RebRL: Reinforcing Discrete Visual Diffusion Models with Rebalanced Timestep Credits

Supplementary Material

A. Derivation Details

A.1. The Approximation

As discussed in the foundational theory of DDMs [19], the exact conditional log-likelihood $\log \pi_\theta(o|o^{t_j}, \mathbf{c})$ is intractable to compute directly. To address this, the approximation in Eq. (7) is grounded in the Evidence Lower Bound (ELBO) formulation:

$$\ell_{\pi_\theta}(o^{t_j}, o|\mathbf{c}) = \frac{1}{t_j} \sum_{k=1}^{|o|} \delta(o_k^{t_j}, \mathbf{m}) \cdot \log \pi_\theta(o_k|o^{t_j}, \mathbf{c}) \quad (13)$$

The exact log-likelihood is typically approximated by its ELBO [19, 37]:

$$\ell_{\pi_\theta}(o^{t_j}, o|\mathbf{c}) \leq \log \pi_\theta(o|o^{t_j}, \mathbf{c}). \quad (14)$$

Consequently, for any specific corruption level t_j, the term $\ell_{\pi_\theta}(o^{t_j}, o|\mathbf{c})$ serves as a tractable local estimator for the log-likelihood contribution. This justifies utilizing the reconstruction loss as a proxy for the intractable log-likelihood in our gradient update.

A.2. Temporal Rebalancing Factor

In Section 4.2, we proposed using polynomial rebalancing factors $\lambda(t_j) = t_j^p$ with orders $p = 1$ and $p = 2$. Here, we provide a theoretical justification for why higher-order terms ($p \geq 3$) were not employed, focusing on the trade-off between gradient scale balance and preservation for fine-grained refinement. To avoid confusion with the cumulative gradient scale (denoted as $w(\Delta t_j)$ in the main text), we focus here on the per-step gradient scale, denoted as α_j.

Recall that the standard DDM loss effectively assigns a scale $\alpha_j \propto 1/t_j$ to the policy gradient of each masked token at reverse step j. When applying a rebalancing factor $\lambda(t_j) = t_j^p$, the per-step scale becomes $\alpha'_j(p) \propto t_j^p \cdot (1/t_j) = t_j^{p-1}$. Based on this derivation, 1-Order rebalance transforms the original scale into a constant contribution ($\alpha'_j \propto 1$), leaving the cumulative gradient scale w growing linearly, as visualized in Fig. 7 (c). 2-Order rebalance introduces a linear decay ($\alpha'_j \propto t_j$). This suppression is desirable because it alleviates the cumulative nature: since tokens unmasked later participate in more steps, reducing their per-step scale helps flatten the cumulative scale curve as shown in Fig. 7 (c).

However, higher orders ($p \geq 3$) lead to gradient vanishing. For $p = 3$, the scale decays quadratically ($\alpha'_j \propto t_j^2$). As generation reaches the refinement stage (e.g., $t_j < 0.1$), the policy gradient is suppressed by two orders of magnitude (< 0.01). This suppression is visually demonstrated in Fig. 9 (c). Unlike the 2-Order case, the 3-Order rebalancing diminishes the late-timestep contributions to negligible levels, confirming the risk of rendering the model unable to learn fine-grained details. This issue is also confirmed empirically in Tab. 3. While the 3-Order rebalancing achieves a decent Overall score (0.80), it notably underperforms the 2-Order case (0.83). Specifically, performance drops in fine-grained attributes like Attribute Binding (0.70 *vs*. 0.76). Therefore, we determine that $p = 2$ represents the optimal setup, providing appropriate balancing without sacrificing the necessary optimization for final refinement.

Table 3. **Performance comparison of different temporal rebalancing strategies on GenEval.**

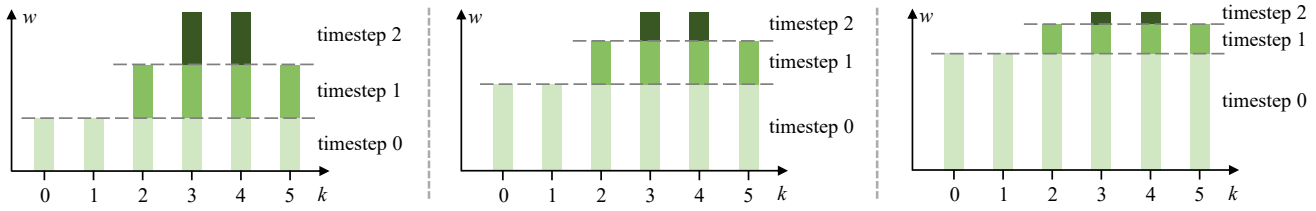
Model	GenEval↑						
	Single.	Two.	Count.	Colors.	Pos.	Attr.	Overall
MMaDA*	0.96	0.76	0.60	0.85	0.61	0.67	0.74
w/ 1-Order.	0.96	0.84	0.77	0.85	0.77	0.73	0.82
w/ 2-Order.	0.98	0.83	0.76	0.88	0.77	0.76	**0.83**
w/ 3-Order.	0.97	0.79	0.73	0.86	0.73	0.70	0.80

B. Another Perspective of Exploration Potential

In the main text, we introduced the concept of "Exploration Potential" to characterize the varying importance of different generation timesteps. Here, we provide a complementary empirical validation of this concept by analyzing the Standard Deviation of the reward score (HPSv3) across the generation process.

The variance of the reward score can also serves as a quantitative representation for exploration potential [11]. A high standard deviation at a specific step implies that the stochastic sampling choices made at that step lead to diverse outcomes with significantly varying quality, indicating a vast search space where decisions are critical for determining image quality. Conversely, a low standard deviation suggests that the generation has converged to a stable state, representing a refinement phase with limited impact on semantic alignment.

We visualize this relationship in Figure 10, plotting the HPS Standard Deviation alongside the Cumulative Gradient Scale. The standard deviation peaks significantly in the initial phase (Steps 1-3) and decays monotonically as generation progresses, confirming that early, high-mask-ratio steps



(a) w/ 1-Order re-balance (b) w/ 2-Order re-balance (c) w/ 3-Order re-balance

Figure 9. Comparison of cumulative gradient scales under different temporal rebalancing orders.

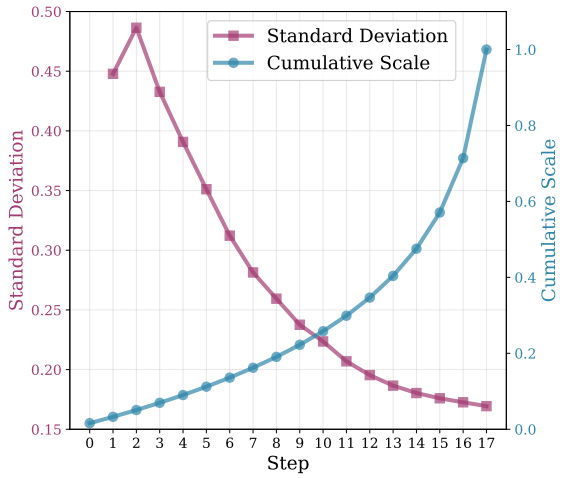


Figure 10. **Analysis of Exploration Potential via Reward Variance.** The crossing curves highlight the severe misalignment: the cumulative gradient scale is minimal where exploration potential is highest.

constitute the most critical stage for exploration. Crucially, the plot reveals that the Cumulative Gradient Scale follows a trend exactly opposite to the Standard Deviation. This provides compelling statistical evidence for the misalignment identified in our work—the GRPO objective without rebalance concentrates the majority of credit on steps where the exploration potential is minimal.

C. Implementation Details

C.1. Reward Function

The reward functions used in our reinforcement learning experiments include:

- **GenEval Score**: We utilize the GenEval score as a reward to enhance compositional generation capabilities, following the methodology established in Flow-GRPO [17].
- **HPSv3**: We employ the HPSv3 model [21] to assess both visual quality and text–image alignment, optimizing for human preference.

C.2. Evaluation

Generation Configuration. For standard image generation, the sampler decodes a sequence of 1024 tokens (which represents an image with resolution 512×512) with 32 sampling steps, and is equipped with classifier-free guidance at a scale of 3.5, consistent with the original MMaDA configuration. The results reported in main text are all obtained using Token-level rebalancing illustrated in Sec. 4.3.

Benchmark Protocols.

- **GenEval**: The evaluation for compositional image generation is conducted on the standard GenEval benchmark, which consists of 553 evaluation prompts.
- **Human Preference Alignment**: For HPSv3, DeQA, and ImageReward, we utilize test prompts from the HPDv2 test set. Following the protocol in TempFlow-GRPO [11], we increase the sampling steps to 64 for high-fidelity generation. It is important to note that the scores for DeQA and ImageReward are also evaluated on HPDv2 test images generated by the model trained with HPSv3 reward; no separate training was conducted specifically for DeQA or ImageReward.

C.3. Data Usage

GenEval. For the GenEval experiments, we utilize instruction tuning data from Blip3-o (which is distilled from GPT-4o) for SFT. We train on this dataset for $\sim$ 1000 steps with global batch size 128, with a learning rate of $3e^{-6}$. We have denoted the corresponding results with explicit SFT mark in the tables. For GRPO, we utilize a dataset of 50,000 randomly composed prompts, consistent with the training data used in Flow-GRPO [18].

Human Preference Alignment. It is noted that experiments related to Human Preference Alignment did not undergo this SFT stage. For HPS optimization, we employ the HPDv2 dataset, following the data usage protocols referenced in TempFlow-GRPO [11].

D. More Results

We provide additional qualitative comparisons in this appendix to complement the results presented in the main text. Fig. 11 visualizes the performance of RebRL on prompts designed to test specific spatial and attribute control. Compared to the baselines, our method consistently excels in attribute binding and positional accuracy. For instance, in the "pink dining table and a black sandwich" prompt (row 2), RebRL is the only method to correctly adhere to both

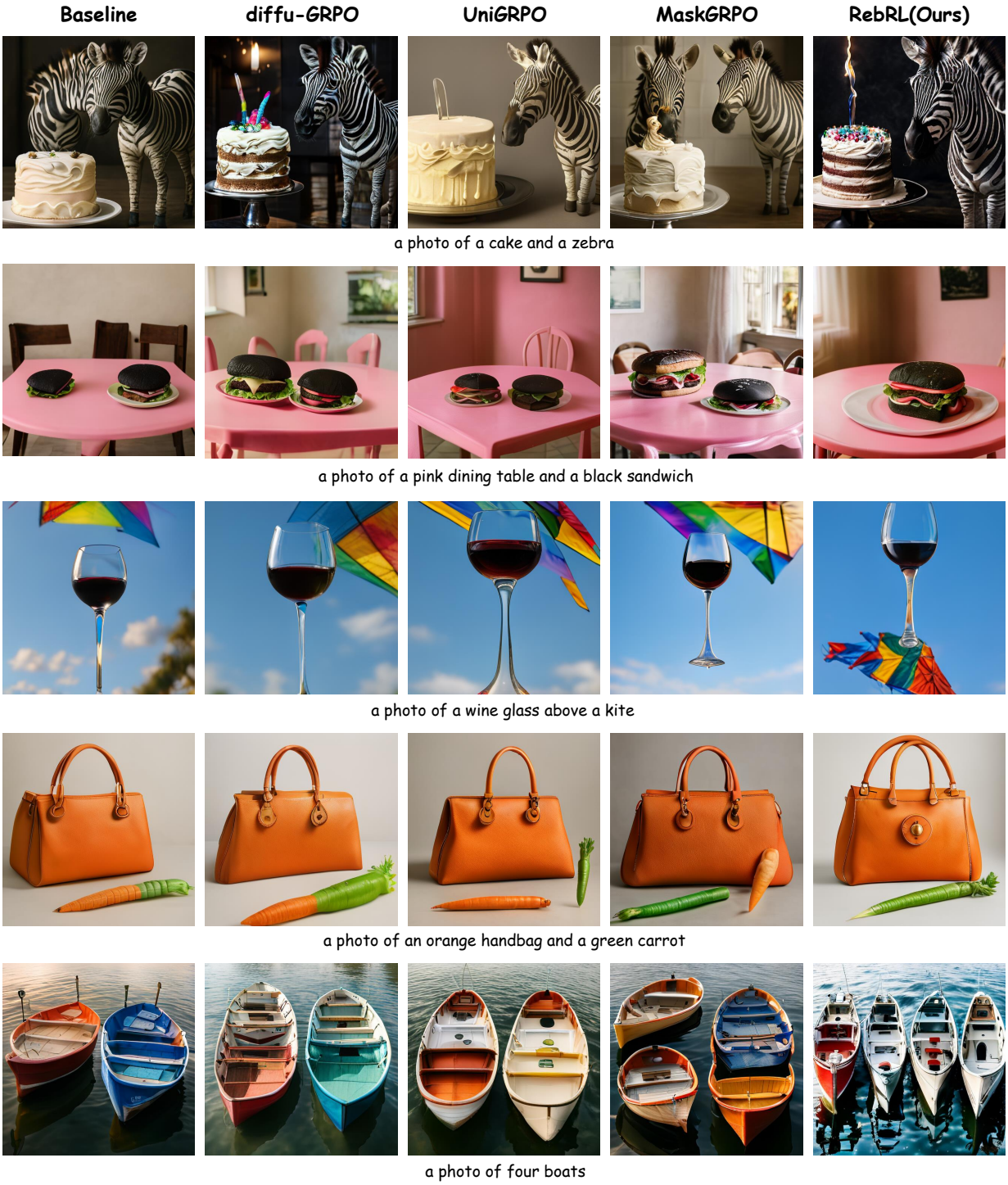


Figure 11. **Qualitative comparison of compositional image generation.** Our method consistently renders more accurate position relationship and capacity of attribute binding.

the specific color requirement and the object placement. Fig. 12 further validates our gains on human preference score, focusing on aesthetic quality and complex concept adherence. RebRL demonstrates superior performance in rendering both natural and conceptual subjects. This is evidenced by the highly detailed textures *e.g.* in the "woolly mammoths" and "dragon" prompts (rows 1 and 2). The overall aesthetic quality and visual impact confirm that our framework leads to a stable and generalized policy optimization that aligns strongly with human preferences.

Figure 12. **Qualitative comparison of human preference alignments.** Our method consistently demonstrates finer details and stronger prompt adherence compared to the baselines.