
Appendix of Flow-GRPO: Training Flow Matching Models via Online RL

Content of Appendix

257	A Mathematical Derivations for Stochastic Sampling using Flow Models	2
258	B Further Details on the Experimental Setup	3
259	B.1 Quality Metrics	3
260	B.2 Model Specification	3
261	B.3 Hyperparameters Specification	3
262	B.4 Compute Resources Specification	4
263	C Extended Experimental Results	4
264	C.1 Flow-GRPO vs. Other Alignment Methods	4
265	C.2 Effect of Denoising Reduction	5
266	C.3 Learning Curves with or without KL	5
267	C.4 Additional Qualitative Results	6
268	C.5 Evolution of Evaluation Images During Flow-GRPO Training	6
269	D Training Sample Visualization with Denoising Reduction	6

270 Our Appendix consists of 4 sections. Readers can click on each section number to navigate to the
271 corresponding section:

- 272 • Section **A** provides detailed derivations of stochastic sampling in flow matching models.
- 273 • Section **B** presents details about our experimental setup.
- 274 • Section **C** offers some additional experimental results, including 1) the comparison with other
275 alignment methods, 2) ablation of denoising reduction on OCR accuracy and pickscore, 3) the
276 learning curves of Flow-GRPO on three tasks, 4) additional qualitative results, and 5) evolution
277 of evaluation images during training.
- 278 • Section **D** Visualization of training samples under the denoising reduction strategy.

279 In addition to this Appendix, we also provide **a local HTML for image comparisons**. We encourage
280 the reviewers to consult this HTML file for a more intuitive assessment of the improvements brought
281 by Flow-GRPO.

A Mathematical Derivations for Stochastic Sampling using Flow Models

We present a detailed proof here. To compute $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})$ in Equation 5 during forward sampling, we adapt flow models to a stochastic differential equation (SDE). While flow models normally follow a deterministic ODE:

$$d\mathbf{x}_t = \mathbf{v}_t dt \quad (10)$$

We consider its stochastic counterpart. Inspired by the derivation from SDE to its probability flow ODE in SGMs [23], we aim to construct a SDE with specific drift and diffusion coefficients so that its marginal distribution matches that of Eq. 10. We begin with the generic form of SDE:

$$d\mathbf{x}_t = f_{\text{SDE}}(\mathbf{x}_t, t)dt + \sigma_t d\mathbf{w}, \quad (11)$$

Its marginal probability density $p_t(\mathbf{x})$ evolves according to the Fokker–Planck equation [70], i.e.,

$$\partial_t p_t(\mathbf{x}) = -\nabla \cdot [f_{\text{SDE}}(\mathbf{x}_t, t)p_t(\mathbf{x})] + \frac{1}{2}\nabla^2[\sigma_t^2 p_t(\mathbf{x})] \quad (12)$$

Similarly, the marginal probability density associated with Eq. 10 evolves:

$$\partial_t p_t(\mathbf{x}) = -\nabla \cdot [\mathbf{v}_t(\mathbf{x}_t, t)p_t(\mathbf{x})] \quad (13)$$

To ensure that the stochastic process shares the same marginal distribution as the ODE, we impose:

$$-\nabla \cdot [f_{\text{SDE}}p_t(\mathbf{x})] + \frac{1}{2}\nabla^2[\sigma_t^2 p_t(\mathbf{x})] = -\nabla \cdot [\mathbf{v}_t p_t(\mathbf{x})] \quad (14)$$

Observing that

$$\begin{aligned} \nabla^2[\sigma_t^2 p_t(\mathbf{x})] &= \sigma_t^2 \nabla^2 p_t(\mathbf{x}) \\ &= \sigma_t^2 \nabla \cdot (\nabla p_t(\mathbf{x})) \\ &= \sigma_t^2 \nabla \cdot (p_t(\mathbf{x}) \nabla \log p_t(\mathbf{x})) \end{aligned} \quad (15)$$

Substituting Eq. 15 to Eq. 14, we arrive at the drift coefficients of the target SDE:

$$f_{\text{SDE}} = \mathbf{v}_t - \frac{1}{2}\sigma_t^2 \nabla \log p_t(\mathbf{x}) \quad (16)$$

Hence we can rewrite the reverse-time SDE in Eq. 11 as:

$$\boxed{d\mathbf{x}_t = \left(\mathbf{v}_t(\mathbf{x}_t) - \frac{\sigma_t^2}{2} \nabla \log p_t(\mathbf{x}_t) \right) dt + \sigma_t d\mathbf{w},} \quad (17)$$

where $d\mathbf{w}$ denotes Wiener process increments and σ_t is the diffusion coefficient that controls the level of stochasticity during sampling.

Once the score $\nabla \log p_t(\mathbf{x}_t)$ is available, the stochastic process can be sampled directly. While the Flow-Matching framework, this score is tied to the velocity field \mathbf{v}_t .

Specifically, let $\dot{\alpha}_t \equiv \partial \alpha_t / \partial t$. All expectations are over $\mathbf{x}_0 \sim X_0$ and $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$, where X_0 is the data distribution.

For the linear interpolation $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \beta_t \mathbf{x}_1$, we have:

$$p_{t|0}(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t \mid \alpha_t \mathbf{x}_0, \beta_t^2 \mathbf{I}), \quad (18)$$

yielding the conditional score:

$$\nabla \log p_{t|0}(\mathbf{x}_t | \mathbf{x}_0) = -\frac{\mathbf{x}_t - \alpha_t \mathbf{x}_0}{\beta_t^2} = -\frac{\mathbf{x}_1}{\beta_t}. \quad (19)$$

The marginal score becomes:

$$\begin{aligned} \nabla \log p_t(\mathbf{x}_t) &= \mathbb{E}[\nabla \log p_{t|0}(\mathbf{x}_t | \mathbf{x}_0) \mid \mathbf{x}_t] \\ &= -\frac{1}{\beta_t} \mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_t]. \end{aligned} \quad (20)$$

304 For the velocity field $\mathbf{v}_t(\mathbf{x}_t)$, we derive:

$$\begin{aligned}
\mathbf{v}_t(\mathbf{x}) &= \mathbb{E} \left[\dot{\alpha}_t \mathbf{x}_0 + \dot{\beta}_t \mathbf{x}_1 \mid \mathbf{x}_t = \mathbf{x} \right] \\
&= \dot{\alpha}_t \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}] + \dot{\beta}_t \mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_t = \mathbf{x}] \\
&= \dot{\alpha}_t \mathbb{E} \left[\frac{\mathbf{x}_t - \beta_t \mathbf{x}_1}{\alpha_t} \mid \mathbf{x}_t = \mathbf{x} \right] + \dot{\beta}_t \mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_t = \mathbf{x}] \\
&= \frac{\dot{\alpha}_t}{\alpha_t} \mathbf{x} - \frac{\dot{\alpha}_t \beta_t}{\alpha_t} \mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_t = \mathbf{x}] + \dot{\beta}_t \mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_t = \mathbf{x}] \\
&= \frac{\dot{\alpha}_t}{\alpha_t} \mathbf{x} - \left(\dot{\beta}_t \beta_t - \frac{\dot{\alpha}_t \beta_t^2}{\alpha_t} \right) \nabla \log p_t(\mathbf{x}),
\end{aligned} \tag{21}$$

305 Substituting $\alpha_t = 1 - t$ and $\beta_t = t$ simplifies Equation 21 to:

$$\mathbf{v}_t(\mathbf{x}) = -\frac{\mathbf{x}}{1-t} - \frac{t}{1-t} \nabla \log p_t(\mathbf{x}). \tag{22}$$

306 Solving for the score yields:

$$\nabla \log p_t(\mathbf{x}) = -\frac{\mathbf{x}}{t} - \frac{1-t}{t} \mathbf{v}_t(\mathbf{x}). \tag{23}$$

307 Substituting Equation 23 into 17 gives the final SDE:

$$d\mathbf{x}_t = \left[\mathbf{v}_t(\mathbf{x}_t) + \frac{\sigma_t^2}{2t} (\mathbf{x}_t + (1-t)\mathbf{v}_t(\mathbf{x}_t)) \right] dt + \sigma_t d\mathbf{w}. \tag{24}$$

308 Applying Euler-Maruyama discretization yields the update rule:

$$\boxed{\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \left[\mathbf{v}_\theta(\mathbf{x}_t, t) + \frac{\sigma_t^2}{2t} (\mathbf{x}_t + (1-t)\mathbf{v}_\theta(\mathbf{x}_t, t)) \right] \Delta t + \sigma_t \sqrt{\Delta t} \epsilon,} \tag{25}$$

309 where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ injects stochasticity.

310 **B Further Details on the Experimental Setup**

311 **B.1 Quality Metrics**

312 The details of quality metrics are as follows:

- 313 • Aesthetic score [56]: a CLIP-based linear regressor that predicts an image’s aesthetic score.
- 314 • DeQA score [57]: a multimodal large language model based image-quality assessment (IQA)
- 315 model that quantifies how distortions, texture damage, and other low-level artefacts affect per-
- 316 ceived quality.
- 317 • ImageReward [32]: a general purpose T2I human preference reward model that captures
- 318 text–image alignment, visual fidelity, and harmlessness.
- 319 • UnifiedReward [58]: a recently proposed unified reward model for multimodal understanding
- 320 and generation that currently achieves state-of-the-art performance on the human preference
- 321 assessment leaderboard.

322 **B.2 Model Specification**

323 The following table lists the base model and the reward models and their corresponding links.

324 **B.3 Hyperparameters Specification**

325 Except for β , GRPO hyperparameters are fixed across tasks. We use a sampling timestep $T = 10$ and
326 an evaluation timestep $T = 40$. Other settings include a group size $G = 24$, an noise level $a = 0.7$
327 and an image resolution of 512. The KL ratio β is set to 0.004 for GenEval and Text Rendering, and
328 0.001 for Pickscore. We use Lora with $\alpha = 64$ and $r = 32$.

Models	Links
SD3.5-M [4]	https://huggingface.co/stabilityai/stable-diffusion-3.5-medium
Aesthetic Score [56]	https://github.com/LAION-AI/aesthetic-predictor
PickScore [19]	https://huggingface.co/yuvalkirstain/PickScore_v1
DeQA score [57]	https://huggingface.co/zhiyuanyou/DeQA-Score-Mix3
ImageReward [32]	https://huggingface.co/THUDM/ImageReward
UnifiedReward [58]	https://huggingface.co/CodeGoat24/UnifiedReward-7b-v1.5

B.4 Compute Resources Specification

We train our model using 24 NVIDIA A800 GPUs. The learning curves in Appendix C.3 provide details on the specific GPU hours.

C Extended Experimental Results

C.1 Flow-GRPO vs. Other Alignment Methods

We compare Flow-GRPO with several alignment methods: supervised fine-tuning (SFT), reward-weighted regression (Flow-RWR [14, 68]), Flow-DPO [14], and their online variants. Flow-GRPO consistently outperforms all baselines by a significant margin. At each step, we generate a group of images using the same group size as in Flow-GRPO. The only difference lies in the update rule:

- **SFT**: Select the highest-reward image in each group and fine-tune on it.
- **Flow-RWR**[14, 68]: Apply a softmax over rewards in each group and perform reward-weighted likelihood maximization.
- **Flow-DPO**[14]: Use the highest-reward image in each group as the chosen sample and the lowest as the rejected, then apply the DPO loss.

Offline variants use a fixed pretrained model for data collection, while online variants update their data collection model every 40 steps. As shown in Figure 6, Flow-GRPO outperforms all other methods. The figure also indicates that DPO and SFT improve over time. In contrast, RWR does not, which aligns with experimental findings on RWR in [12]. Additionally, Online DPO surpasses offline DPO, aligning with [15]’s finding that online DPO performs better. For the second-best online DPO, a hyperparameter search on its key parameter β revealed that smaller values are not always optimal; excessively small β values can cause training collapse.

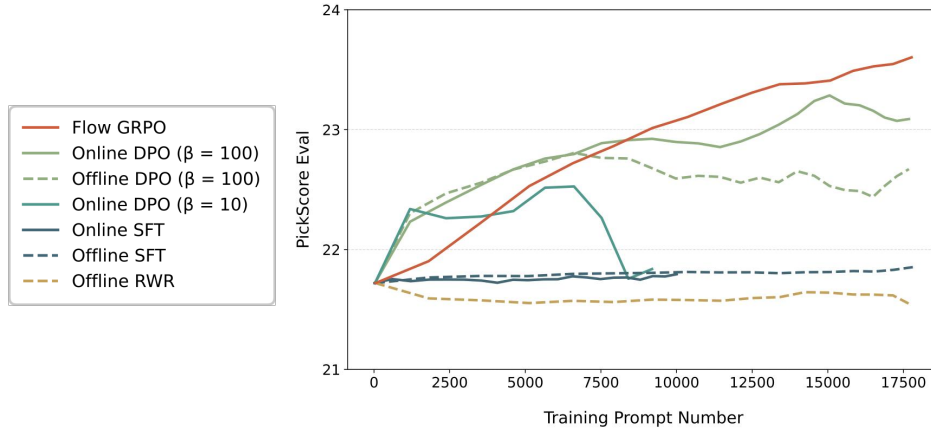


Figure 6: Comparison of Flow-GRPO and Other Alignment Methods.

350 C.2 Effect of Denoising Reduction

351 We show the extended Denoising Reduction ablations of Visual Text Rendering and Human Preference Alignment tasks in Figure 7.

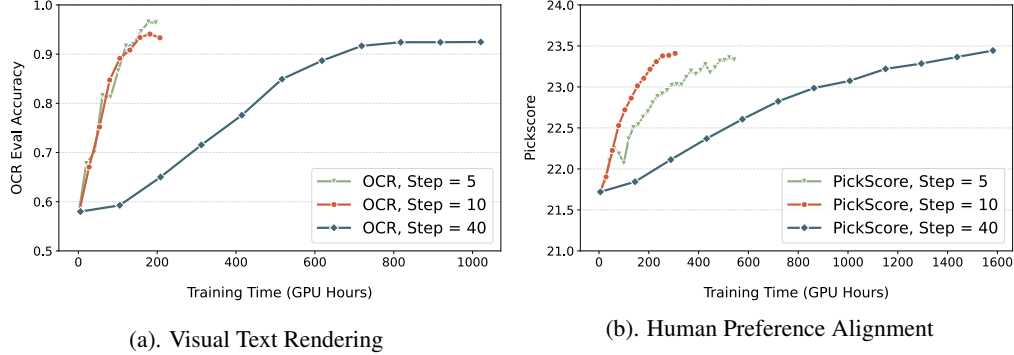


Figure 7: Effect of Denoising Reduction

352

353 C.3 Learning Curves with or without KL

354 Figure 8 shows learning curves for three tasks, with and without KL. These results emphasize that
 355 KL regularization is not empirically equivalent to early stopping. Adding appropriate KL can achieve
 356 the same high reward as the KL-free version and maintain image quality, though it requires longer training.

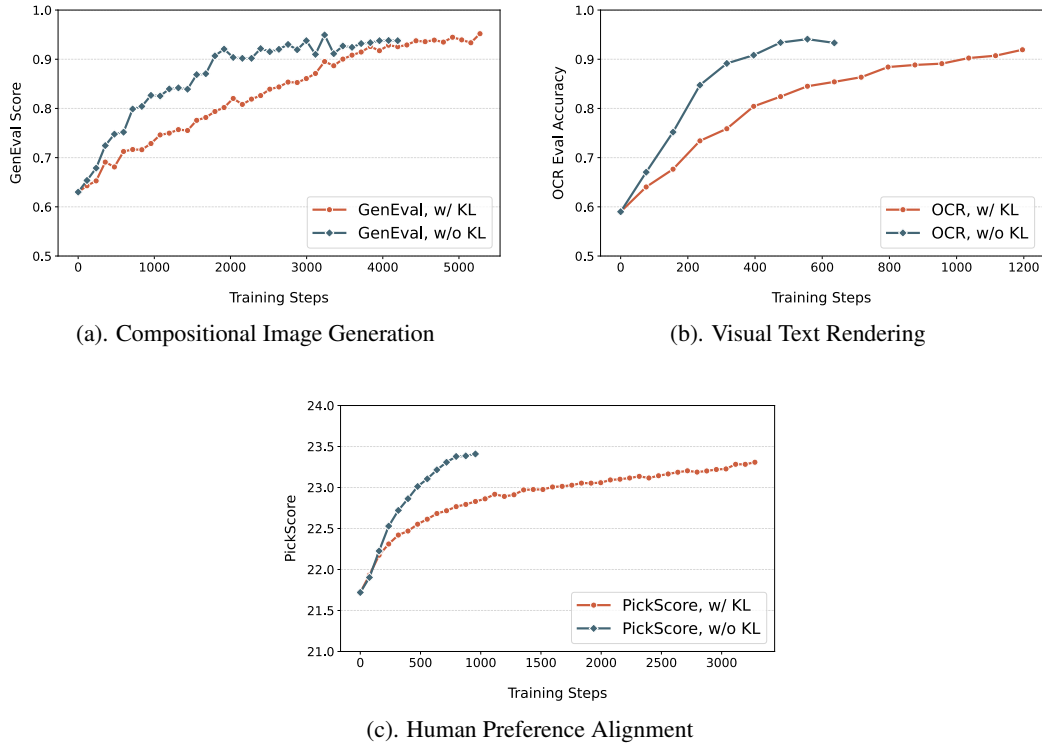


Figure 8: Learning Curves with and without KL. KL penalty slows early training yet effectively suppresses reward hacking.

357

C.4 Additional Qualitative Results

Figures 9, 10 & 11 qualitatively compare SD3.5-M with its Flow-GRPO enhanced versions (with and without KL regularization) using GenEval, OCR and PickScore rewards, respectively. Flow-GRPO with KL regularization improves the target capability while maintaining image quality and minimizing reward-hacking. Conversely, removing the KL constraint significantly degrades image quality and diversity.

C.5 Evolution of Evaluation Images During Flow-GRPO Training

To better understand the training dynamics of our proposed Flow-GRPO framework, we visualize the evolution of generated samples corresponding to fixed evaluation prompts at regular intervals during training in Figure 12, 13 & 14. For consistency, all visualizations are produced using a 40-step ODE-based sampling schedule. These qualitative results provide a visual representation of how the model progressively improves its generation quality and alignment with task objectives over time.

D Training Sample Visualization with Denoising Reduction

In this section, we compare images obtained with SDE sampling at various steps against those produced by ODE sampling, and offer an intuitive view of the denoising reduction strategy. Figure 15 presents SD3.5-Medium samples under four inference settings: (a) ODE sampling with 40 steps; (b) SDE sampling with 40 steps; (c) SDE sampling with 10 steps; (d) SDE sampling with 5 steps.

The 40-step ODE and SDE runs yield visually indistinguishable images, confirming that our SDE sampler preserves quality. Shortening the SDE schedule to 10 and 5 steps introduces conspicuous artifacts, like color drift and fine details blur. Contrary to expectation that such low-quality samples might hinder optimization, it actually do just the opposite and accelerate optimization. Because Flow-GRPO relies on relative preferences, it still extracts a useful reward signal, while the shorter trajectories signifantly cut wall-clock time. Consequently, Flow-GRPO with denoising reduction strategy converges more quickly on both layout-oriented benchmarks such as GenEval and quality-focused metrics such as PickScore, without sacrificing final performance.

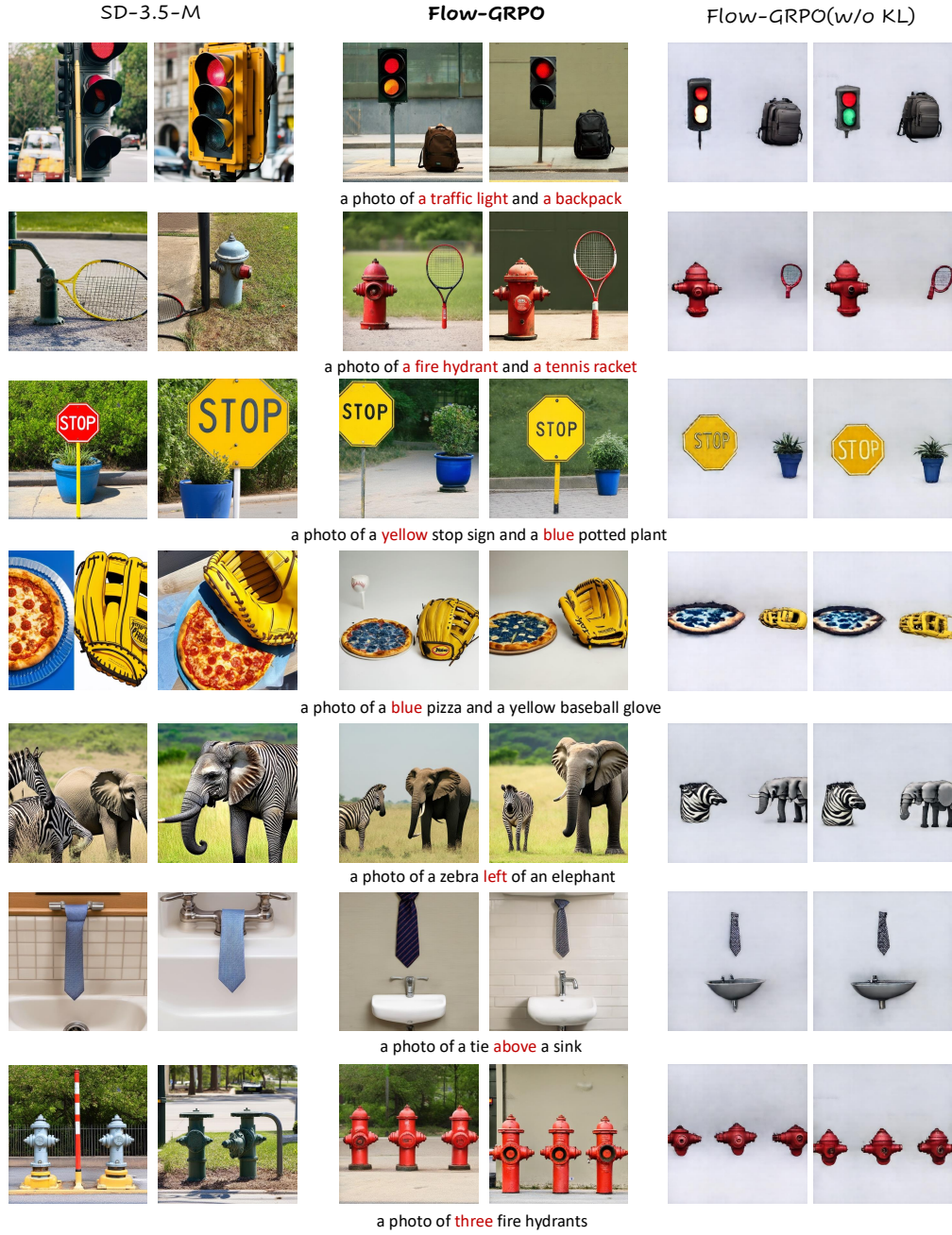


Figure 9: Additional Qualitative comparison between the SD3.5-M and SD3.5-M + Flow-GRPO trained with **GenEval** reward.

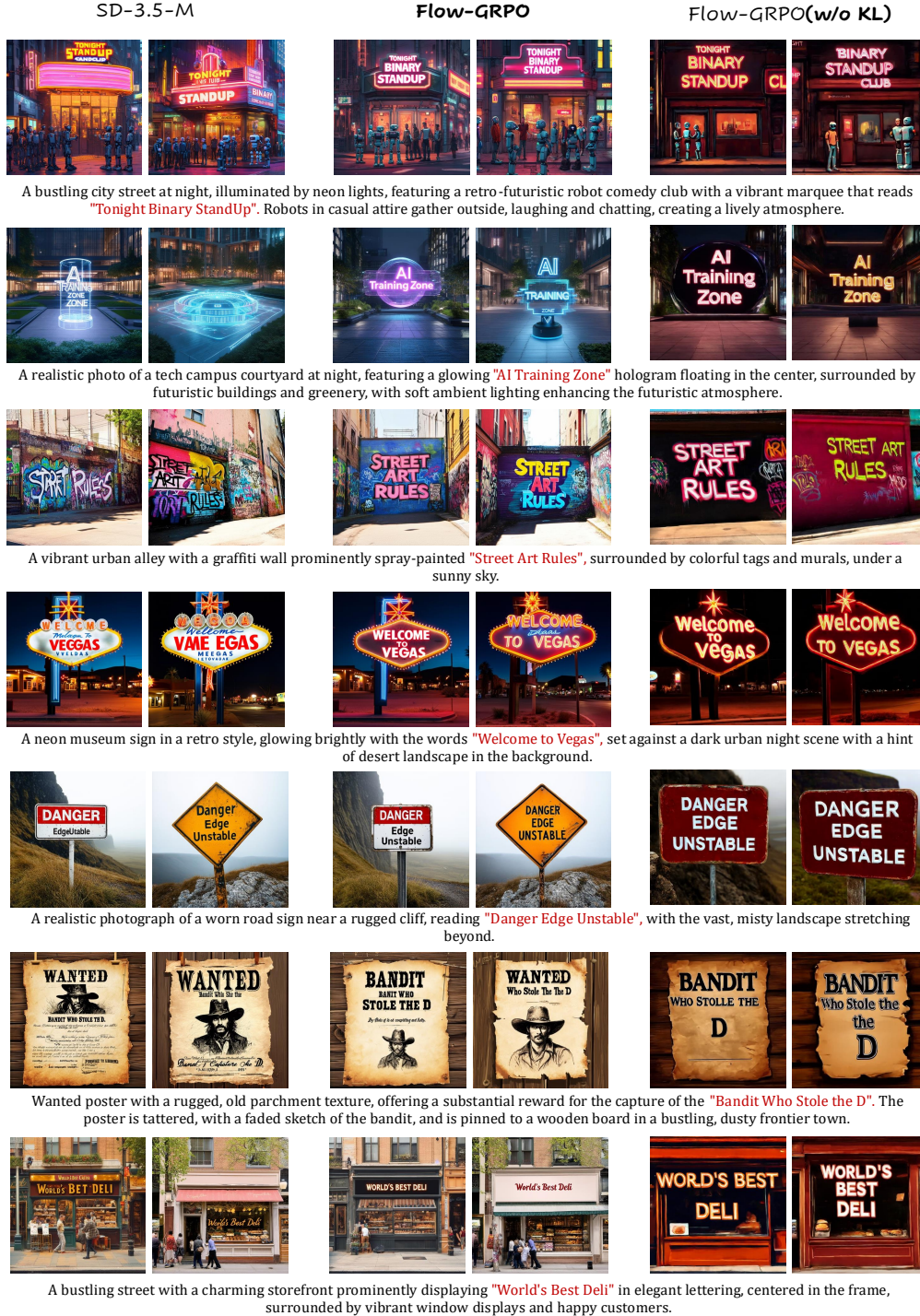


Figure 10: Additional Qualitative comparison between the SD3.5-M and SD3.5-M + Flow-GRPO trained with OCR reward.

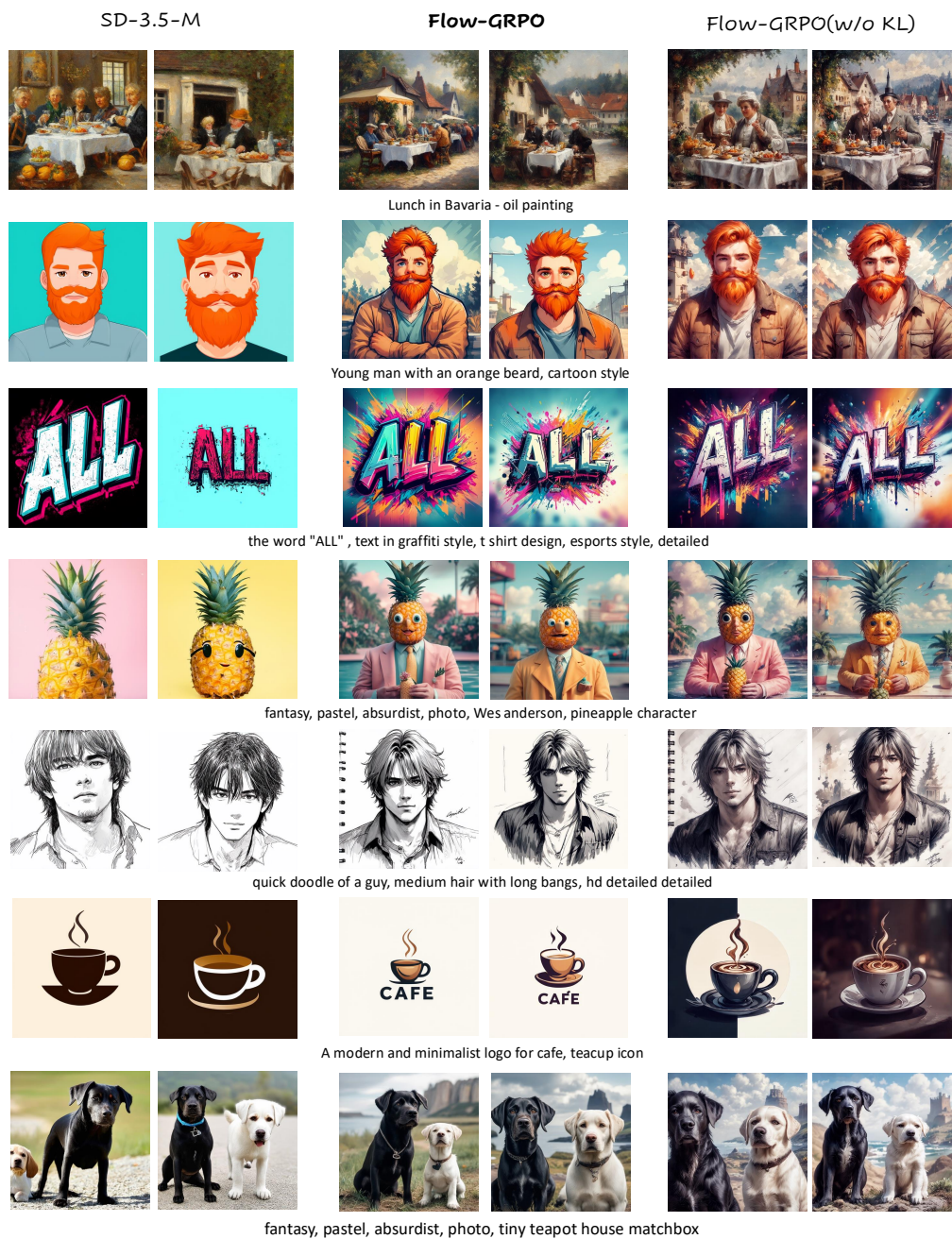


Figure 11: Additional Qualitative comparison between the SD3.5-M and SD3.5-M + Flow-GRPO trained with **PickScore** reward.



Figure 12: We visualize the generated samples across successive training iterations during the optimization of SD3.5-Medium on the **GenEval** task.



Figure 13: We visualize the generated samples across successive training iterations during the optimization of SD3.5-Medium on the **OCR** task.



Figure 14: We visualize the generated samples across successive training iterations during the optimization of SD3.5-Medium on the **PickScore** task.

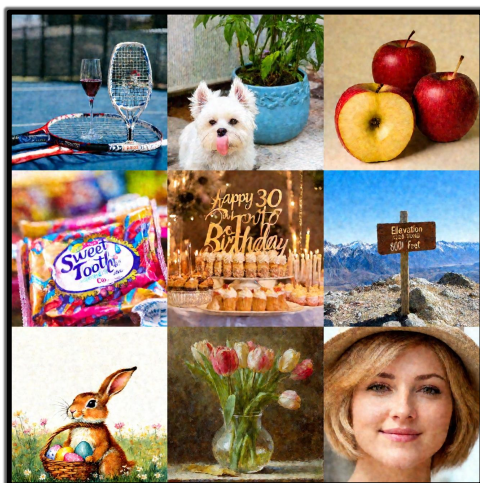
ODE Sampling
(Step = 40)



SDE Sampling
(Step = 40)



SDE Sampling
(Step = 10)



SDE Sampling
(Step = 5)



Figure 15: Visualization of training samples under difference inference settings.

References

- [1] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [2] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [3] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
- [4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- [5] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [6] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.
- [7] Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025.
- [8] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.
- [9] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.
- [10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [11] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [12] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.
- [13] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
- [14] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.
- [15] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.
- [16] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [17] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

- [18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [19] 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.
- [20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [23] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [24] Kuaishou. Kling ai. <https://klingai.kuaishou.com/>, 2024.
- [25] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- [26] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024.
- [27] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- [28] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.
- [29] Carles Domingo-Enrich, Michal Drozdal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. *arXiv preprint arXiv:2409.08861*, 2024.
- [30] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.
- [31] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.
- [32] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [33] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.
- [34] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.
- [35] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

- [36] Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback. *arXiv preprint arXiv:2412.02617*, 2024.
- [37] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- [38] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.
- [39] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.
- [40] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihang Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024.
- [41] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2024.
- [42] Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *arXiv preprint arXiv:2402.10210*, 2024.
- [43] Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024.
- [44] Jiacheng Zhang, Jie Wu, Weifeng Chen, Yatai Ji, Xuefeng Xiao, Weilin Huang, and Kai Han. Onlinevpo: Align video diffusion model with online video-centric preference optimization. *arXiv preprint arXiv:2412.15159*, 2024.
- [45] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [46] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.
- [47] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [48] Shashank Gupta, Chaitanya Ahuja, Tsung-Yu Lin, Sreya Dutta Roy, Harrie Oosterhuis, Maarten de Rijke, and Satya Narayan Shukla. A simple and effective reinforcement learning method for text-to-image diffusion fine-tuning. *arXiv preprint arXiv:2503.00897*, 2025.
- [49] Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024.
- [50] Po-Hung Yeh, Kuang-Huei Lee, and Jun-Cheng Chen. Training-free diffusion model alignment with sampling demons. *arXiv preprint arXiv:2410.05760*, 2024.
- [51] Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Tuning-free alignment of diffusion models with direct noise optimization. *arXiv preprint arXiv:2405.18881*, 2024.

- [52] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pages 32483–32498. PMLR, 2023.
- [53] Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. F5r-tts: Improving flow matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*, 2025.
- [54] Jaihoon Kim, Taehoon Yoon, Jisung Hwang, and Minhyuk Sung. Inference-time scaling for flow models via stochastic generation and rollover budget forcing. *arXiv preprint arXiv:2503.19385*, 2025.
- [55] Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. Seedream 2.0: A native chinese-english bilingual image generation foundation model. *arXiv preprint arXiv:2503.07703*, 2025.
- [56] Chrisoph Schuhmann. Laion aesthetics, Aug 2022.
- [57] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. *arXiv preprint arXiv:2501.11561*, 2025.
- [58] 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.
- [59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [60] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [61] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- [62] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- [63] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- [64] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.
- [65] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.
- [66] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.
- [67] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025.

- 572 [68] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression:
573 Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.
- 574 [69] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-
575 compbench++: An enhanced and comprehensive benchmark for compositional text-to-image
576 generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- 577 [70] Bernt Øksendal and Bernt Øksendal. *Stochastic differential equations*. Springer, 2003.