Flow-GRPO:
Training Flow Matching Models via Online RL

Jie Liu*, Gongye Liu*, Jiajun Liang, Yangguang Li,
Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, Wanli Ouyang✉

香港中文大學
The Chinese University of Hong Kong

TSINGHUA UNIVERSITY 1911

KlingAI

NANJING UNIVERSITY 1902

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

1. **Existing post-training algorithms for FM models**

2. **Flow-GRPO: Training FM models via online RL**

3. **Experiments**

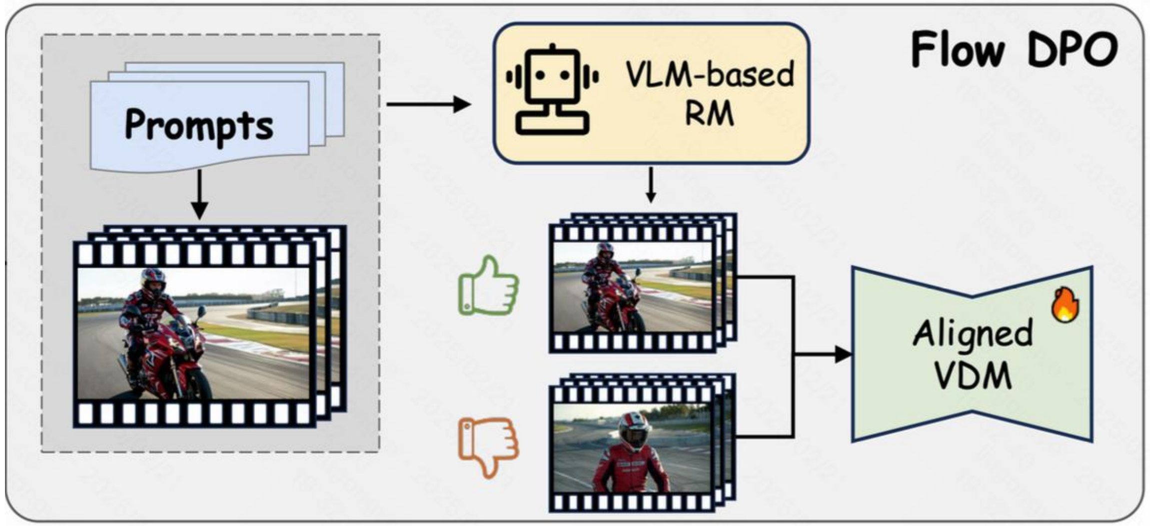
Post-Training Algorithms

- Evolution in LLM
    - 2024: DPO / Online-DPO dominated alignment methods
    - 2025: GRPO / PPO replaced DPO across LLMs
        - Offering better stability, on-policy updates, and higher reward efficiency.

<u>Can an online RL method like GRPO can be applied to Flow Matching Models?</u>

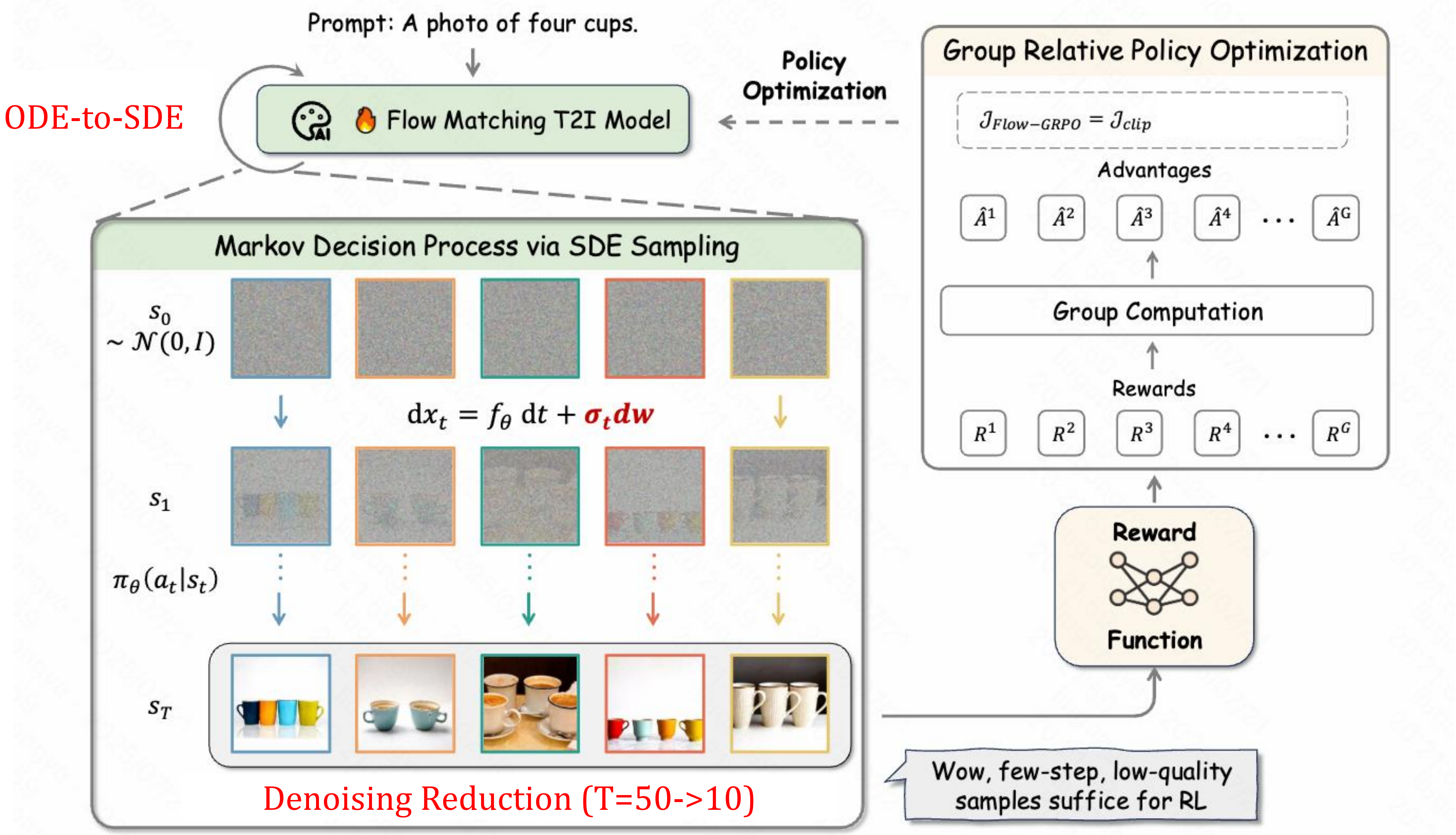
Existing Post-Training Algorithms: DPO

- Limitations of Flow-DPO
  - DPO: Learns directly from preference pairs → low data efficiency, limited performance gain.
  - GRPO: Trains a reward model to relabel online generated new samples → more stable training with steadily improving rewards.



$$-\mathbb{E}\left[\log\sigma\left(-\frac{\beta_t}{2}\Big(\|\boldsymbol{v}^w - \boldsymbol{v}_\theta(\boldsymbol{x}_t^w, t)\|^2 - \|\boldsymbol{v}^w - \boldsymbol{v}_{\mathrm{ref}}(\boldsymbol{x}_t^w, t)\|^2 - \big(\|\boldsymbol{v}^l - \boldsymbol{v}_\theta(\boldsymbol{x}_t^l, t)\|^2 - \|\boldsymbol{v}^l - \boldsymbol{v}_{\mathrm{ref}}(\boldsymbol{x}_t^l, t)\|^2\big)\Big)\right)\right],$$

Post-Training Algorithms: GRPO



Challenges & Solutions

Challenge1: GRPO require a statistic sampling, but FM is trained for deterministic sampling.

Solution1: ODE-to-SDE

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{v}_t \mathrm{d}t,$$

$$-\nabla \cdot [f_{\mathrm{SDE}}\, p_t(\boldsymbol{x})] + \frac{1}{2}\nabla^2[\sigma_t^2 p_t(\boldsymbol{x})] = -\nabla \cdot [\boldsymbol{v}_t(\boldsymbol{x}_t, t) p_t(\boldsymbol{x})]$$

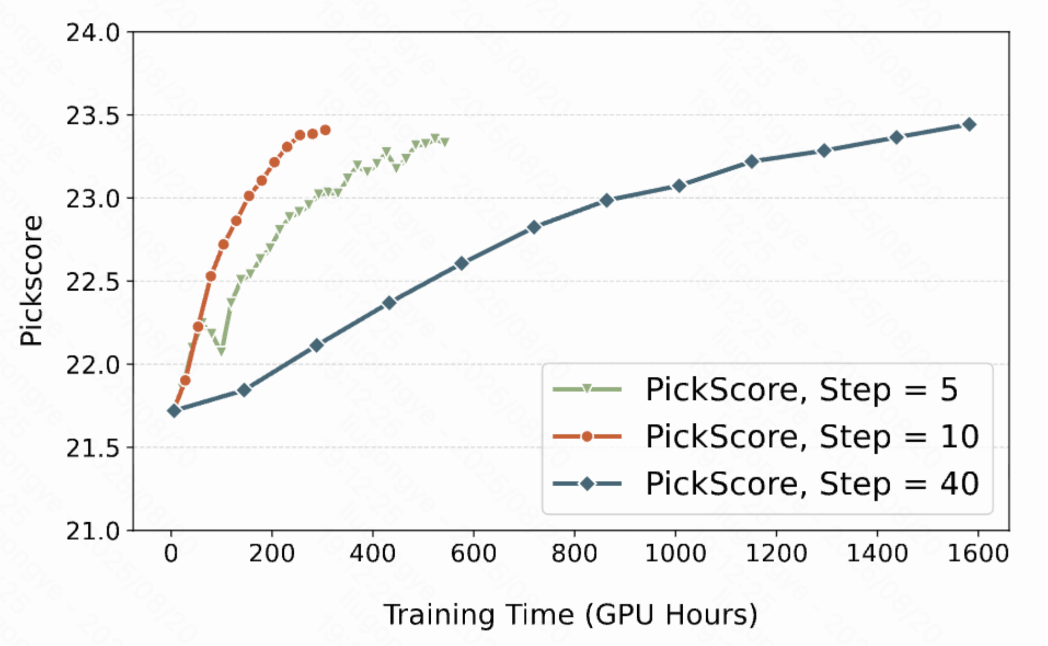
with equal marginal distribution

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

Challenges & Solutions

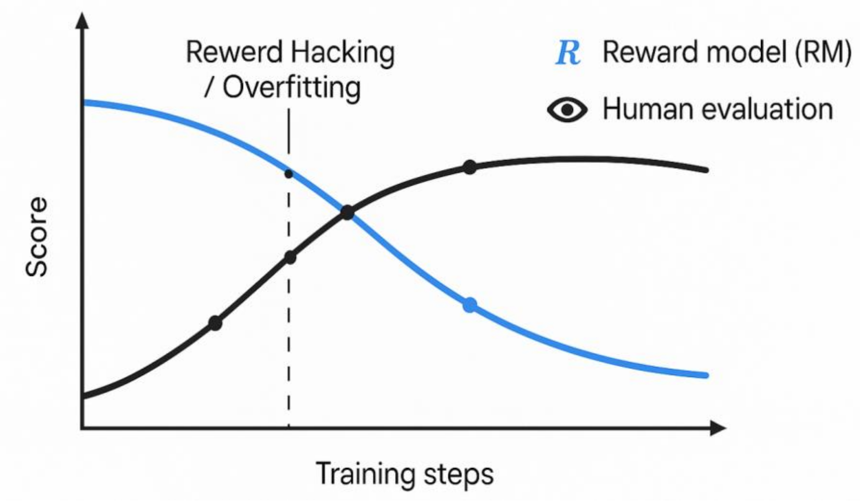
Challenge2: GRPO requires multiple samplings, and each sampling takes about 50 steps -> Too slow! 😩

Solution2: Denoising Reduction. Just 10-step training are sufficient to boost performance for 40-step inference -> Fast! 😊

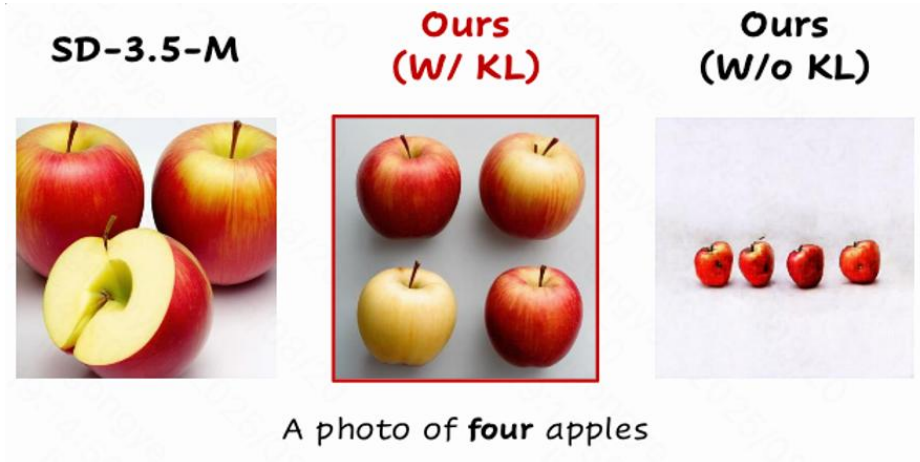


Challenges & Solutions

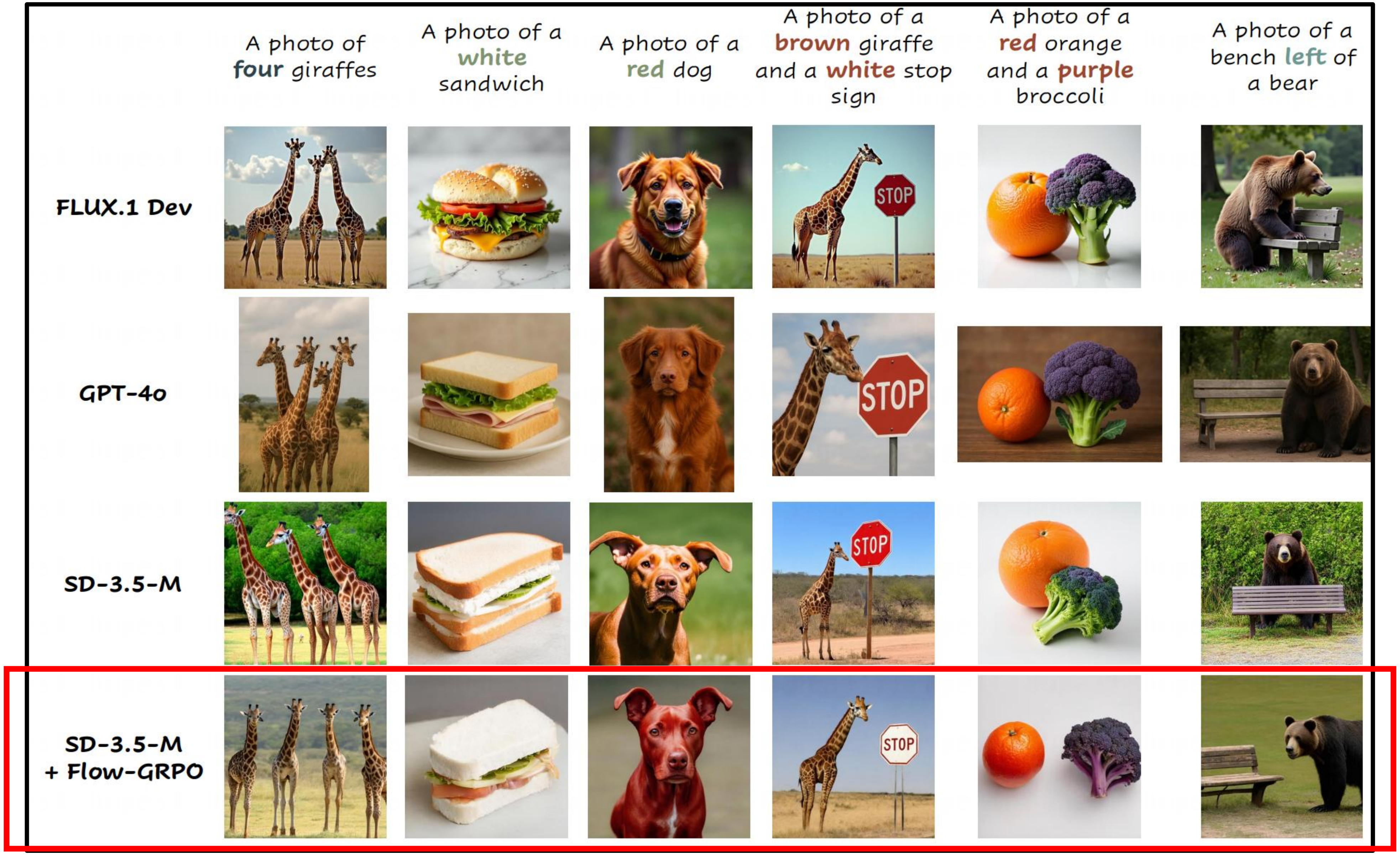
Challenge3: Reward Hacking problem-> Overfitting to the imperfect rm



Solution3: KL divergence can effectively prevent reward hacking.



Experiments



Thanks

paper

code