Test-Time Reinforcement Learning for Flow Matching

Jili Chen[1 2], Changqin Huang[3], Qionghao Huang[1], Yaxin Tu[3], Zhonglong Zheng[2], Xiaodi Huang[4]

[1] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University
[2] School of Computer Science and Technology, Zhejiang Normal University
[3] College of Education, Zhejiang University
[4] School of Computing, Mathematics and Engineering, Charles Sturt University

---

Presenter：Jili Chen

CONTENTS

01

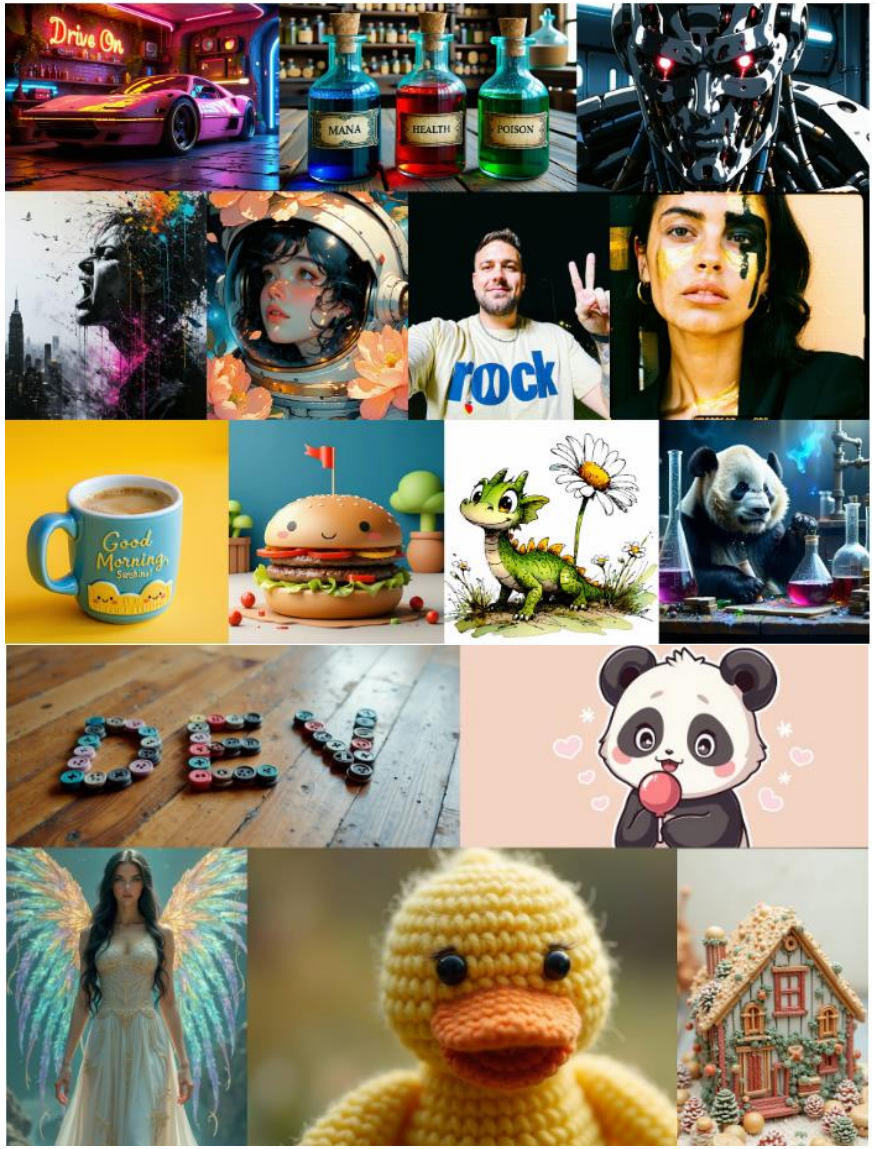
Problem and Motivation

Problem and Motivation

Mainstream T2I
Flow-matching+RL

SD3.5

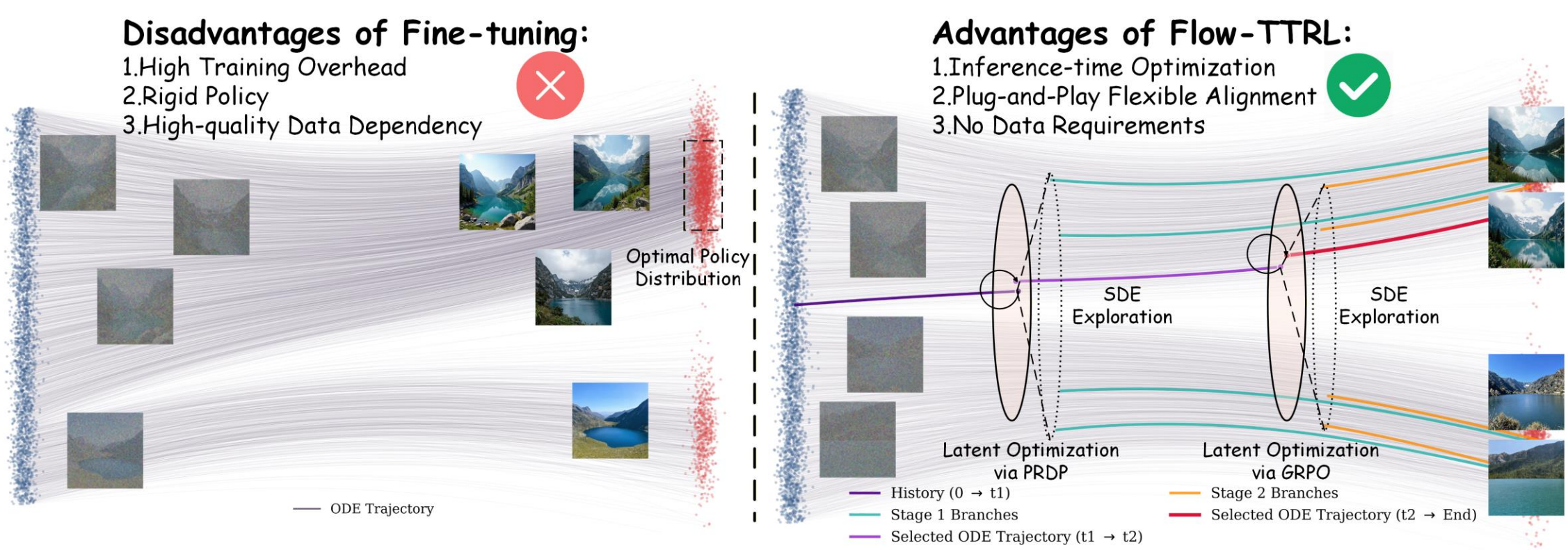
Flux.1 Dev



w/o Flow-DPO vs. w/ Flow-DPO

Problem and Motivation

Flow Matching Dominates
but RL Alignment Is Costly

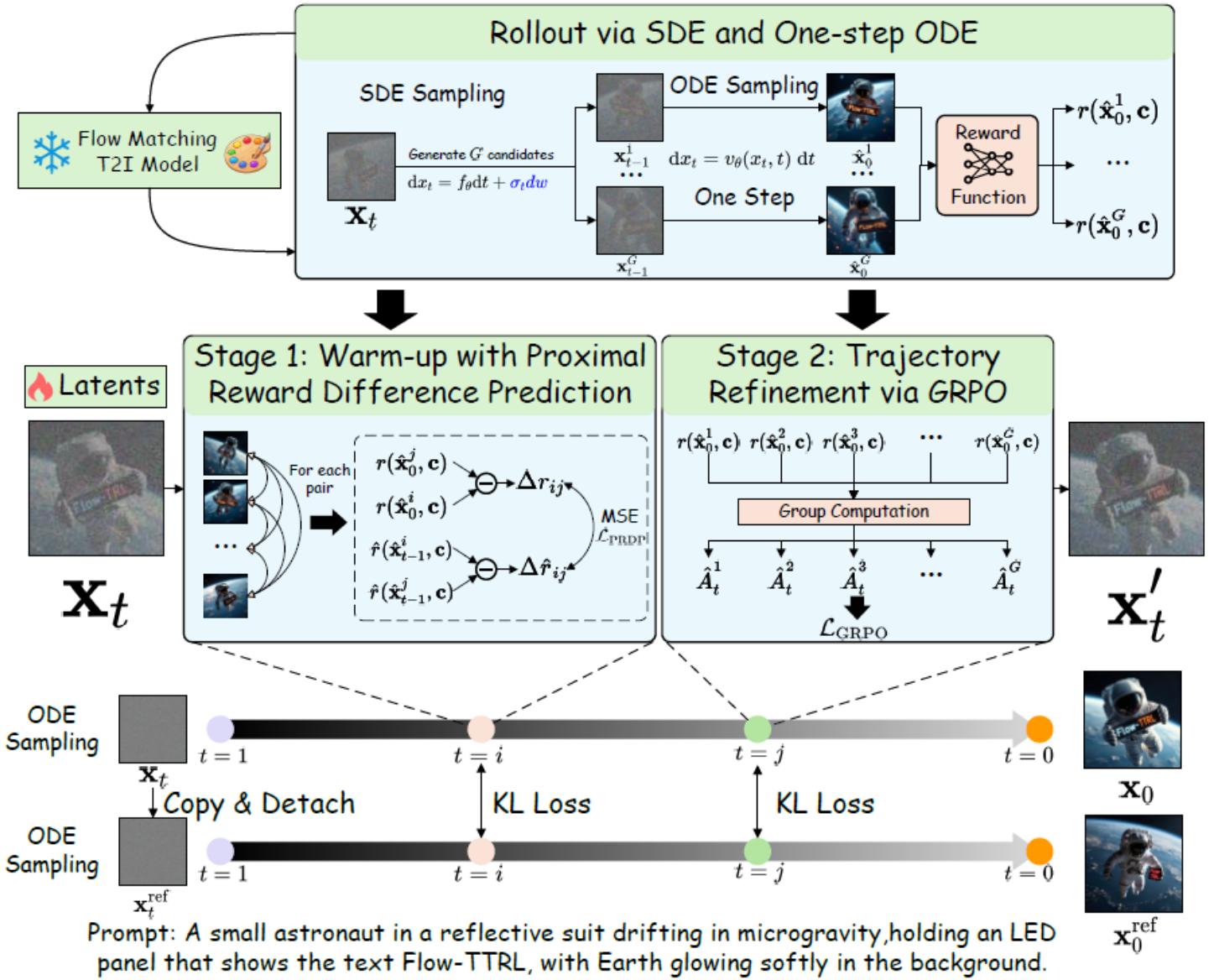


How to design a reinforcement learning mechanism that **dynamically navigates** the latent manifold to locate **high-reward regions** during inference **without modifying network weights**?

02

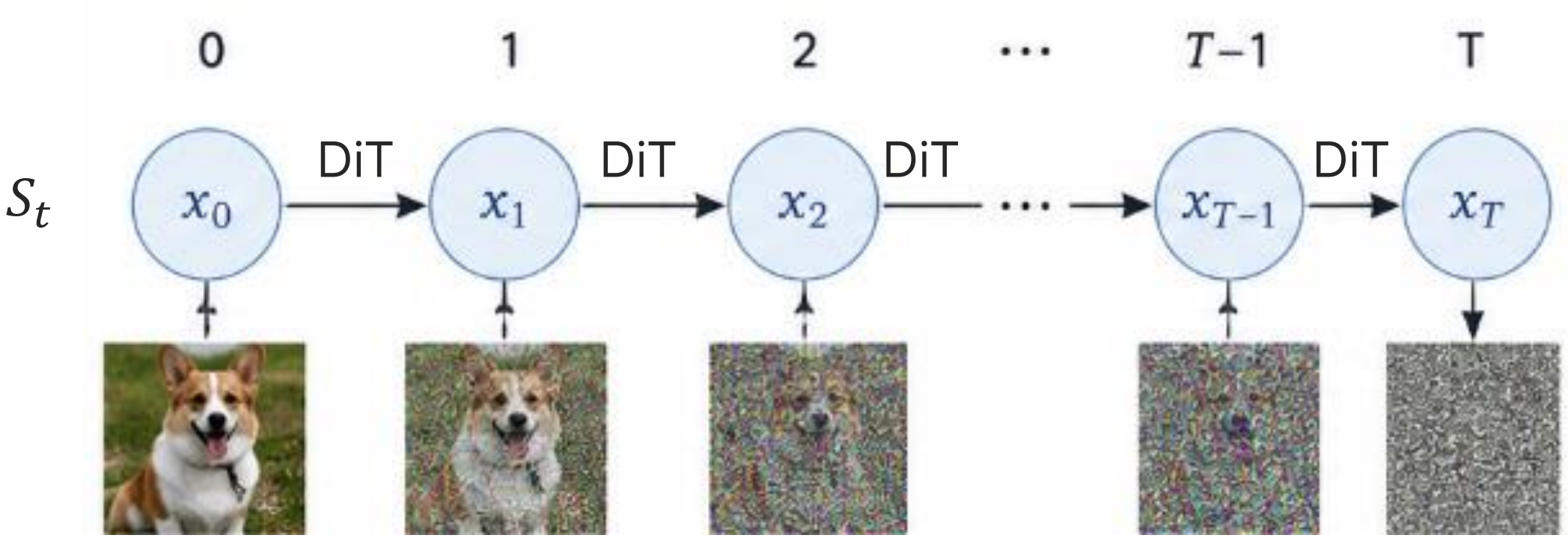
Flow-TTRL Framework

Overall Pipeline of Flow-TTRL



Prompt: A small astronaut in a reflective suit drifting in microgravity,holding an LED panel that shows the text Flow-TTRL, with Earth glowing softly in the background.

Reformulating Denoising as Markov Decision Process

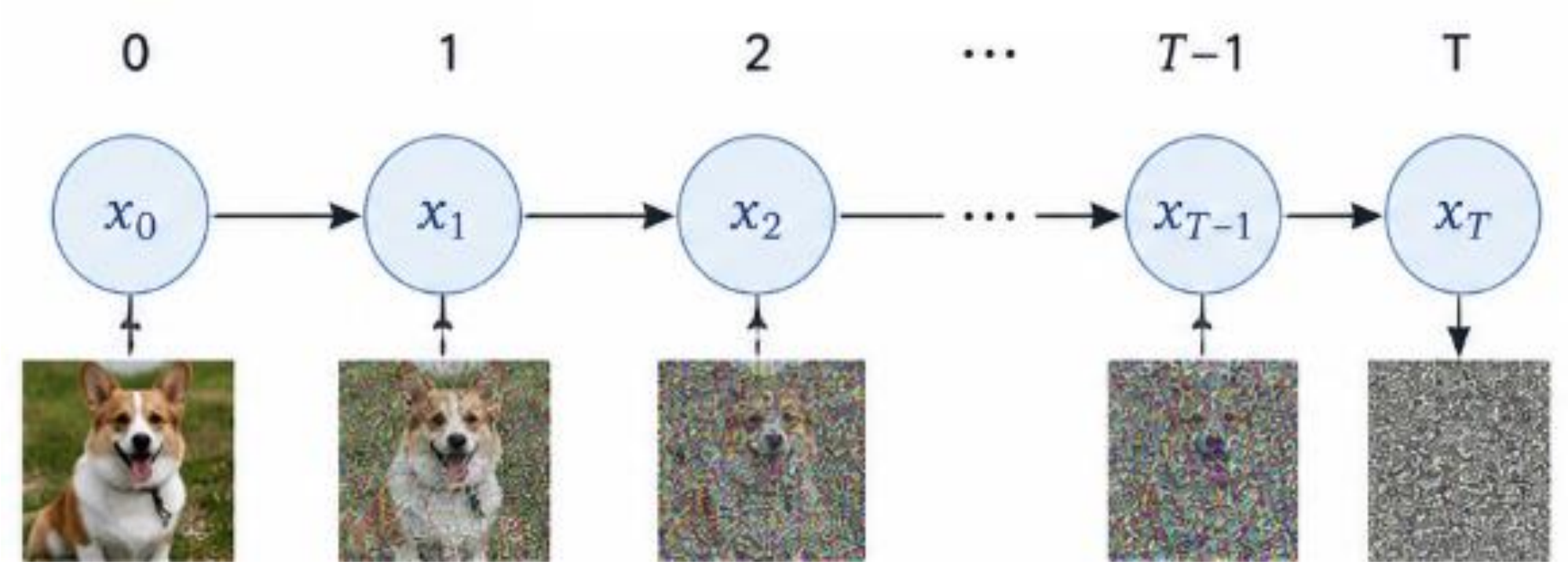


$$\pi(\mathbf{a}_t|s_t) \triangleq p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$$

$$P(s_{t+1}|s_t, \mathbf{a}_t) \triangleq (\delta_c, \delta_{t-1}, \delta_{\mathbf{x}_{t-1}})$$

Action Rectification through Latent Optimization

$$\mathbf{x}_{t-\Delta t} = \Phi(\mathbf{x}_t, v_\theta(\mathbf{x}_t, t)).$$

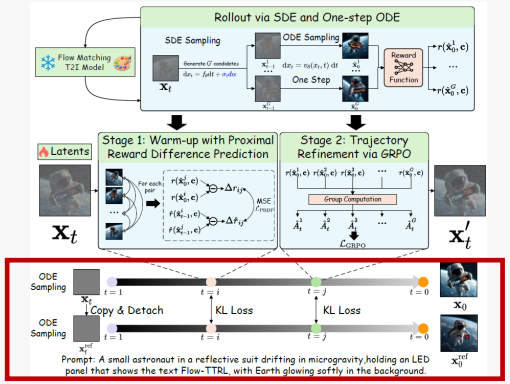


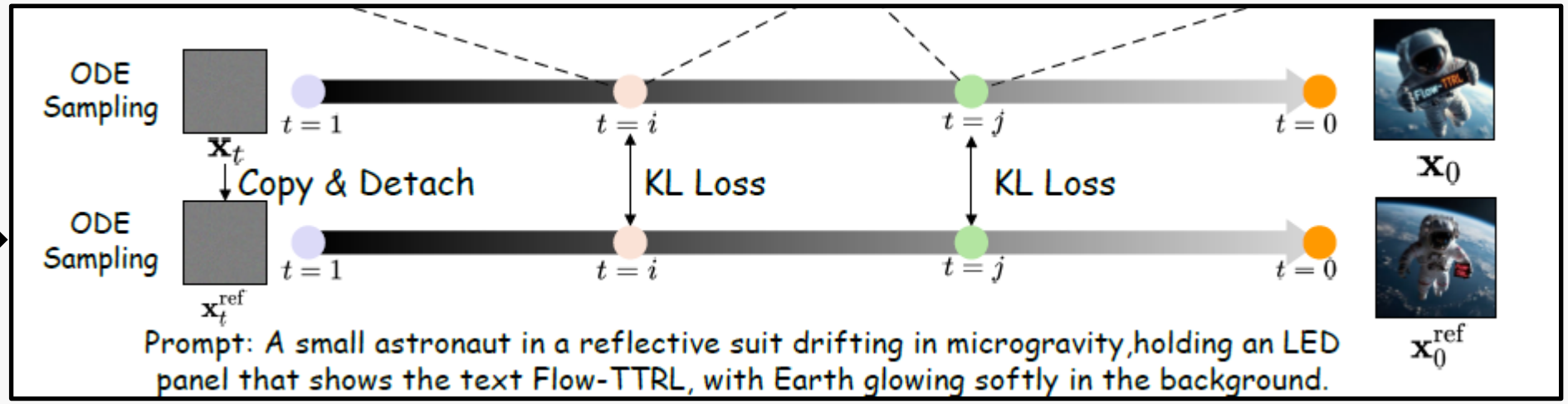
$$v_\theta(\mathbf{x}_t + \delta_t, t) \approx v_\theta(\mathbf{x}_t, t) + \underbrace{\nabla_{\mathbf{x}_t} v_\theta(\mathbf{x}_t, t) \cdot \delta_t}_{\text{Action Rectification}}.$$

$$S_{t-1} \rightarrow \text{DiT} \rightarrow a_t \xrightarrow{+\boldsymbol{\delta_t}} S_t \rightarrow \text{DiT} \rightarrow a_{t+1}$$

$$\mathbf{x}_t^* = \mathbf{x}_t + \eta \nabla_{\mathbf{x}_t} R(\Phi_\theta(\mathbf{x}_{T:0}, \mathbf{c})),$$

Reference Latent





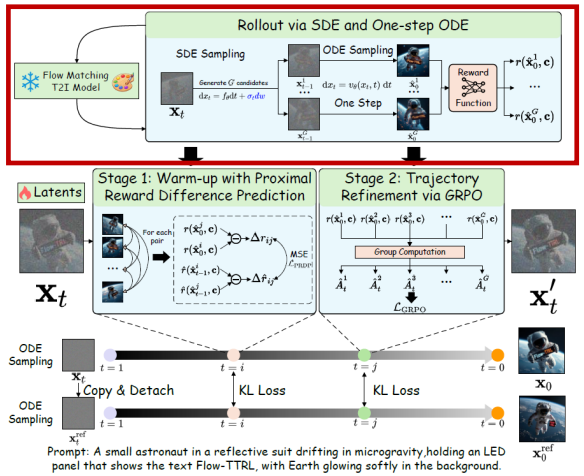
Traditional Global Parameter Update

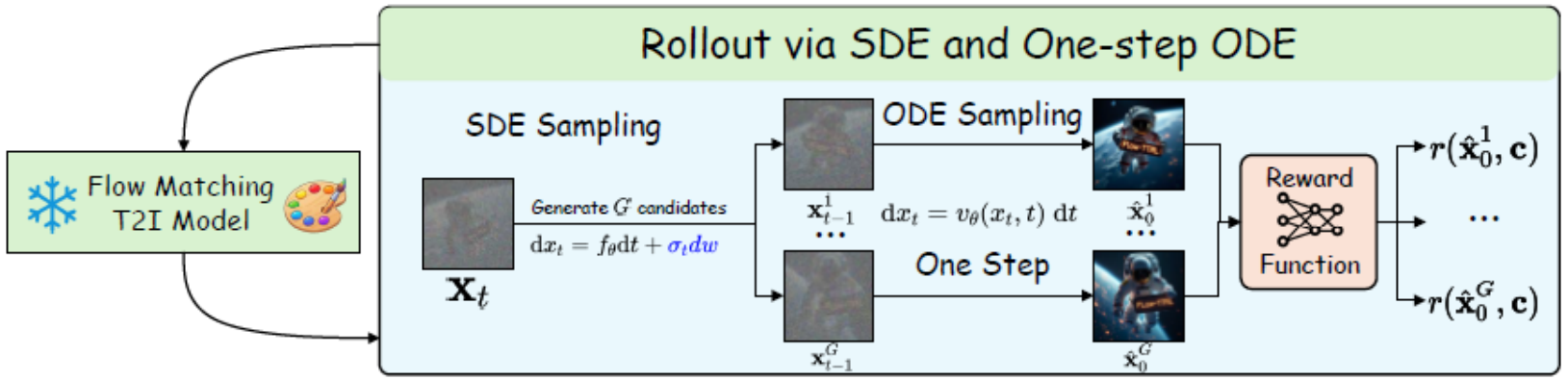
$$\max_{\pi_\theta} \ \mathbb{E}_{\mathbf{x}_0,\mathbf{c}}\Big[r(\mathbf{x}_0,\mathbf{c}) - \beta\,\mathrm{KL}\big(\pi_\theta(\mathbf{x}_0|\mathbf{c}) \,\|\, \pi_{\mathrm{ref}}(\mathbf{x}_0|\mathbf{c})\big)\Big],$$

Latent Update in Flow-TTRL

$$\max_{\mathbf{x}_t} \ \mathbb{E}_{\mathbf{x}_{t-1}\sim\pi(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{c})}\Big[r\big(\Phi_\theta(\mathbf{x}_{t-1:0},\mathbf{c})\big)\Big] - \beta\,\mathrm{KL}\Big(\pi(\cdot|\mathbf{x}_t,\mathbf{c}) \,\|\, \pi(\cdot|\mathbf{x}_t^{\mathrm{ref}},\mathbf{c})\Big).$$

Rollout and Optimization with the RL Objective





SDE Rollout

$$x_{t+\Delta t} = x_t + [v_\theta(x_t, t) + \frac{\sigma_t^2}{2t}(x_t + (1-t)v_\theta(x_t, t))]\Delta t + \sigma_t\sqrt{\Delta t}\epsilon,$$

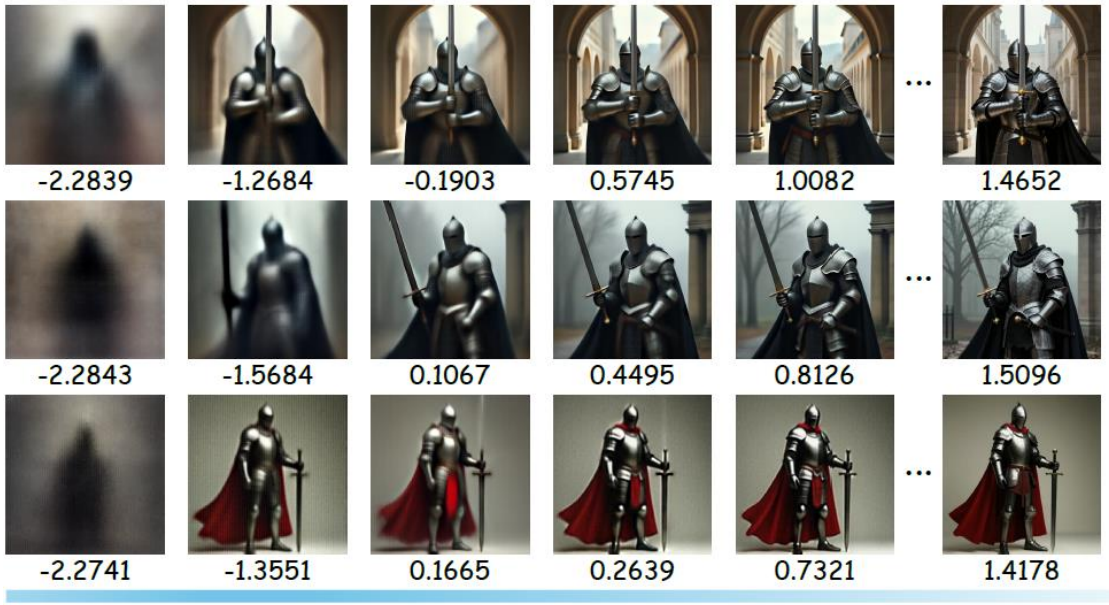
One-step ODE

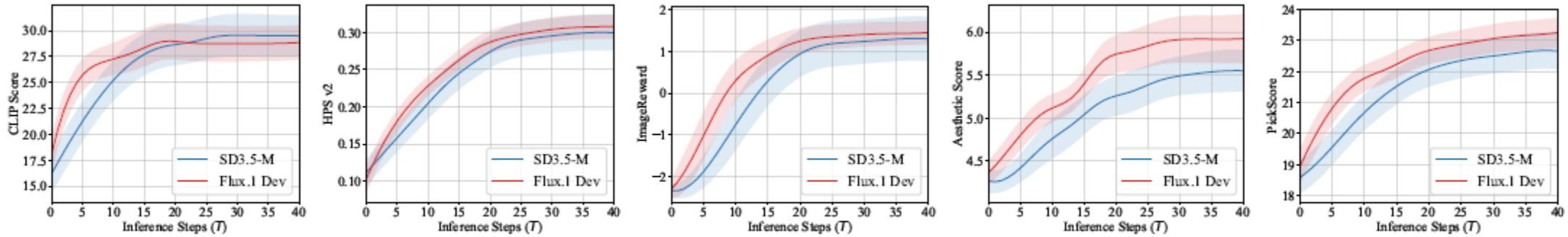
$$\hat{\mathbf{x}}_0^i = \mathbf{x}_{t-1}^i - (t-1)\cdot v_\theta(\mathbf{x}_{t-1}^i, t-1)$$

03

Two-Stage Optimization via PRDP and GRPO

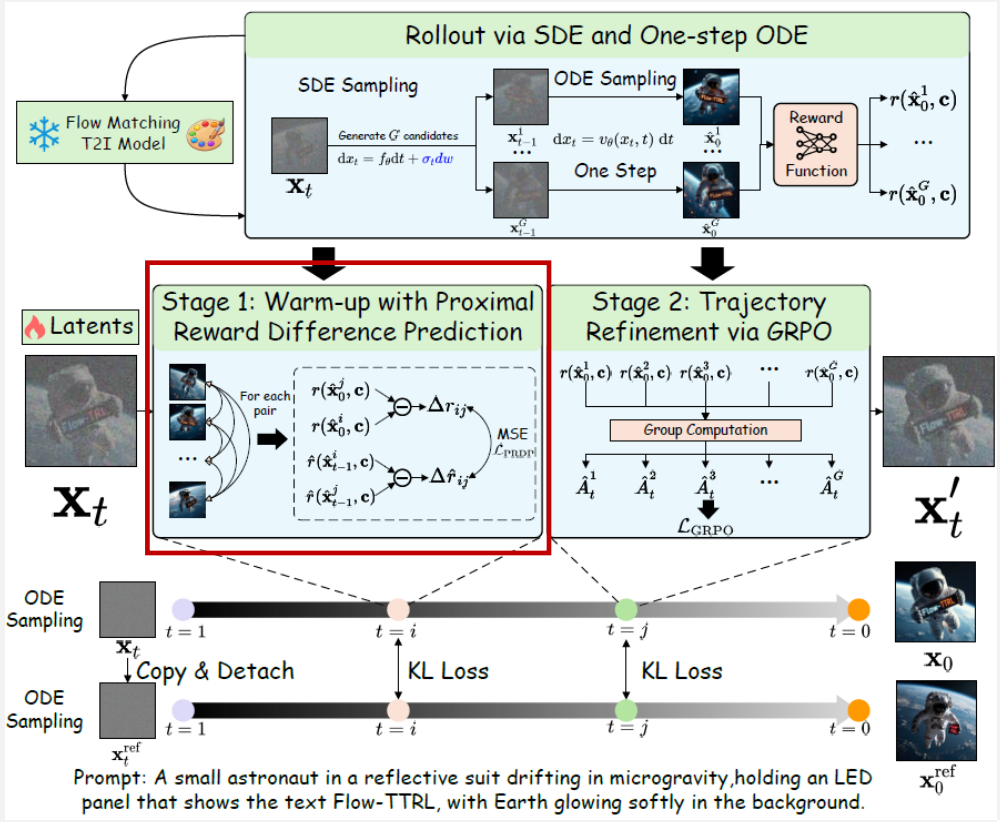
Rationale for Two-Stage PRDP and GRPO

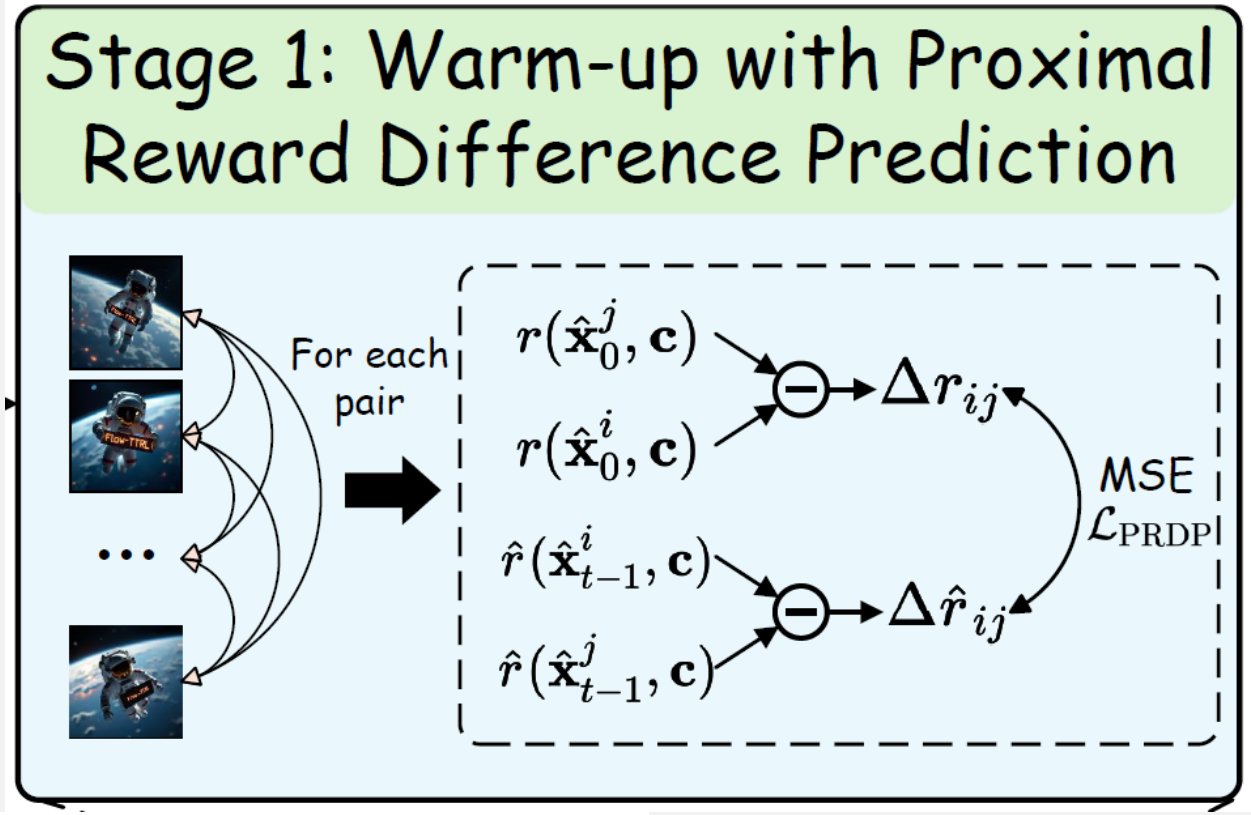




(a) CLIP Score *(b)* HPS v2 *(c)* ImageReward *(d)* Aesthetic *(e)* PickScore

Stage 1: Warm-up with Proximal Reward Difference Prediction





$$\hat{r}\left(\mathbf{x}_{t-1}, \mathbf{c}\right) \triangleq \log \frac{\pi\left(\mathbf{x}_{t-1} \middle| \mathbf{x}_t, \mathbf{c}\right)}{\pi\left(\mathbf{x}_{t-1} \middle| \mathbf{x}_t^{\text{ref}}, \mathbf{c}\right)}$$

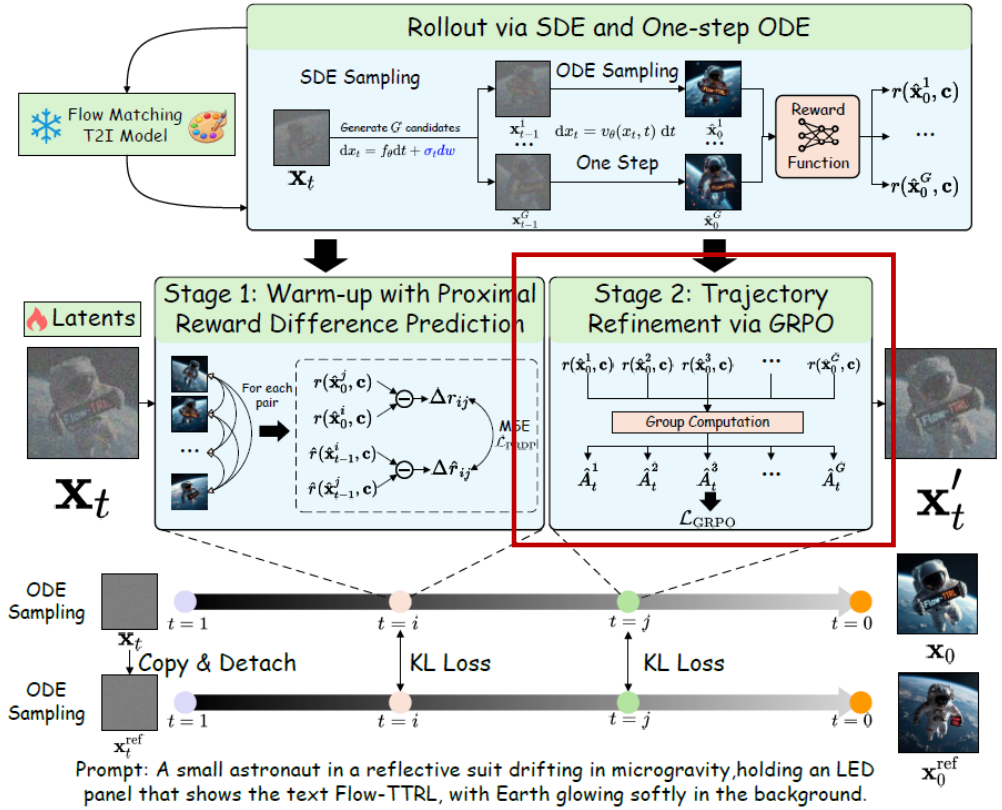
Define Reward

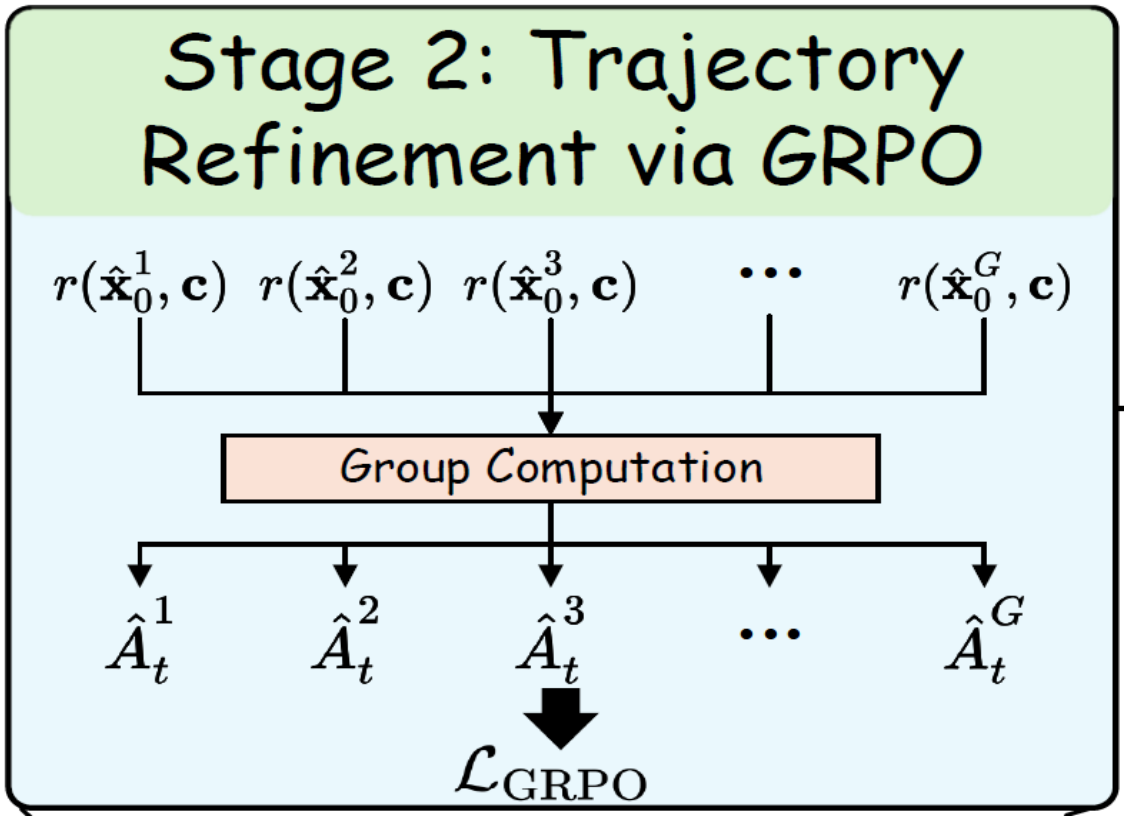
$$\mathcal{L}_{\text{PRDP}}(\mathbf{x}_t) = \mathbb{E}_{i,j}\left[\max\left(\left(\Delta\hat{r}_{ij} - \frac{\Delta r_{ij}}{\beta}\right)^2, \left(\text{clip}\left(\Delta\hat{r}_{ij}, \Delta\hat{r}_{ij}^{\text{old}} - \varepsilon, \Delta\hat{r}_{ij}^{\text{old}} + \varepsilon\right) - \frac{\Delta r_{ij}}{\beta}\right)^2\right)\right]$$

Final Loss

Proof in Appendix A.1.

Stage 2: Trajectory Refinement via Group Relative Policy Optimization





$$\hat{A}_t^i = \frac{r(\hat{\mathbf{x}}_0^i, \mathbf{c}) - \text{mean}(\{r(\hat{\mathbf{x}}_0^k, \mathbf{c})\}_{k=1}^G)}{\text{std}(\{r(\hat{\mathbf{x}}_0^k, \mathbf{c})\}_{k=1}^G) + \epsilon}.$$

Advantage Calculation

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{i\sim G}\Bigg[\max\left(-\hat{A}_t^i \rho_t^i, -\hat{A}_t^i \cdot \text{clip}(\rho_t^i, 1-\varepsilon, 1+\varepsilon)\right) + \beta \frac{\|\bar{\mathbf{x}}_{t-1}^i - \bar{\mathbf{x}}_{t-1}^{\text{ref}}\|^2}{2\sigma_t^2}\Bigg].$$

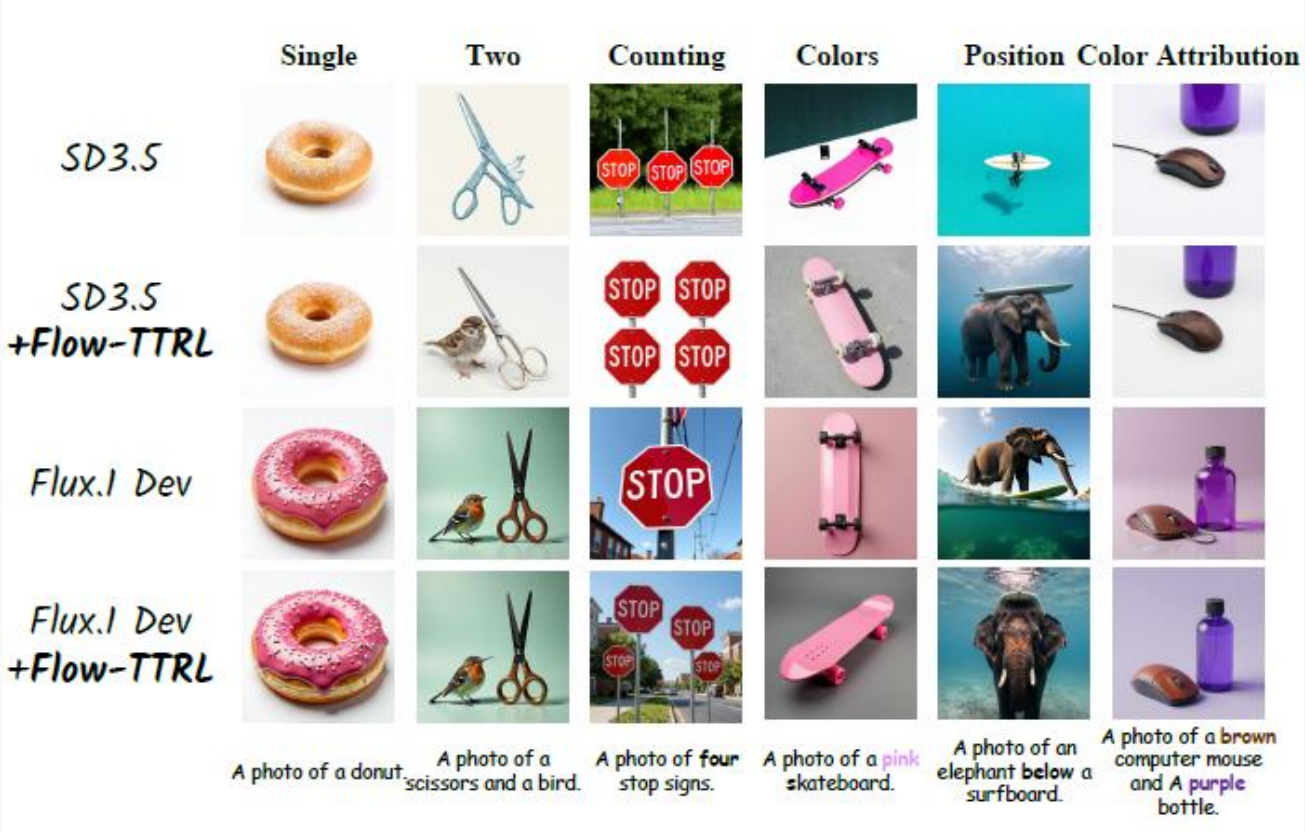
Final Loss

04

Experimental Validation

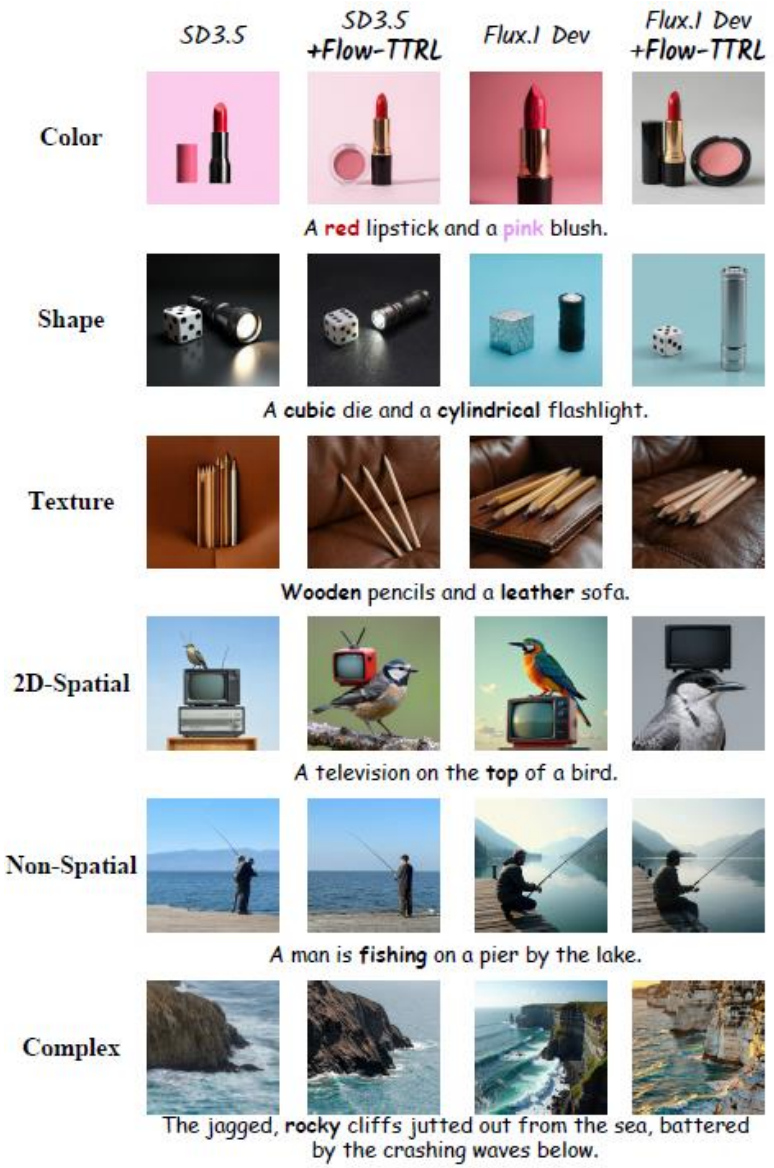
Results on GenEval

Model	Overall↑	Single↑	Two↑	Counting↑	Colors↑	Position↑	Color Attrib.↑
Diffusion Models							
DALL-E 2 (Ramesh et al., 2022)	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DALL-E 3 (Betker et al., 2023)	0.67	0.96	0.87	0.47	0.83	0.43	0.45
Autoregressive Models							
Emu3 (Wang et al., 2024)	0.54	0.98	0.71	0.34	0.81	0.17	0.21
Show-o (Xie et al., 2024)	0.53	0.95	0.52	0.49	0.82	0.11	0.28
GPT-4o (Hurst et al., 2024)	0.84	0.99	0.92	0.85	0.92	0.75	0.61
JanusFlow (Ma et al., 2025)	0.63	0.97	0.59	0.45	0.83	0.53	0.42
Janus-Pro-7B (Chen et al., 2025)	0.80	0.99	0.89	0.59	0.90	0.79	0.66
Flow Matching Models							
Flux.1 Dev (Labs et al., 2025)	0.66	0.98	0.81	0.74	0.79	0.22	0.45
SD 3.5-M (Esser et al., 2024)	0.63	0.98	0.78	0.50	0.81	0.24	0.52
SANA-1.5 4.8B (Xie et al., 2025)	0.81	0.99	0.93	0.86	0.84	0.59	0.65
SD3.5-L (Esser et al., 2024)	0.71	0.98	0.89	0.73	0.83	0.34	0.47
Test-time Optimization Methods							
SD 3.5-M + DAS (Kim et al., 2025)	0.83	**1.00**	0.98	0.89	0.86	0.57	0.67
SD 3.5-M + TITAN-Guide (Simon et al., 2025)	0.74	**1.00**	0.91	0.76	0.85	0.32	0.60
Flux.1 Dev + DAS (Kim et al., 2025)	0.80	**1.00**	0.96	0.81	0.90	0.37	0.73
Flux.1 Dev + TITAN-Guide (Simon et al., 2025)	0.76	0.99	0.90	0.83	0.88	0.34	0.64
Flow-RL based Methods							
SD 3.5-M + Flow-GRPO	0.95	**1.00**	0.99	0.95	0.92	**0.99**	0.86
SD 3.5-M + TempFlow-GRPO	**0.97**	**1.00**	**1.00**	**0.96**	**0.95**	**0.99**	**0.91**
Flux.1 Dev + Flow-TTRL	0.83	**1.00**	0.97	0.94	0.90	0.41	0.77
SD 3.5-M + Flow-TTRL	0.87	**1.00**	0.99	0.95	0.90	0.55	0.85



Results on T2I-CompBench

Model	Attribute Binding			Object Relationship		Complex
	Color↑	Shape↑	Texture↑	2D-Spatial↑	Non-Spatial↑	
DALL-E 3 (Betker et al., 2023)	0.7785	0.6205	0.7036	0.2865	0.3003	0.3773
Janus-Pro-7B (Chen et al., 2025)	0.5145	0.3323	0.4069	0.1566	0.3137	—
Emu3 (Wang et al., 2024)	0.7913	0.5846	0.7422	—	—	—
Flux.1 Dev (Labs et al., 2025)	0.7407	0.5718	0.6922	0.2863	0.3127	0.3703
SD 3.5-M (Esser et al., 2024)	0.7994	0.5669	0.7338	0.2850	0.3146	0.3542
SD 3.5-M + Flow-GRPO (Liu et al., 2025a)	0.8379	0.6130	0.7236	**0.5447**	0.3195	—
SD 3.5-M + DAS (Kim et al., 2025)	0.8561	0.6482	0.7717	0.3679	0.3294	0.3847
SD 3.5-M + TITAN-Guide (Simon et al., 2025)	0.8468	0.6235	0.7689	0.3725	0.3276	0.3791
Flux.1 Dev + DAS (Kim et al., 2025)	0.8371	0.6779	0.7689	0.3662	0.3161	0.3951
Flux.1 Dev + TITAN-Guide (Simon et al., 2025)	0.8217	0.6511	0.7673	0.3786	0.3154	0.3853
Flux.1 Dev + Flow-TTRL	0.8804	0.6717	0.7958	0.4390	0.3229	**0.4179**
SD 3.5-M + Flow-TTRL	**0.9042**	**0.7361**	**0.8261**	0.4414	**0.3319**	0.4045



Results on PartiPrompts、Pick-a-Pic、Drawbench

PartiPrompts

Model	Human Preference Alignment			Aesthetic Quality	T2I Alignment
	PickScore↑	HPS v2↑	ImageReward↑	Aesthetic↑	CLIP↑
SD 3.5 Medium	22.21	0.275	0.890	5.442	28.01
Flux.1 Dev	22.84	0.316	1.205	5.843	27.27
SD 3.5 Medium + BoN (5min)	22.50	0.289	1.112	5.532	28.84
Flux.1 Dev + BoN (5min)	23.02	0.318	1.308	5.960	27.98
SD 3.5 Medium + Flow-TTRL	22.69	0.301	1.365	5.541	**29.27**
Flux.1 Dev + Flow-TTRL	**23.17**	**0.323**	**1.472**	**5.963**	28.51

Pick-a-Pic

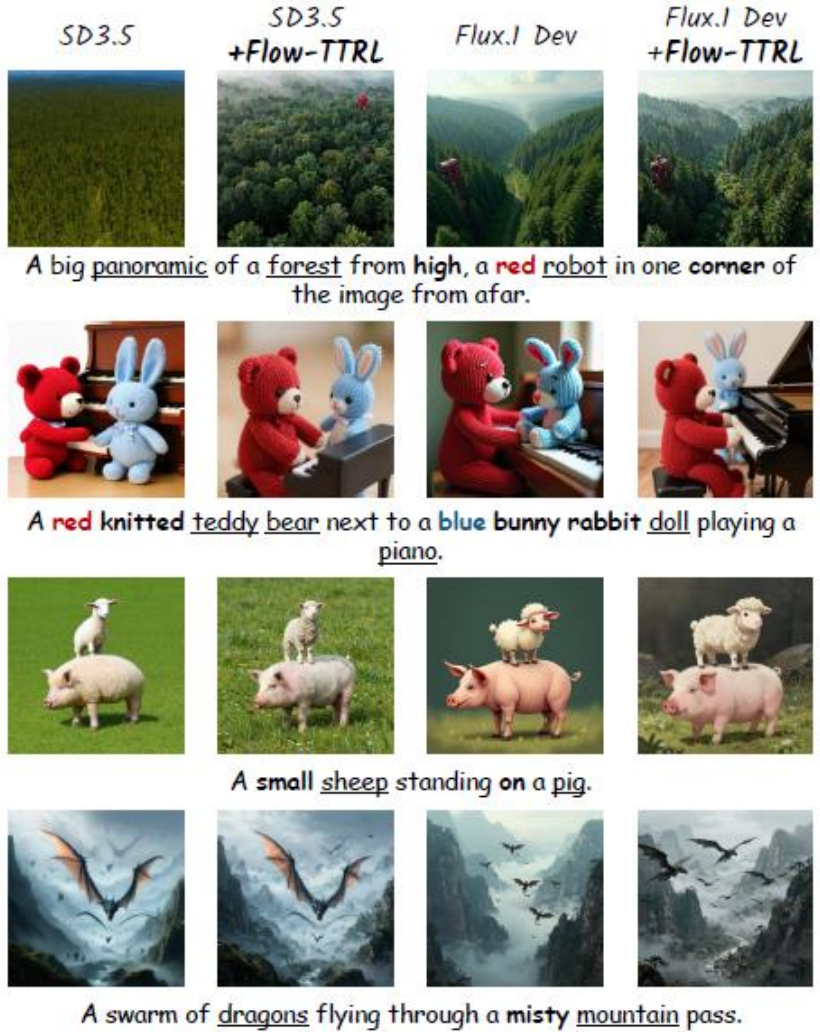
Model	Human Preference Alignment			Aesthetic Quality	T2I Alignment
	PickScore↑	HPS v2↑	ImageReward↑	Aesthetic↑	CLIP↑
SD 3.5 Medium	21.89	0.285	0.727	5.631	26.92
Flux.1 Dev	22.35	0.316	0.929	6.086	26.02
SD 3.5 Medium + BoN (5min)	22.01	0.289	0.839	5.735	27.47
Flux.1 Dev + BoN (5min)	22.51	0.314	1.081	6.180	26.86
SD 3.5 Medium + Flow-TTRL	22.24	0.299	1.187	5.733	**28.09**
Flux.1 Dev + Flow-TTRL	**22.67**	**0.322**	**1.340**	**6.185**	27.37

Drawbench

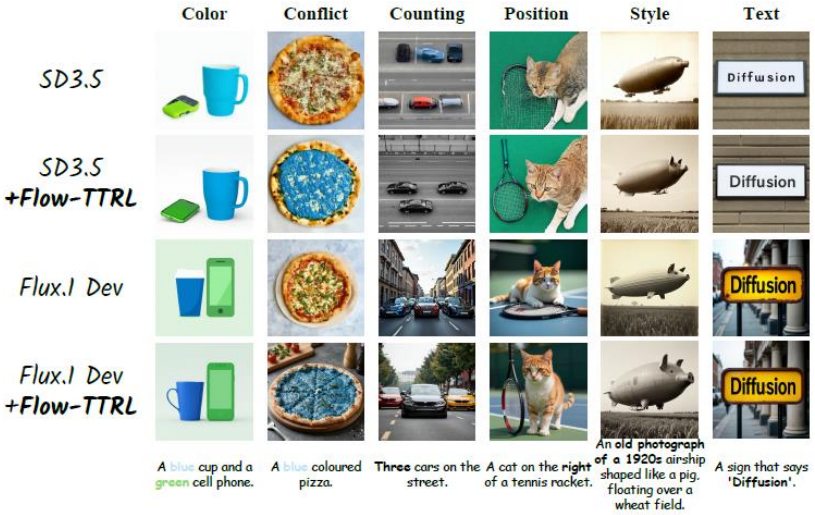
Model	Human Preference Alignment			Aesthetic Quality	T2I Alignment
	PickScore↑	HPS v2↑	ImageReward↑	Aesthetic↑	CLIP↑
SD 3.5 Medium	22.14	0.272	0.630	5.181	28.87
Flux.1 Dev	22.59	0.305	0.931	5.722	27.30
SD 3.5 Medium + BoN (5min)	22.27	0.275	0.757	5.199	29.82
Flux.1 Dev + BoN (5min)	22.77	0.308	1.048	**5.812**	28.03
SD 3.5 Medium + Flow-TTRL	22.50	0.288	1.082	5.257	**30.12**
Flux.1 Dev + Flow-TTRL	**23.03**	**0.315**	**1.252**	5.778	29.00

Results on PartiPrompts、Pick-a-Pic、Drawbench

PartiPrompts



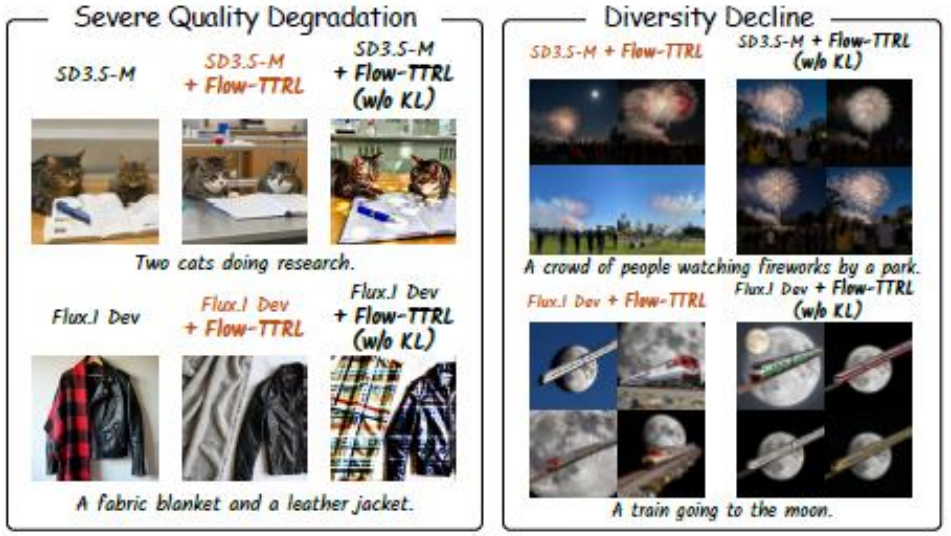
Pick-a-Pic

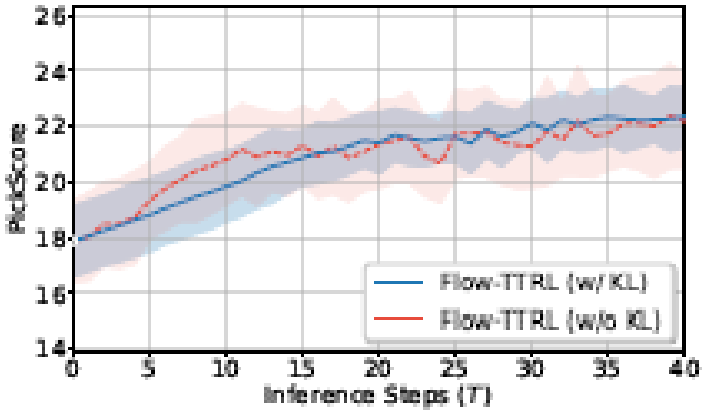


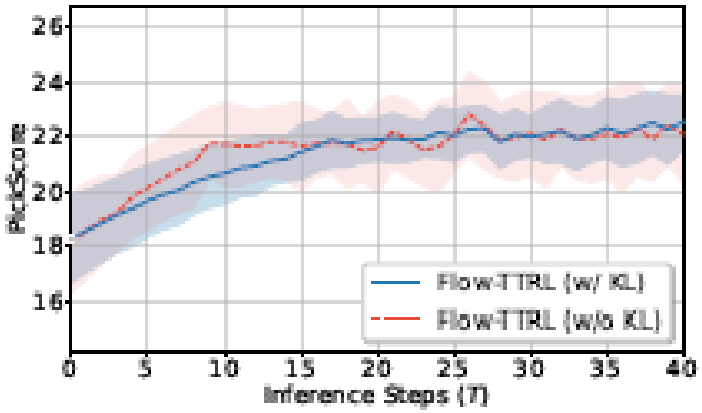
Drawbench

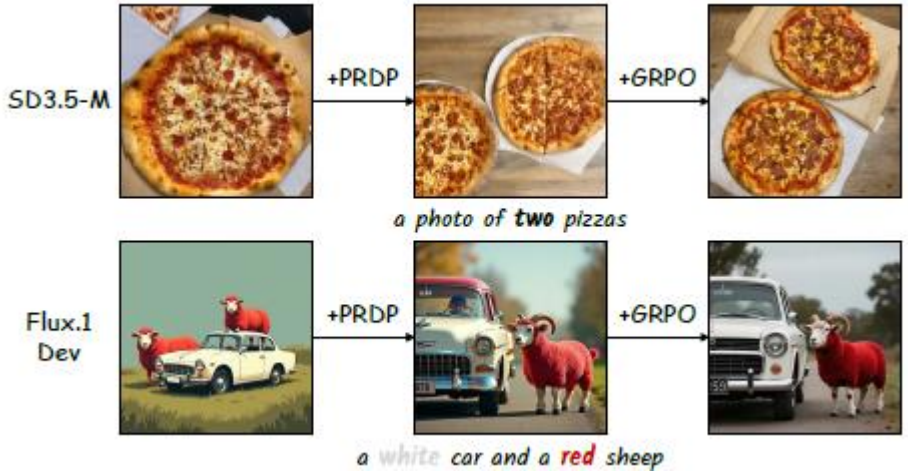
Ablation Results

Model	GenEval↑	Complex↑	Human Preference Alignment			Aesthetic Quality
			PickScore↑	HPS v2↑	ImageReward↑	Aesthetic↑
SD3.5-M + Flow-TTRL	**0.87**	**0.4045**	**22.24**	**0.299**	1.187	5.733
SD3.5-M + Flow-TTRL (w/o KL)	0.71	0.3569	22.11	0.287	0.934	5.328
SD3.5-M + Flow-TTRL (w/o PRDP)	0.81	0.3792	22.17	0.291	1.221	5.553
SD3.5-M + Flow-TTRL (w/o GRPO)	0.72	0.3612	22.07	0.289	0.896	5.721
SD3.5-M + Flow-TTRL (w/o NR)	0.85	0.3895	22.13	0.294	**1.243**	**5.892**
SD3.5-M	0.63	0.3542	21.89	0.285	0.727	5.631
Flux.1 Dev + Flow-TTRL	**0.83**	**0.4179**	22.67	**0.322**	**1.340**	6.185
Flux.1 Dev + Flow-TTRL (w/o KL)	0.69	0.3697	22.39	0.304	1.114	5.893
Flux.1 Dev + Flow-TTRL (w/o PRDP)	0.76	0.3861	22.61	0.302	1.307	5.981
Flux.1 Dev + Flow-TTRL (w/o GRPO)	0.70	0.3765	22.41	0.317	0.897	6.034
Flux.1 Dev + Flow-TTRL (w/o NR)	0.82	0.3997	**22.69**	0.316	1.329	**6.235**
Flux.1 Dev	0.66	0.3703	22.35	0.316	0.929	6.086

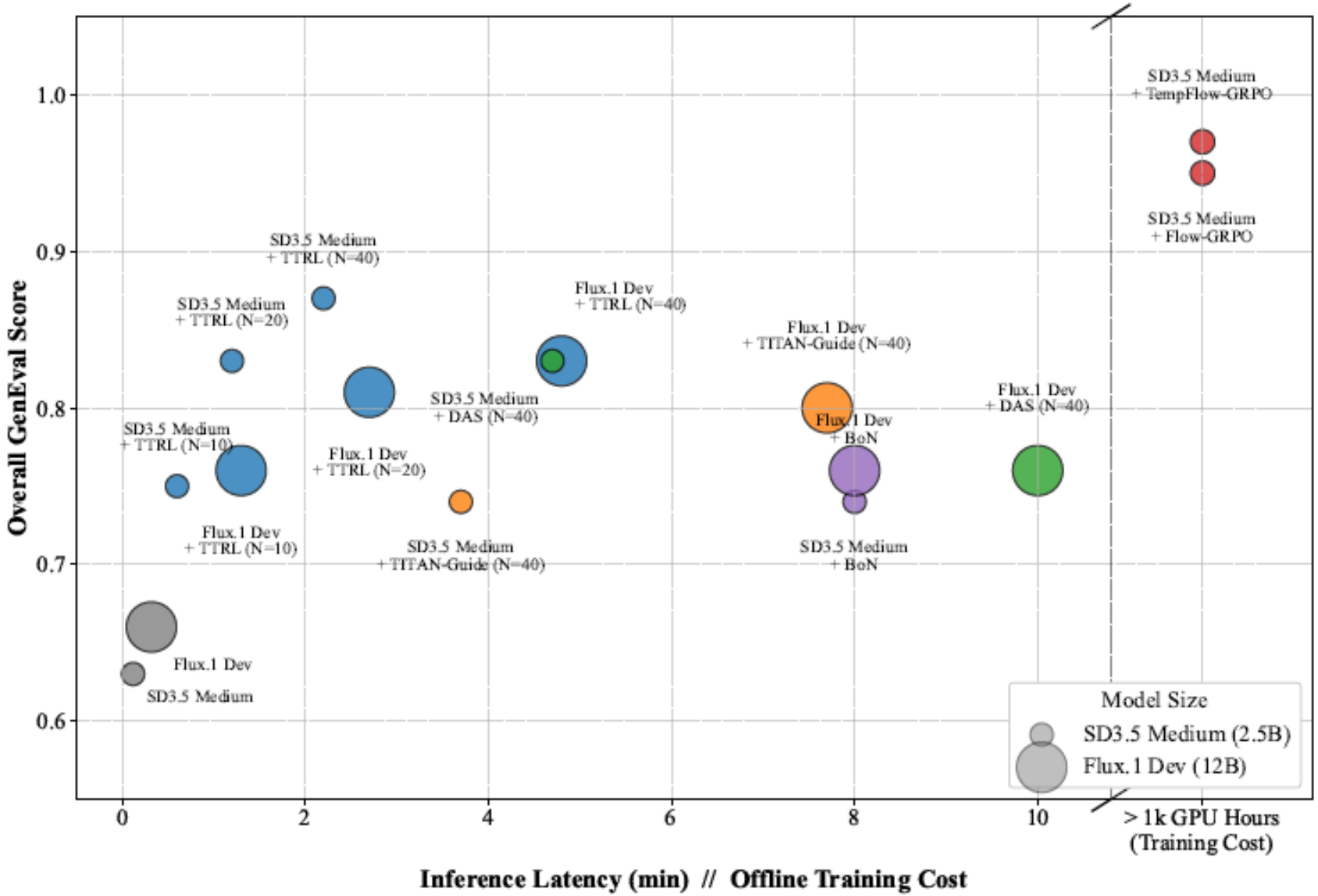




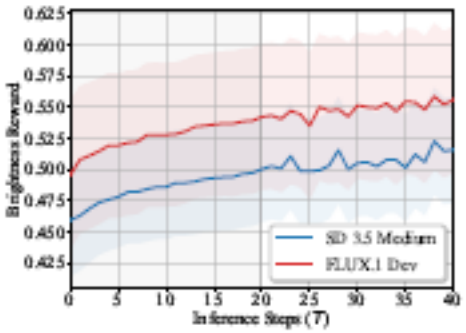




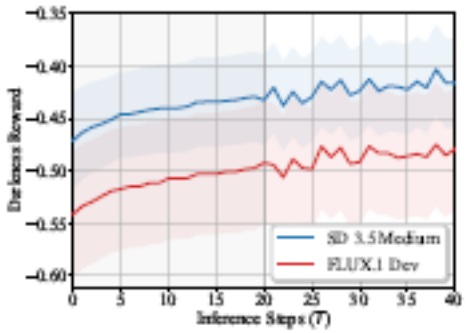
Complexity


Overall GenEval Score
1.0
0.9
0.8
0.7
0.6
0
2
4
6
8
10
> 1k GPU Hours
(Training Cost)
Inference Latency (min) // Offline Training Cost
SD3.5 Medium
+ TempFlow-GRPO
SD3.5 Medium
+ Flow-GRPO
SD3.5 Medium
+ TTRL (N=40)
SD3.5 Medium
+ TTRL (N=20)
SD3.5 Medium
+ TTRL (N=10)
SD3.5 Medium
+ DAS (N=40)
Flux.1 Dev
+ TTRL (N=40)
Flux.1 Dev
+ TTRL (N=20)
Flux.1 Dev
+ TTRL (N=10)
SD3.5 Medium
+ TITAN-Guide (N=40)
Flux.1 Dev
+ TITAN-Guide (N=40)
Flux.1 Dev
+ BoN
SD3.5 Medium
+ BoN
Flux.1 Dev
+ DAS (N=40)
Flux.1 Dev
SD3.5 Medium
Model Size
SD3.5 Medium (2.5B)
Flux.1 Dev (12B)

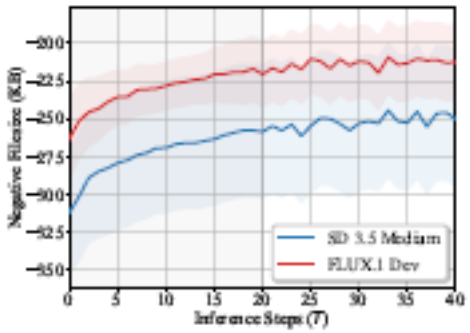
Experiments on Verifiable Rewards



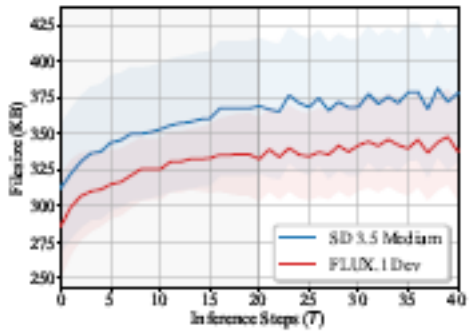
(a) Brightness



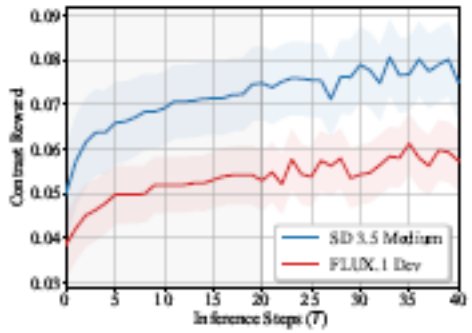
(b) Darkness



(c) Compressibility



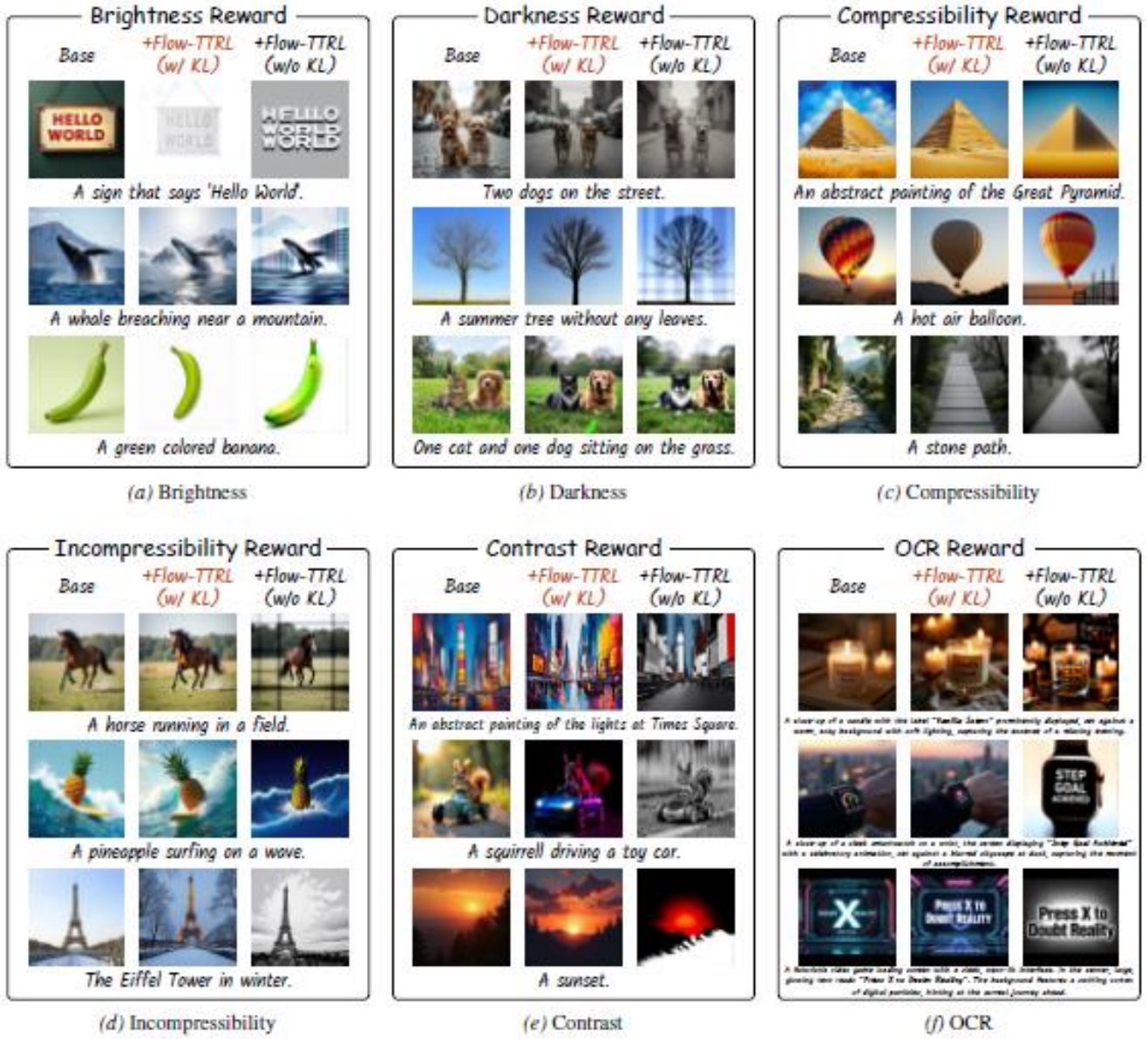
(d) Incompressibility



(e) Contrast

Metric	SD 3.5	SD 3.5 + Flow-TTRL	Flux.1 Dev	Flux.1 Dev + Flow-TTRL	Flow-GRPO
OCR Score ↑	0.59	0.78	0.70	0.84	**0.91**

Experiments on Verifiable Rewards

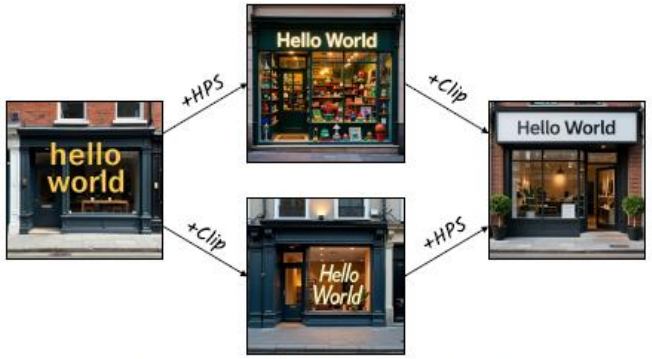


(a) Brightness (b) Darkness (c) Compressibility

(d) Incompressibility (e) Contrast (f) OCR

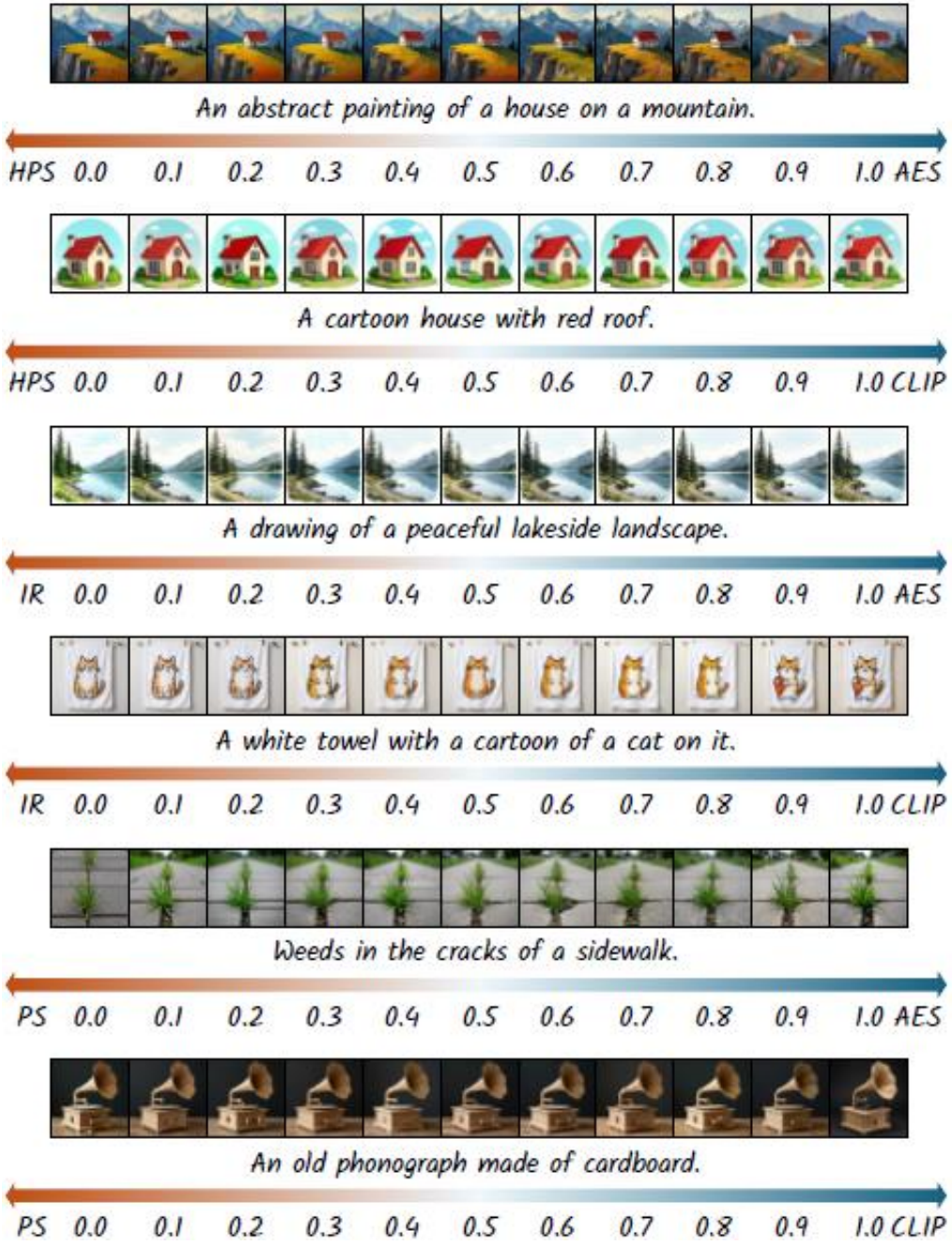
Multi-Reward Combination

Model	Human Preference Alignment			Aesthetic Quality	T2I Alignment
	PickScore↑	HPS v2↑	ImageReward↑	Aesthetic↑	CLIP↑
SD 3.5 Medium	22.14	0.272	0.630	5.181	28.87
+ Flow-TTRL	**22.50**	**0.288**	**1.082**	5.257	30.12
+ Flow-TTRL w/o HPS v2	22.17	0.268	0.723	5.093	**30.59**
+ Flow-TTRL w/o CLIPScore	22.39	0.287	0.981	**5.347**	28.91
Flux.1 Dev	22.59	0.305	0.931	5.722	27.30
+ Flow-TTRL	**23.03**	**0.315**	1.252	**5.778**	**29.00**
+ Flow-TTRL w/o HPS v2	22.62	0.304	0.912	5.683	28.83
+ Flow-TTRL w/o CLIPScore	22.83	0.314	**1.325**	5.707	27.43

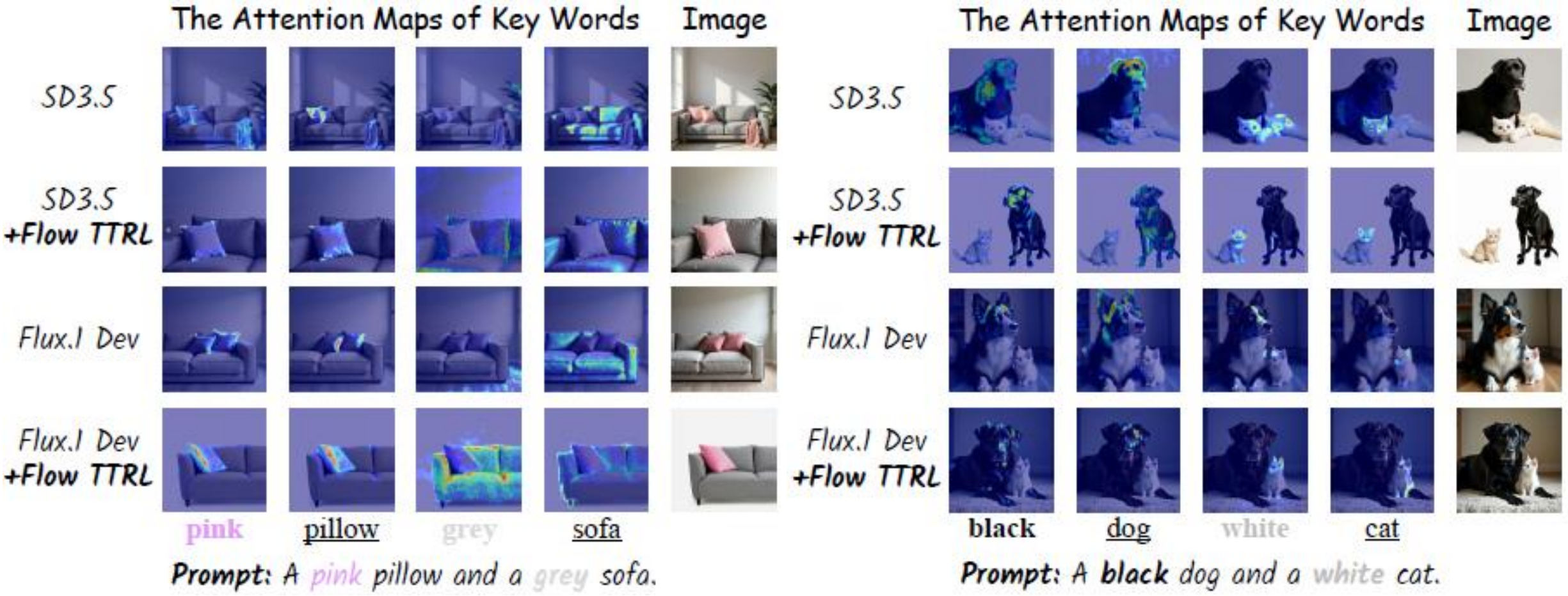


A storefront with 'Hello World' written on it.

In late afternoon in January in New England, a man stands in the shadow of a maple tree.



Attention Visualization



05

Main Contributions and Conclusions

Main Contributions and Conclusions

01

The First Test-Time Reinforcement Learning Framework

Flow-TTRL reinterprets latent trajectories as an implicit policy to enable flexible alignment without the need for traditional fine-tuning or curated training data.

02

Coarse-to-fine Trajectory Optimization

Flow-TTRL employs a coarse-to-fine strategy (PRDP followed by GRPO) that balances structural stability with precise aesthetic refinement during the trajectory optimization process.

03

Competitive Training-Free Performance

We demonstrate that Flow-TTRL achieves highly competitive results on GenEval and T2I-CompBench, yielding performance comparable to established RL-based finetuning methods and proprietary models. Furthermore, it consistently enhances human preference alignment and image fidelity across all five evaluated datasets.

Thanks

Presenter： Jili Chen

Code

Contact