive diversity degradation during fine-tuning.
- **(13) Human Evaluation Experiment.** See Fig. 12. We present a human evaluation experiment assessing our method against the baseline methods.

3. Supplementary Experiments

3.1. Why AdaScope Improves Both Quality and Efficiency ?

Computational Savings: Our method reduces training computational costs by adaptively pruning uninformative early denoising samples and late-stage steps where returns have saturated. In the early stage of denoising, the image's semantic structure has not yet formed, leading to insufficient training signals. In the late stage, when the latent representation has largely stabilized and reward optimization has reached diminishing returns, further optimization becomes meaningless. By employing this adaptive pruning strategy, we effectively reduce the computational resources required for training.

Quality Improvement: In the early denoising stage, the sample's semantic structure has not been shaped. Meanwhile, reward reshaping, which propagates the final reward back to all previous steps to stabilize training, leads to a severe action-reward attribution mismatch. Since the reward only appears at the final step but is assigned to all preceding actions, the attribution bias grows larger in the earlier steps. Consequently, using these trajectory segments for RL training may cause the policy to make erroneous advantage estimates based on these semantically vague and ineffective states. On the other hand, in the final denoising stages, the marginal gain from RL fine-tuning diminishes, and continued training increases the risk of overfitting to non-critical image details.

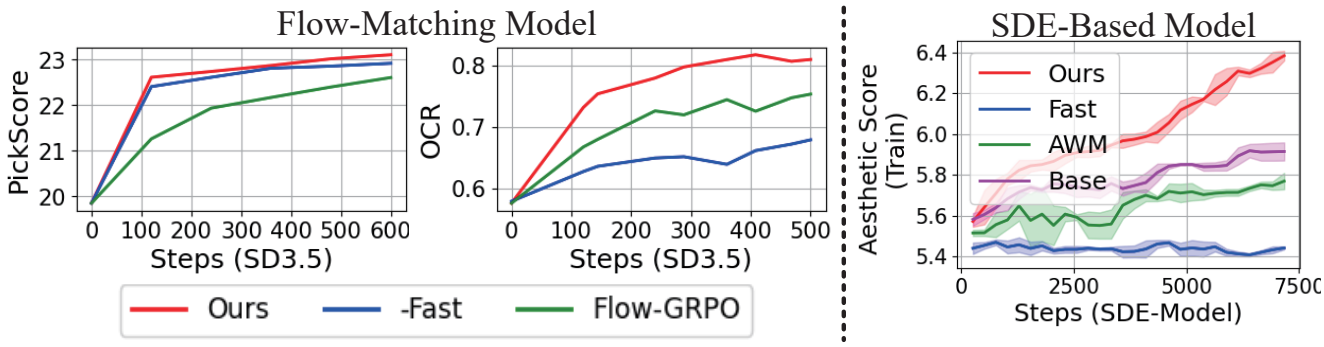

Figure 11. Left: Results on Flow-Matching Model. Right: on SDE Model.

Overall: Our method eliminates samples that are not only computationally redundant but also low-quality, thereby improving policy in RL fine-tuning. Because these samples are pruned, we simultaneously reduce computational costs and enhance generation quality, thereby achieving a dual optimization of both quality and efficiency.

3.2. More Quantitative Results

To quantitatively evaluate the generative performance as extensively as possible, we introduce 4 evaluating dimensions with 13 distinct metrics, including: **Aesthetic Preference:** AES, PS, IR, and HPSv2 [55]. **Image Fidelity:** ClipScore [37] (Clip), Fréchet Inception Distance (FID) [14], and improved F1 Score (iFS) [26]. **Generative Diversity:** LPIPS [64], TCE [21], and Inception Score [44] (IS). **Compositional Richness:** NIQE [33], BRISQUE (BRI) [32], and Spectual Entropy (SE) [29].

As shown in Tab. 3, our method consistently achieves superior overall performance among 13 metrics, with **12** top2 effectiveness.

3.3. Comparisons with Flow-matching-native RL baselines.

Theoretical Analysis. Flow-GRPO-Fast adopts a hybrid SDE–ODE optimization strategy, where short SDE segments are randomly inserted into ODE trajectories. This design presents two key limitations. First, its SDE–ODE switching mechanism is tailored for flow-matching formulations and does not naturally generalize to standard SDE-based diffusion models (e.g., SD1.5, SDXL), which remain competitive in practice. Second, the selection of SDE intervals is heuristic, relying on random sampling without considering denoising dynamics or reward-driven signals, thus lacking principled guidance for identifying informative training regions. In contrast, AdaScope introduces an adaptive interval selection mechanism based on structural transitions and reward evolution, supported by theoretical analysis (Sec. 3.4, Theorems 1–2). This enables more effective identification of high-quality training intervals and leads to a model-agnostic, plug-and-play framework applicable across different diffusion paradigms.

Empirical Results. We validate the above differences through extensive experiments. On flow-matching models (Fig. 11-Left), using SD3.5 as the backbone with PickScore and OCR rewards, AdaScope consistently outperforms existing methods, benefiting from its adaptive design. On classical SDE-based models (Fig. 11-Right), we align all hyperparameters and use η to simulate SDE–ODE transitions for Fast and AWM. Under this setting, both methods exhibit limited effectiveness, further indicating their restricted applicability. In contrast, AdaScope maintains stable and consistent improvements due to its backbone-agnostic design.

3.4. Visual Impression of Generation Diversity

As shown in Fig. 13, we comprehensively evaluate the generative diversity in visual quality. It can be observed that our method consistently exhibits superior diversity in posture, color, form, and style while maintaining the prompt-image alignment.

3.5. Subjective Study

As shown in Fig. 12, we evaluate the generated images subjectively with Human and Vision Language Model (VLM). All images were generated by the fine-tuned SDv15 model from HPSv2 prompts using different methods. The evaluation includes five dimensions: Structural Faithfulness (SF), Aesthetic Appeal (AA), Fine-grained Detail (FD), Semantic Alignment (SA),

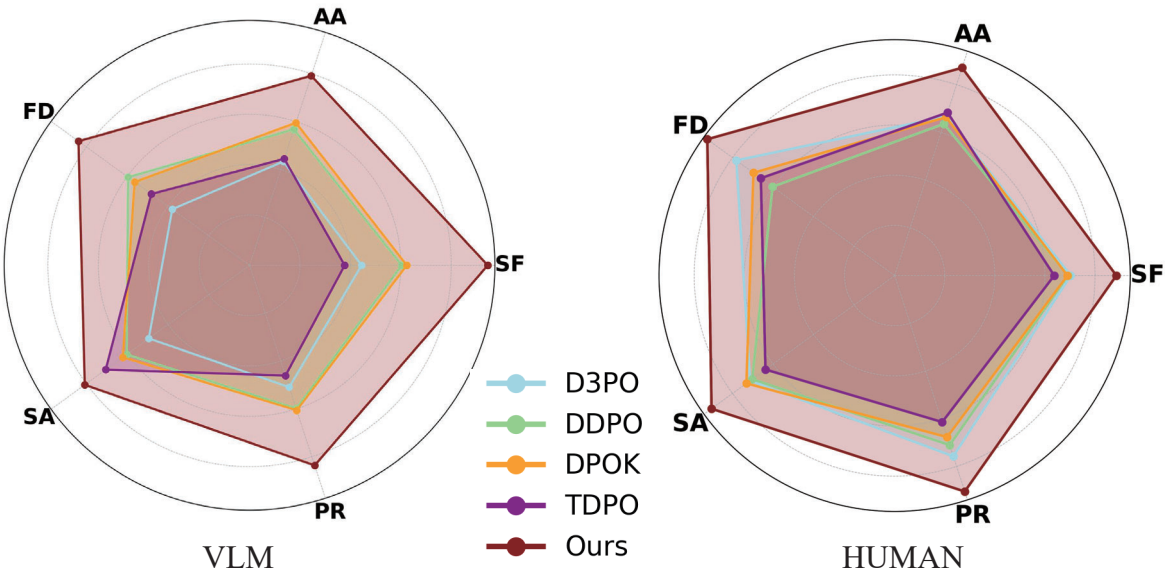

Figure 12. Subjective Evaluation with Human and VLM.

and Prompt Responsiveness (PR). They are rated independently by human evaluators and ChatGPT. The results demonstrate the superior generative quality of our method.

3.6. Showcase Prompt Table

Considering that the specific semantics of the prompts can substantially affect the assessment of the actual quality of generated images, it is necessary to assess the performance superiority of our method based on both images and their corresponding prompts. Therefore, in Tab. 4, we provide a detailed list of the prompts that were not explicitly described in the main text.

4. Proof

4.1. Proof of Theorem 1.

This is proved in the sec.5 of *Reverse-time diffusion equation models* by Anderson.

4.2. Proof of Theorem 2.

We do the direct calculation:

Table 4. Detailed prompts used for generated images in Fig. 13.

Image	**Prompt**
Row 1, Col 1	A young girl standing on a rooftop, blowing dandelions that transform into glowing comets, shooting across the night sky, dreamy fantasy artwork.
Row 1, Col 2	A boy lying on the grass in a field, listening to music with glowing headphones, fireflies surrounding him.
Row 2, Col 1	A little girl painting a rainbow bridge from the classroom window into the sky, playful magical fairytale art, hopeful and inspiring.
Row 2, Col 2	Five birds in the park.
Row 3, Col 1	Four roses.
Row 3, Col 2	A rabbit near a pool.

Assumptions (forward diffusion / Markov Gaussian):

$$x_0 \sim \mathcal{N}(0, \Sigma), \epsilon_t \sim \mathcal{N}(0, I), \epsilon' \sim \mathcal{N}(0, I)$$

all independent, and

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon_t$$

$$x_{t+\tau} \,|\, x_t = \sqrt{\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, x_t + \sqrt{1-\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, \epsilon'.$$

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon_t, \qquad x_{t+\tau} = \sqrt{\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, x_t + \sqrt{1-\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, \epsilon'.$$

1) Expand $x_{t+\tau}$ in terms of $(x_0, \epsilon_t, \epsilon')$:

$$\begin{aligned} x_{t+\tau} &= \sqrt{\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\Big(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon_t\Big) + \sqrt{1-\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, \epsilon' \\ &= \sqrt{\bar{\alpha}_{t+\tau}}\, x_0 + \sqrt{\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, \sqrt{1-\bar{\alpha}_t}\, \epsilon_t + \sqrt{1-\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, \epsilon'. \end{aligned}$$

2) Cross-covariance $\mathrm{Cov}(x_t, x_{t+\tau})$: Using independence and $\mathrm{Cov}(x_0) = \Sigma,\ \mathrm{Cov}(\epsilon_t) = I,\ \mathrm{Cov}(\epsilon') = I$,

$$\begin{aligned} \mathrm{Cov}(x_t, x_{t+\tau}) &= \mathrm{Cov}\Big(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon_t,\ \sqrt{\bar{\alpha}_{t+\tau}}\, x_0 + \sqrt{\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, \sqrt{1-\bar{\alpha}_t}\, \epsilon_t + \sqrt{1-\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, \epsilon'\Big) \\ &= \sqrt{\bar{\alpha}_t \bar{\alpha}_{t+\tau}}\, \mathrm{Cov}(x_0, x_0) + \sqrt{1-\bar{\alpha}_t}\, \sqrt{\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, \sqrt{1-\bar{\alpha}_t}\, \mathrm{Cov}(\epsilon_t, \epsilon_t) \\ &= \sqrt{\bar{\alpha}_t \bar{\alpha}_{t+\tau}}\, \Sigma + \sqrt{\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, (1-\bar{\alpha}_t)\, I. \end{aligned}$$

Therefore, componentwise,

$$\mathrm{Cov}\Big(x_t^{(i)}, x_{t+\tau}^{(j)}\Big) = \sqrt{\bar{\alpha}_t \bar{\alpha}_{t+\tau}}\, \Sigma_{ij} + \sqrt{\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, (1-\bar{\alpha}_t)\, \delta_{ij}.$$

3) Marginal variances at each time:

$$\mathrm{Var}\Big(x_t^{(i)}\Big) = \mathrm{Var}\Big(\sqrt{\bar{\alpha}_t}\, x_0^{(i)} + \sqrt{1-\bar{\alpha}_t}\, \epsilon_t^{(i)}\Big) = \bar{\alpha}_t\, \Sigma_{ii} + (1-\bar{\alpha}_t),$$

$$\mathrm{Var}\Big(x_{t+\tau}^{(j)}\Big) = \mathrm{Var}\Big(\sqrt{\bar{\alpha}_{t+\tau}}\, x_0^{(j)} + \sqrt{1-\bar{\alpha}_{t+\tau}}\, \tilde{\epsilon}^{(j)}\Big) = \bar{\alpha}_{t+\tau}\, \Sigma_{jj} + (1-\bar{\alpha}_{t+\tau}),$$

(where $\tilde{\epsilon}$ is standard normal noise independent of x_0.)

4) Correlation:

$$\begin{aligned} \mathrm{Corr}\Big(x_t^{(i)}, x_{t+\tau}^{(j)}\Big) &= \frac{\mathrm{Cov}\Big(x_t^{(i)}, x_{t+\tau}^{(j)}\Big)}{\sqrt{\mathrm{Var}(x_t^{(i)})\, \mathrm{Var}(x_{t+\tau}^{(j)})}} \\ &= \frac{\sqrt{\bar{\alpha}_{t+\tau}\bar{\alpha}_t}\, \Sigma_{ij} + \sqrt{\frac{\bar{\alpha}_{t+\tau}}{\bar{\alpha}_t}}\, (1-\bar{\alpha}_t)\, \delta_{ij}}{\sqrt{\big(\bar{\alpha}_t \Sigma_{ii} + (1-\bar{\alpha}_t)\big)\big(\bar{\alpha}_{t+\tau} \Sigma_{jj} + (1-\bar{\alpha}_{t+\tau})\big)}}. \end{aligned}$$

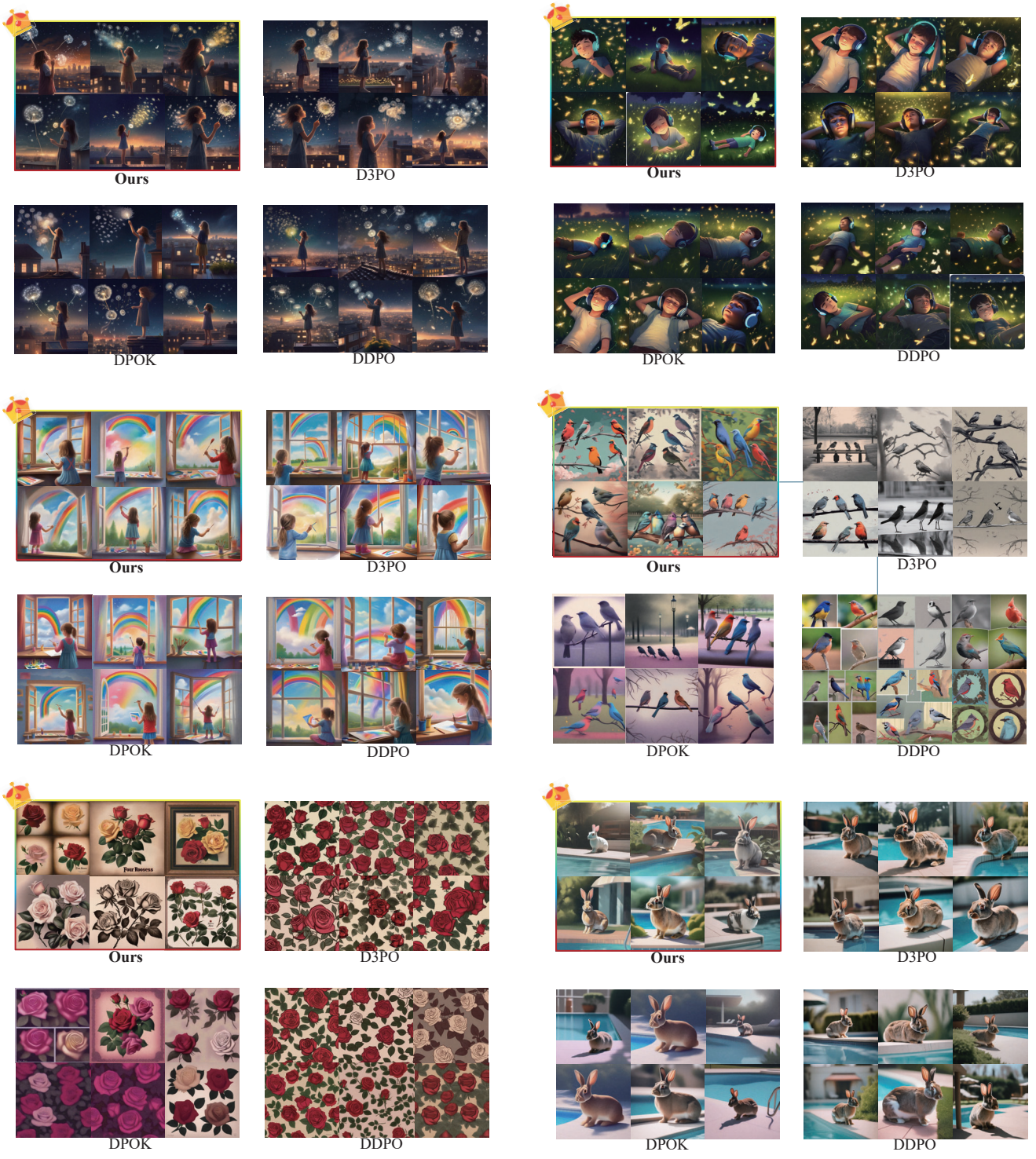

Figure 13. **Diversity Evaluation:** Our method demonstrates the highest level of variation under these prompts, producing outputs with a wide range of artistic styles, figure posture, object positioning, and background colors. In contrast, D3PO predominantly generates grayscale backgrounds or same posture, DPOK consistently incorporates purple tones into its visual style, and DDPO tends to produce collage-like compositions within a single image.