E-GRPO: High Entropy Steps Drive Effective Reinforcement Learning for Flow Models

Supplementary Material

A. Ablation Study on the Entropy Threshold τ

In our main paper, we introduce an adaptive entropy threshold τ to separate timesteps into high-entropy and low-entropy groups. During sampling, consecutive low-entropy steps are merged until their entropy reaches the threshold. This threshold serves as a critical hyperparameter in our entropy-driven step-merging strategy. To claim the effectiveness of the proposed method and assess its sensitivity to τ, we conducted a series of experiments with different threshold values. Specifically, we trained E-GRPO with τ set to 0 (meaning all steps are treated equally and no merging ooccurs), 1.8, 2.0, 2.2 (our default setting), and 2.6 under the HPS reward configuration. The results are summarized in Table 4.

As shown in Table 4, the model behaves noticeably differently under varying threshold values. As τ increases, the achievable HPS score also improves, indicating the effectiveness of entropy as a guidance signal during training. However, when τ becomes excessively large, a long sequence of steps may be merged, occasionally combining steps that still contain useful entropy or gradient information. This leads to overly coarse updates and, consequently, a slight degradation in performance. Notably, our default choice of $\tau = 2.2$ strikes an effective balance between leveraging entropy for guidance and avoiding excessive merging, yielding the best overall performance in our experiments.

B. Additional Visualizations

B.1. More Quality Results

To further demonstrate the superiority of our proposed E-GRPO, we provide additional qualitative comparisons with baseline methods in Figure 6 and Figure 7. As illustrated in these figures, E-GRPO consistently produces results that are more faithful to the text prompts. For example, under the prompt " An award-winning portrait of a lemon in a muted, space age style reminiscent of the 1930s." E-GRPO successfully generates a portrait that combines a space-age aesthetic with the intended compositional structure. Likewise, for the prompt "A lot of buildings on each side of the road, with a very curvy road in the middle." our method captures the "curvy" characteristic more accurately and achieves higher aesthetic quality compared with baseline methods. These results further validate that by focusing on high-entropy steps, E-GRPO enables more effective exploration and better alignment with complex human preferences.

Table 4. **Ablation study on the entropy threshold τ.** Results are reported under the HPS reward setting. A threshold of $\tau = 0$ corresponds to the baseline without step merging. Our default choice ($\tau = 2.2$) achieves the overall best performance. Best results in each column are highlighted in **bold**.

Threshold (τ)	HPS	CLIP	PickScore	ImageScore
0 (No Merging)	0.384	0.349	0.230	1.297
1.8	0.383	0.352	0.232	1.293
2.0	0.384	0.344	0.231	1.269
2.2 (Ours)	**0.391**	**0.355**	**0.233**	**1.324**
2.6	0.388	0.355	0.233	1.320

B.2. Failure Cases

Despite the robustness of E-GRPO, we observe several recurring failure patterns when handling challenging prompts.
Reward Hacking. As discussed in the main paper, using only the HPS reward tends to produce overly saturated images, making the CLIP reward necessary as a counterbalance. Nevertheless, reward hacking still occurs in some cases. For instance, in the prompts shown in Figure 8, such as "A jellyfish sleeping in a space station pod. " and "The image depicts alien flowers and plants surrounded by visceral exoskeletal formations in front of mythical mountains with dramatic contrast lighting, created with surreal hyper detailing in a 3D render. ", the model occasionally introduces human faces or humanoid shapes that should not be present. These artifacts reflect the model's tendency to exploit biases in the reward models, a limitation that is common across many RL-based training frameworks. Improving reward model reliability will be crucial for advancing RL in visual generation.

Overall, these observations highlight several key challenges faced by RL-based visual generation systems. Future research may explore solutions guided by these identified limitations.

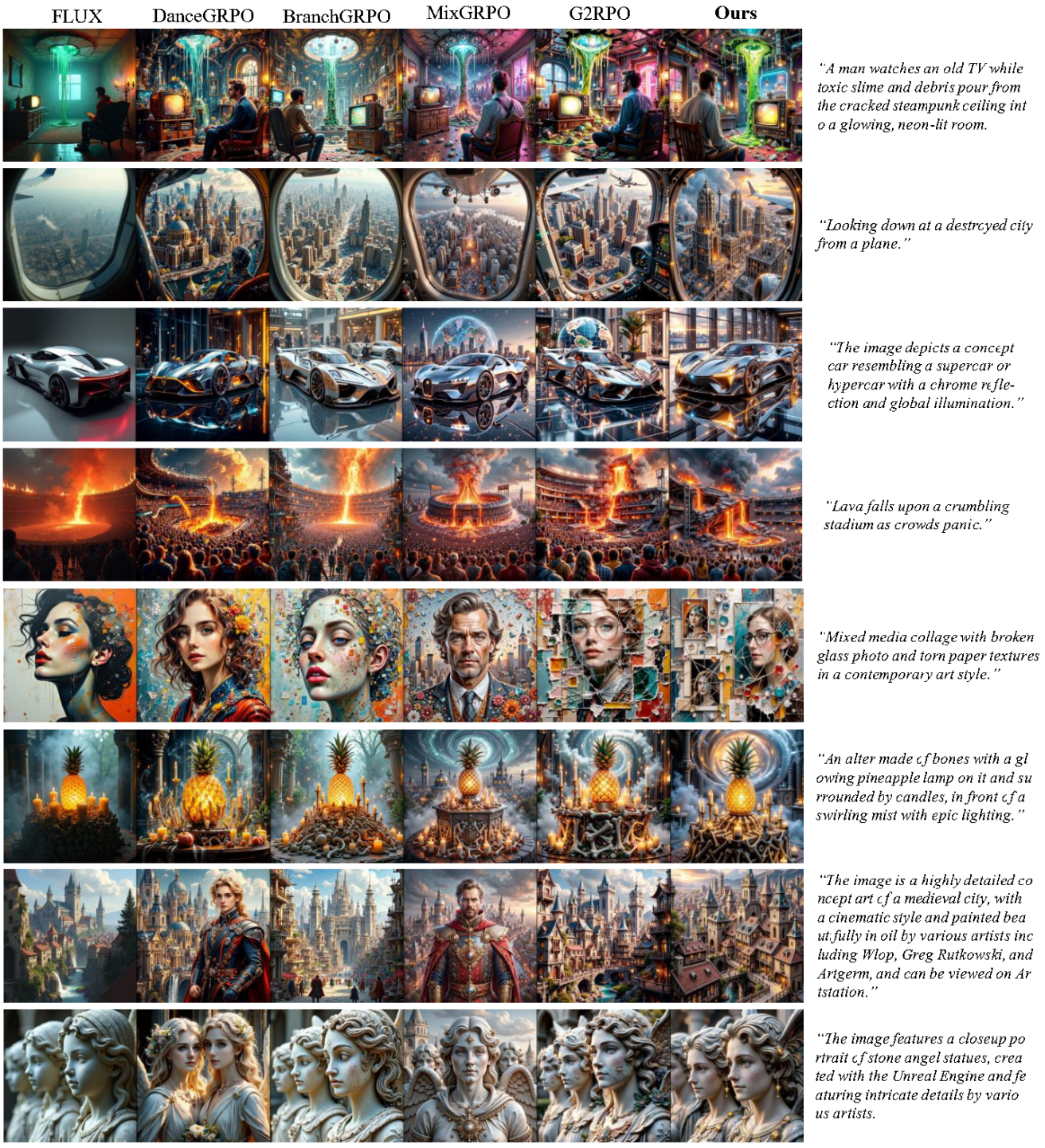

Figure 6. **Additional visualization comparisons between E-GRPO and other baseline methods.**

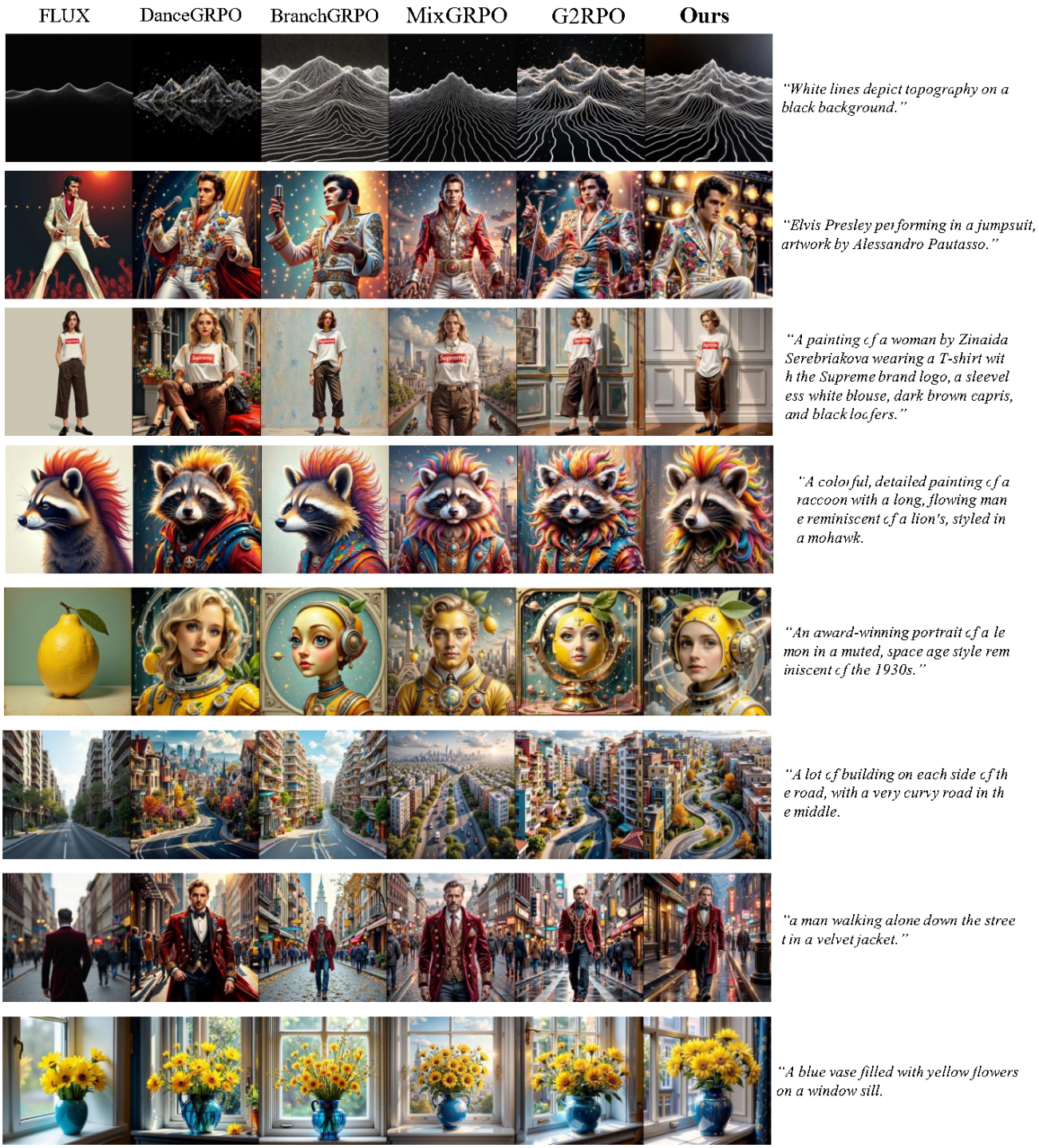

Figure 7. **Additional visualization comparisons between E-GRPO and other baseline methods.**

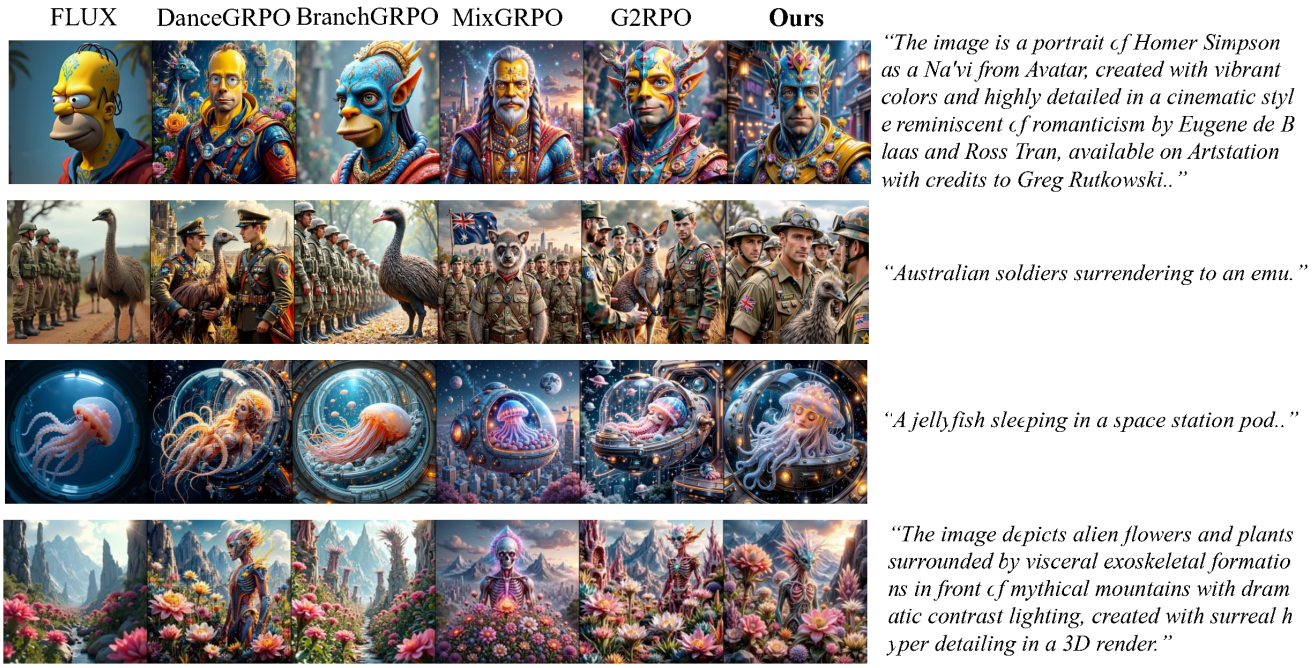

Figure 8. **Failure cases of E-GRPO**