Fine-Grained GRPO for Precise Preference Alignment in Flow Models

Supplementary Material

In the supplementary material, we present additional implementation details (Section 1), additional quantitative comparison with baselines (Section 2), visual ablation study of MGAI module (Section 3), additional qualitative evaluation (Section 4), more visual samples of G^{2}RPO (Section 5), Comparison of reward curves of the training phase (Section 6), as well as the limitation of our method (Section 7), as a supplement to the main paper.

1. Additional Implementation Details

Tab. 1 presents the detailed hyperparameter configuration used in our experiments. We keep the same hyperparameter configuration across all experiments.

Table 1. **Hyperparameter settings used in all experiments.**

Parameter	Value	Parameter	Value
Random seed	42	Learning rate	2×10^{-6}
Train batch size	1	Weight decay	1×10^{-4}
Warmup steps	0	Mixed precision	bfloat16
Dataloader workers	4	Max grad norm	1.0
Eta	0.7	Sampler seed	1223627
Group size	12	Scheduler shift	3
Sampling steps	16	Adv. clip max	5.0
Init same noise	Yes	Granularity Λ	$\{1, 2, 3\}$
The number of GPUs	16	Clip range	1×10^{-4}

2. Additional Quantitative Evaluation

To further demonstrate the robustness and superiority of G^{2}RPO over baseline methods, as well as the comprehensive enhancement effects brought by multi-granularity evaluation to GRPO training, we employ the latest UniGenBench++ [4] as the benchmark for evaluation. UniGenBench++ is a unified and versatile benchmark for image generation that integrates diverse prompt themes with a comprehensive suite of fine-grained evaluation criteria. The benchmark encompasses 10 primary dimensions, covering semantic evaluation, image quality assessment, text alignment, and other aspects, to provide a complete and thorough evaluation of generative models. Official offline evaluation model *UniGenBench-EvalModel-qwen-72b-v1* is used as the VLM for evaluation.

As shown in Tab. 2, our Singular Stochastic Sampling strategy (i.e., G^{2}RPO without MGAI) enhances the precision of reward signals during GRPO training, thereby aligning the model's output more closely with the reward models. This leads to significant improvements in several evaluation metrics, including Attribute, Relation, and Text. However, relying solely on single-granularity alignment tends to overfit to the biases inherent in the reward models. This causes the model to generate outputs that increasingly collapse into a narrower domain, resulting in degraded performance on metrics such as Style and Layout.

Notably, our Multi-Granularity Advantage Integration module (i.e., MGAI) provides a more comprehensive evaluation of the sampling directions within a group, enabling the selection of samples that exhibit advantages at both coarse-grained (structural) and fine-grained (textural) levels. This leads to more robust model updates and effectively mitigates the risk of domain collapse. Ultimately, our G^{2}RPO achieves substantial overall performance improvements on UniGenBench++.

3. Visual Ablation Study of MGAI

In this section, We conduct a visual ablation of the Multi-Granularity Advantage Integration (MGAI) module. Fig. 1 compares eight pairs of samples, with each pair sharing the same prompt and random seed. Without MGAI, G^{2}RPO performs single-granularity denoising, whose trajectory is locked to a fixed step budget. Group-wise advantage estimation under this setting is easily biased: the reward model over-attends to fine details while ignoring coarse structural coherence, so samples with higher reward frequently exhibit distorted textures or global misalignment. Consequently, single-granularity generators tend to yield images that exhibit excessive textural detail while suffering from structural fragility.

The proposed MGAI alleviates this by re-weighting each group sample only when it simultaneously surpasses its peers at both coarse and fine scales, forcing the policy to improve detail fidelity and global layout jointly. VLM-centric metrics (Unified Reward [5] and UniGenBench++) capture this multi-dimensional gain, showing large boosts in prompt adherence, detail accuracy, and overall quality. Conversely, uni-dimensional metrics such as HPS-V2.1 [6], CLIP Score [3], and Pick Score [1] can already be overfitted with our singular stochastic sampling strategy, hence these metrics alone cannot fully capture the holistic gains conferred by MGAI.

4. Additional Qualitative Comparisons

In this section, we present more qualitative comparison results between our G^{2}RPO and existing flow-based GRPO methods [2, 7], as illustrated in Fig. 2, Fig. 3, and Fig. 4. G^{2}RPO achieves superior visual fidelity and text-image alignment.

Table 2. **Quantitative Comparison on UniGenBench++ [4]**. Best scores are in **bold**, second-best in underlined.

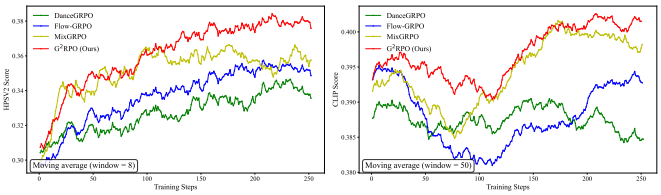
Model	Overall	Style	World Know.	Attribute	Action	Relation.	Logic.Reason.	Grammar	Compound	Layout	Text
Flux.1-dev	61.59	**84.60**	86.87	66.77	62.74	67.13	29.36	**60.83**	47.04	71.08	39.48
DanceGRPO	66.06	76.00	87.82	73.72	68.82	75.38	38.76	59.63	64.95	80.78	34.77
MixGRPO	66.60	80.80	87.97	74.04	68.82	75.00	38.99	59.49	64.82	**81.34**	34.77
G^{2}RPO w/o MGAI	66.32	73.70	86.08	75.53	67.49	77.28	39.45	60.70	65.21	76.68	41.09
G^{2}RPO	**69.21**	76.20	**89.08**	**79.91**	**71.10**	**78.17**	**42.66**	58.82	**70.36**	79.29	**46.55**

5. Gallery of G^{2}RPO

In this section, we provide more visual samples of the proposed G^{2}RPO to demonstrate its generation capability, as shown in Fig. 5, Fig. 6, and Fig. 7. Text prompts used to generate images are randomly sampled from UniGenBench++ [4].

6. Convergence Curves

The reward convergence curves of G^{2}RPO *v.s.* baselines during multi-reward training are shown below. G^{2}RPO demonstrates superior training dynamics: it climbs faster (efficiency), achieves a higher ceiling (effectiveness), and has a more stable curve (stability).



7. Limitation

Despite the advancements of our G^{2}RPO in human preference alignment, it faces certain constraints. Specifically, G^{2}RPO incurs additional sampling time due to multi-granularity sampling. We quantify the additional computational cost introduced by the granularity set $\Lambda = \{1, 2, 3\}$. Let $\mathcal{M} = \{m_1, \ldots, m_k\}$ be the SDEs timesteps for GRPO training. The standard single-granularity schedule requires $S_1 = \sum_{m \in \mathcal{M}} m$ denoising steps. Then, the tri-granular schedule introduces the additional step counts:

$$S_2 = \sum_{m \in \mathcal{M}} \left\lfloor \frac{m}{2} \right\rfloor, \quad S_3 = \sum_{m \in \mathcal{M}} \left\lfloor \frac{m}{3} \right\rfloor.$$

Total steps and relative overhead are therefore

$$T_{\{1,2,3\}} = S_1 + S_2 + S_3, \quad \Delta_{\{1,2,3\}} = \frac{S_2 + S_3}{S_1 + S_2 + S_3}.$$

For example, with $\mathcal{M} = \{16, 15, ..., 9\}$, we have $T_{\{1,2,3\}} = 184$ and $\Delta_{\{1,2,3\}} \approx 45.7\%$. Although our MGAI module adds a moderate training-time overhead, this cost is incurred only once and leaves inference latency unchanged. Meanwhile, as shown in Fig.1 at the main paper, G^{2}RPO achieves markedly faster alignment with the reward model compared with baseline method.

References

[1] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 1

[2] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025. 1

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

[4] Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, et al. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025. 1, 2

[5] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. 1

[6] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 1

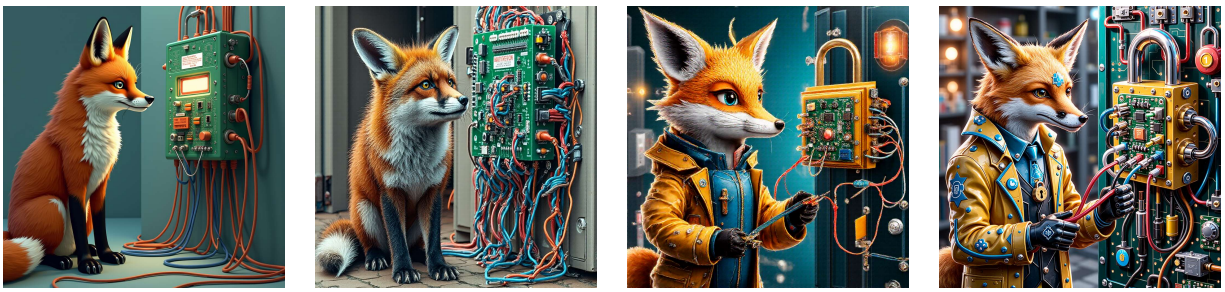
[7] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 1

Figure 1. **Visual ablation study of MGAI module.**

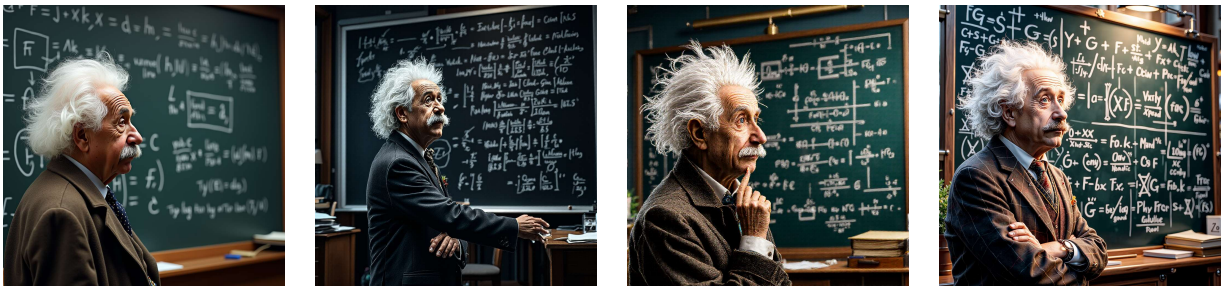
Flux.1-dev	DanceGRPO	MixGRPO	**G^2 RPO (Ours)**

"Please create a sculpture. The main body is a robot that imitates Rodin's "The Thinker". The whole body is made of transparent glass and has complex golden gears running inside, in a steampunk style."

"A golden Labrador retriever is leaping excitedly on the green grass, chasing a soap bubble that glows with a rainbow in the sun, National Geographic photography style."



"A biochemically modified fox faced a complex electronic lock. Instead of forcibly destroying it, it observed the wires and unplugged one of the key wires."



"Physicist Albert Einstein mused in front of a blackboard filled with complex formulas and his trademark messy white hair stood out under the light."

Figure 2. **Qualitative comparison with existing GRPO methods (1/3).**

Flux.1-dev | DanceGRPO | MixGRPO | **G^2 RPO (Ours)**

"In the Japanese animation style, an anthropomorphic Shiba Inu wearing a chef's uniform is skillfully pinching an exquisite sushi with his front paws."

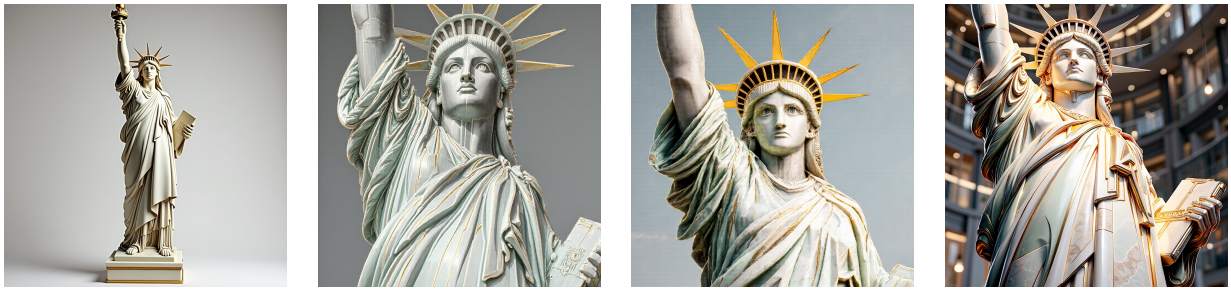
"Next to a huge rough concrete square, there is a small and exquisite glass bird. The picture adopts a minimalist style with clear light and shadow."

"A crystal wall clock in the shape of an ancient Roman Colosseum, inside the clock is a miniature city."

"Please design a magnificent Baroque Opera House interior, center stage, a hippo wearing a tutu is elegantly curtain call."

Figure 3. **Qualitative comparison with existing GRPO methods (2/3).**

Flux.1-dev | DanceGRPO | MixGRPO | **G^2RPO (Ours)**



"An Art Deco sculpture, the body of the Statue of Liberty is composed of smooth white marble and shiny gold lines."

"Generate pictures: Please design a poster for an environmental theme, showing a polar bear standing alone on a piece of ice floe that is about to melt, looking out at the silhouette of the industrial city in the distance, in Memphis style."

"An ancient mysterious treasure chest. The box is covered with moss and vines, and a faint golden light shines through the keyhole. It has a game asset in a dark and realistic style."

"In the surrealist photography style, an old man with a root-like beard is stroking a deer made of crystal, with small white flowers blooming on its antlers."

Figure 4. **Qualitative comparison with existing GRPO methods (3/3).**

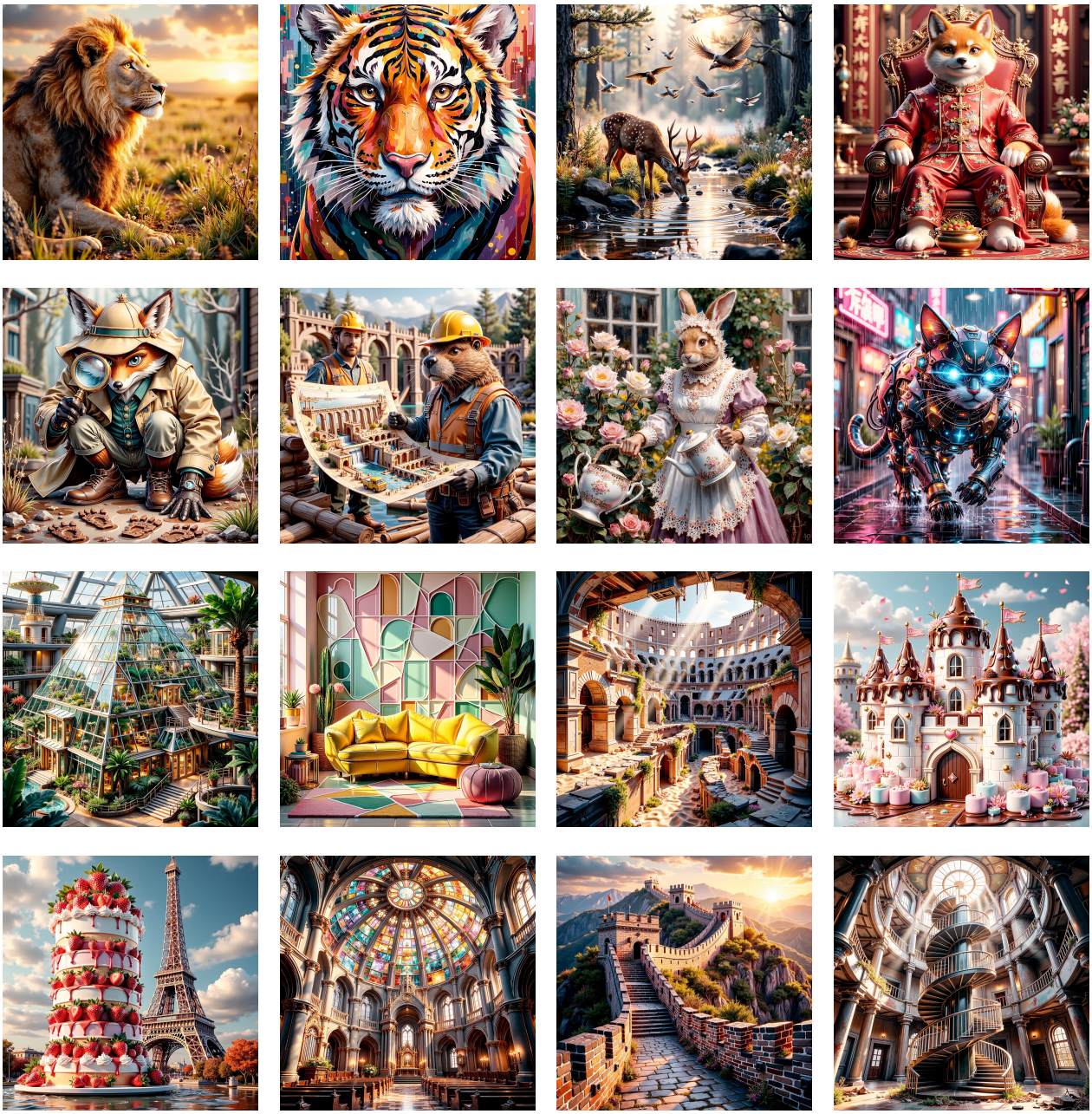


Figure 5. **Gallery of G^{2}RPO (1/3).**

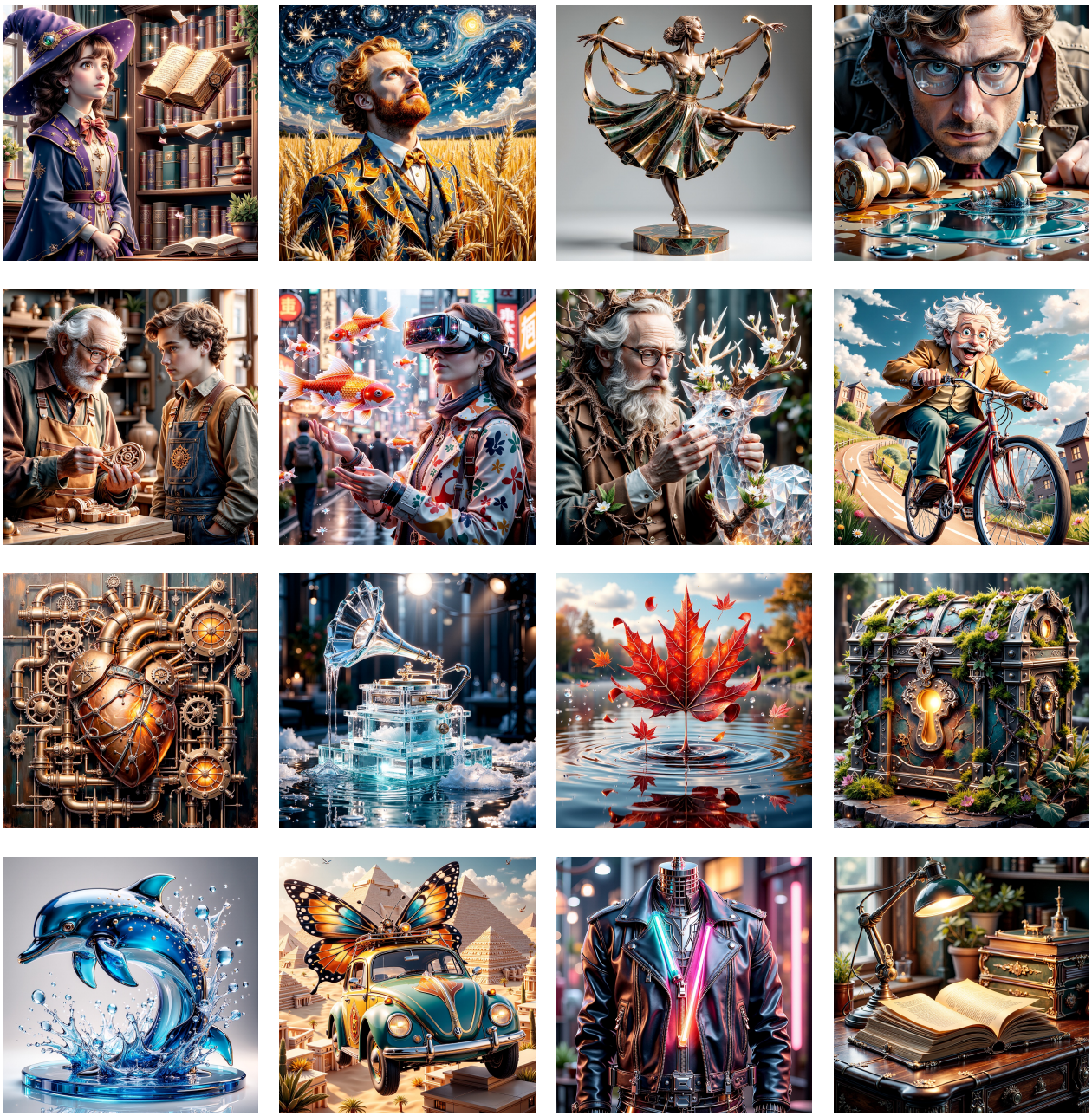


Figure 6. **Gallery of G^{2}RPO (2/3).**

Figure 7. **Gallery of G^{2}RPO (3/3).**