Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models

Authors: Jiajun Fan, Tong Wei, Chaoran Cheng, Yuxin Chen, Ge Liu

Presenter: Jiajun Fan

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Introduction

Pre-training

Knowledge Accumulation

Post-Training

Instruction Finetuning

Instruction Following

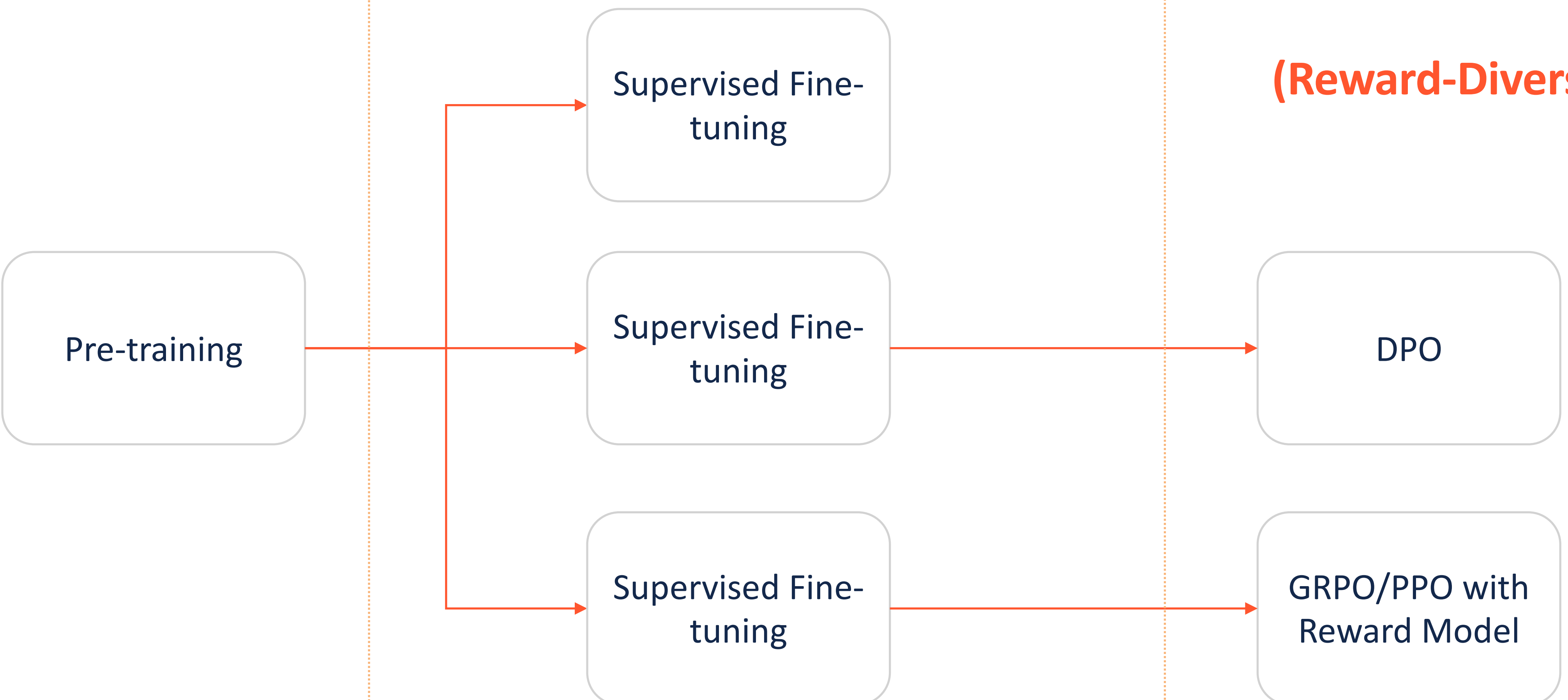
RL Post-training (Our Focus)

Text-Image Alignment

Reasoning Capability

Guided Generation

(Reward-Diversity Trade-off)



Current Limitations & Motivation

RLHF

The conventional RL objective in Equation (1) can be rewritten as a combination of two loss terms:

$$\mathcal{L}_{\text{RLHF}}(\theta) = \mathcal{L}_{\text{RL}}(\theta) + \beta \cdot \mathcal{L}_{\text{D}}(\theta) \tag{2}$$

LLM Fine-tuning

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathcal{L}_{\text{PG}}(\theta) + \beta \cdot D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \tag{3}$$

where $\mathcal{L}_{\text{PG}}(\theta)$ represents a clipped policy gradient loss based on group-level advantage estimation.

Diffusion/Flow Fine-tuning

$$\mathcal{L}_{\text{ORW-CFM-W2}}(\theta) = \mathcal{L}_{\text{ORW}} + \beta \cdot \mathbb{E}_{c,t,x_t}[|\mathbf{v}_\theta(x_t,t,c) - \mathbf{v}_{\text{ref}}(x_t,t,c)|^2]$$

where $\mathcal{L}_{\text{ORW}} = \mathbb{E}_{c,x_1,t,x_t}[\omega(x_1,c) * |\mathbf{v}_\theta(x_t,t,c) - \mathbf{u}_t|^2]$ is the reward weighted loss and $\mathbf{v}_\theta$ and $\mathbf{v}_{\text{ref}}$ are the velocity fields of the fine-tuned and reference policies, respectively.

Fixed Divergence Regularization

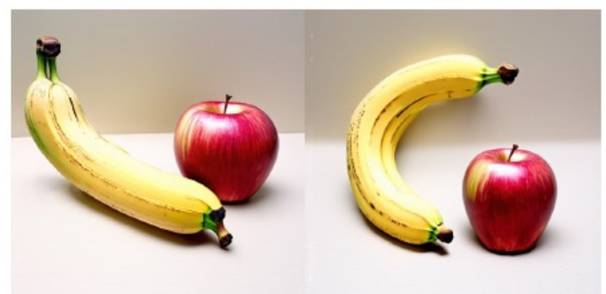
Methods with fixed regularization suffer from over-optimization without diversity (e.g.$\beta \to 0$) or under-optimization (e.g., $\beta \to \infty$)

Can we find a way to adaptively adjust β? Yes! Try our ADRPO

Diversity & Quality Collapse

SD3

ORW-CFM-W2



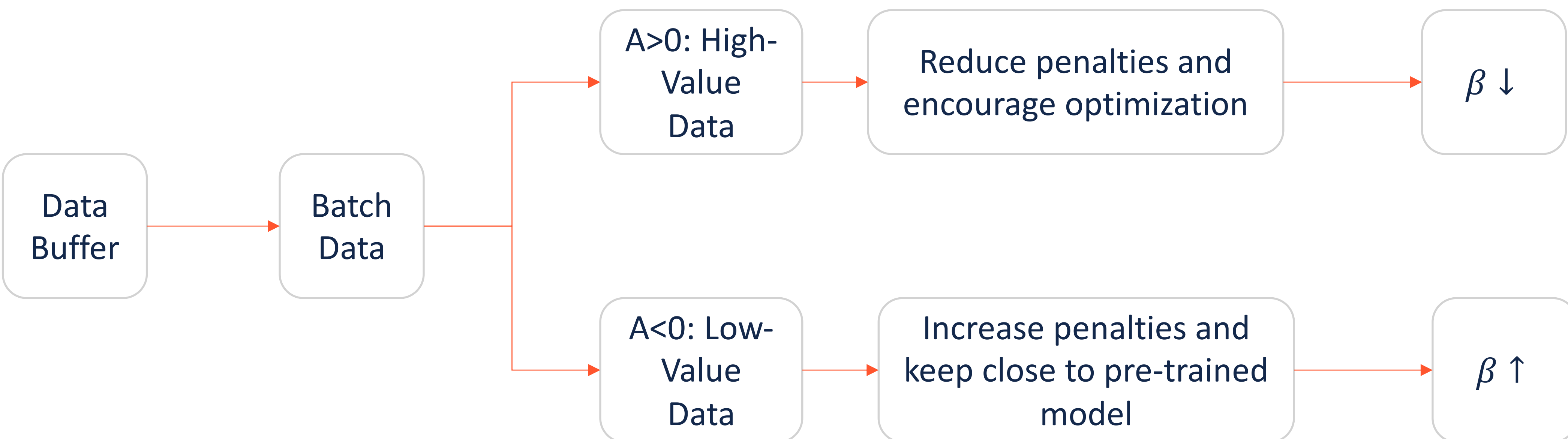
ADRPO for Flow Matching

$$\mathcal{L}_{\text{RLHF}}(\theta) = \mathcal{L}_{\text{RL}}(\theta) + \beta \cdot \mathcal{L}_{\text{D}}(\theta) \tag{2}$$

Can we find a way to adaptively adjust β without extra computation?

$$\mathcal{L}_{\text{ADRPO}}(\theta) = \mathcal{L}_{\text{RL}}(\theta) + (\beta_0 - A) \cdot \mathcal{L}_{\text{D}}(\theta)$$

$$\begin{aligned}\mathcal{L}_{\text{ADRPO-FM}}(\theta) = \mathbb{E}_{c\sim p(c),t\sim U(0,1),x_1\sim p_\theta^{n-1},x_t\sim p_t(x_t|x_1,c)}&[A(x_1,c)\cdot|\mathbf{v}_\theta(x_t,t,c)-\mathbf{u}_t|^2] \\ +(\beta_0 - A(x_1,c))\cdot\mathbb{E}_{c,t,x_t}&[|\mathbf{v}_\theta(x_t,t,c)-\mathbf{v}_{\text{ref}}(x_t,t,c)|^2]\end{aligned} \tag{6}$$



Experimental Results

Table 1: Comparison of text-to-image generation methods across different evaluation metrics. Best scores are highlighted in blue, second-best in green. We report standard errors estimated over 3 random seeds. ClipDiversity measures the mean pairwise distance of CLIP embeddings [25, 14].

Method	Task Metrics		Image Quality	Human Preference		
	ClipScore↑ [25]	ClipDiversity↑ [25]	Aesthetic↑ [39]	BLIPScore↑ [12]	ImageReward↑ [39]	PicScore↑ [21]
Base Model						
SD3 (2B) [13]	29.27±0.42	5.08±0.52	5.53±0.09	0.501±0.007	0.97±0.13	20.81±0.09
Other Flow Matching Models						
FLUX.1-Dev (12B) [43]	31.72±0.48	4.29±0.42	5.95±0.05	0.492±0.004	1.11±0.10	21.83±0.11
SANA-1.5 (4.8B) [38]	32.18±0.36	4.31±0.50	5.89±0.12	0.526±0.006	1.45±0.08	21.85±0.15
SD3 Fine-tuning Methods						
SD3+RAFT [9]	29.35±0.27	1.85±0.19	4.54±0.04	0.512±0.001	0.22±0.08	19.21±0.02
SD3+DPO [37]	31.30±0.52	4.78±0.46	5.82±0.05	0.509±0.005	1.48±0.10	21.31±0.10
SD3+ORW-CFM-W2 [14]	31.42±0.39	3.86±0.37	5.29±0.05	0.542±0.006	1.22±0.10	20.97±0.11
SD3+ADRPO (Ours)	32.97±0.46	5.13±0.47	6.27±0.06	0.567±0.004	1.61±0.05	22.78±0.15

Base Model: SD3 (2B)
Reward Model: Clip Score

Experimental Results

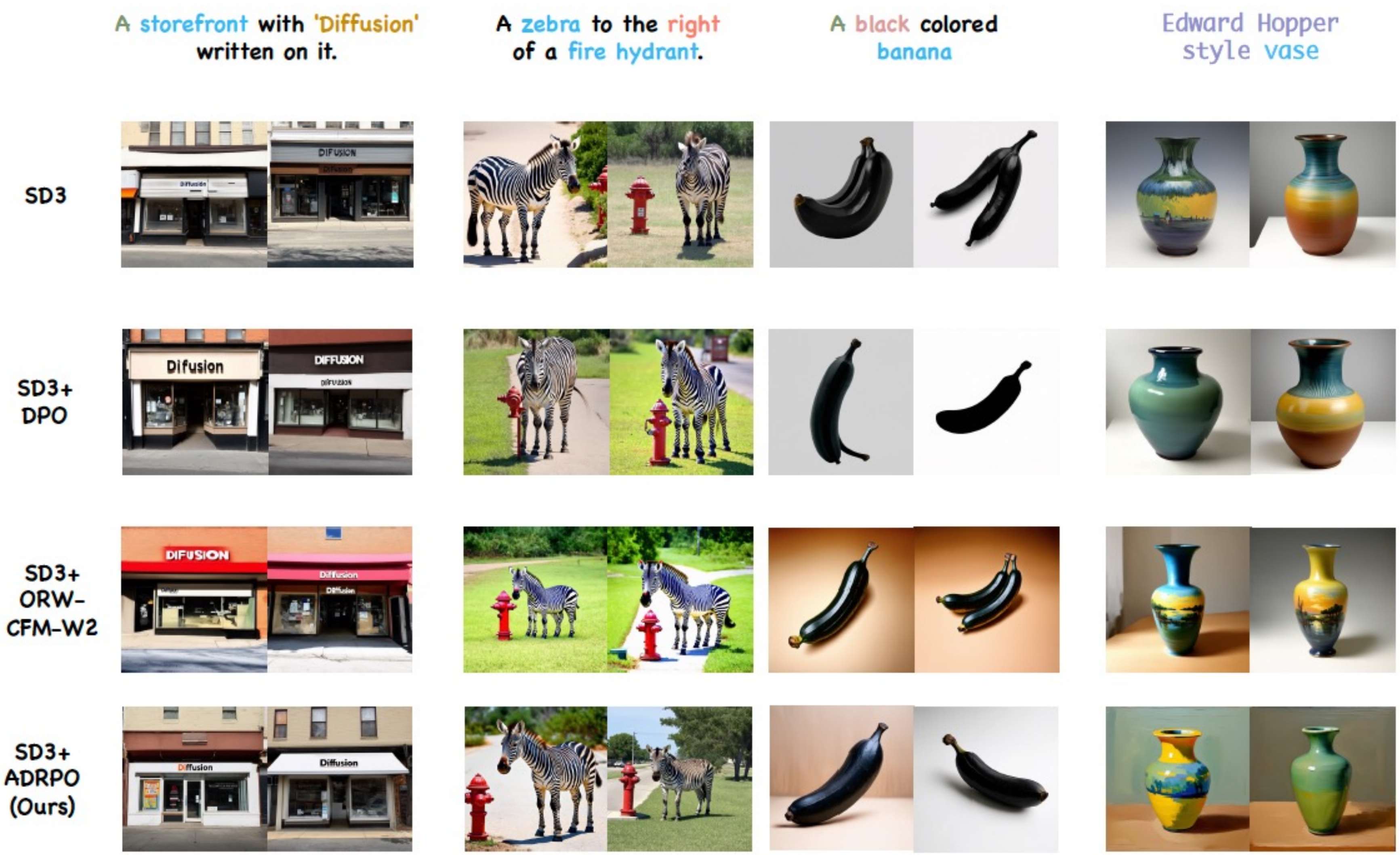


Figure 2: **Qualitative Comparison with Other RL Fine-tuning Methods.** Our ADRPO demonstrates superior performance in Artistic Style Rendering, Text Rendering, Attribute Binding, Coloring, Counting and Position. We use a similar DPO method as described in [8] to fine-tune SD3 models.

Visualizing Exploration-Exploitation Trade-off in Policy Optimization

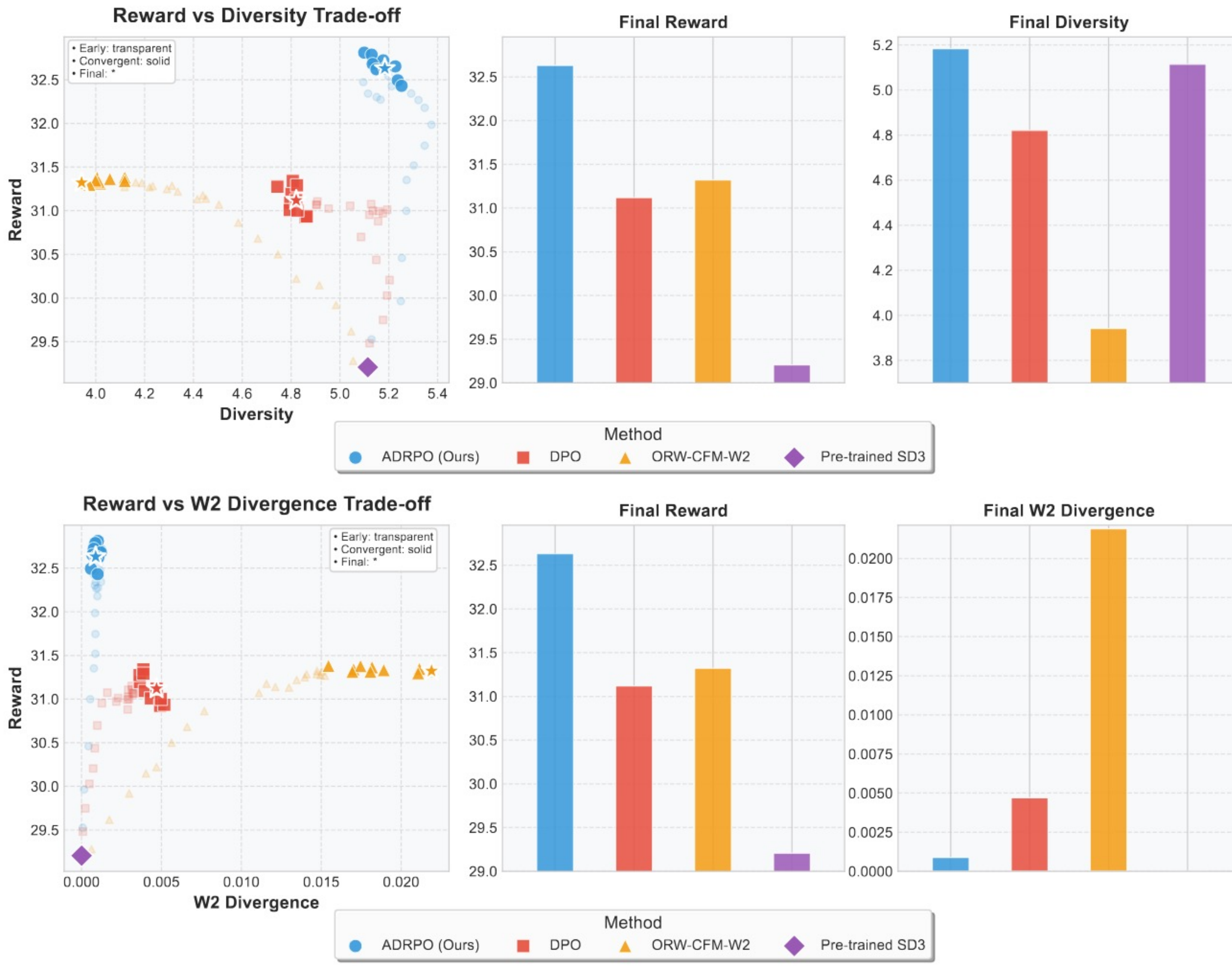


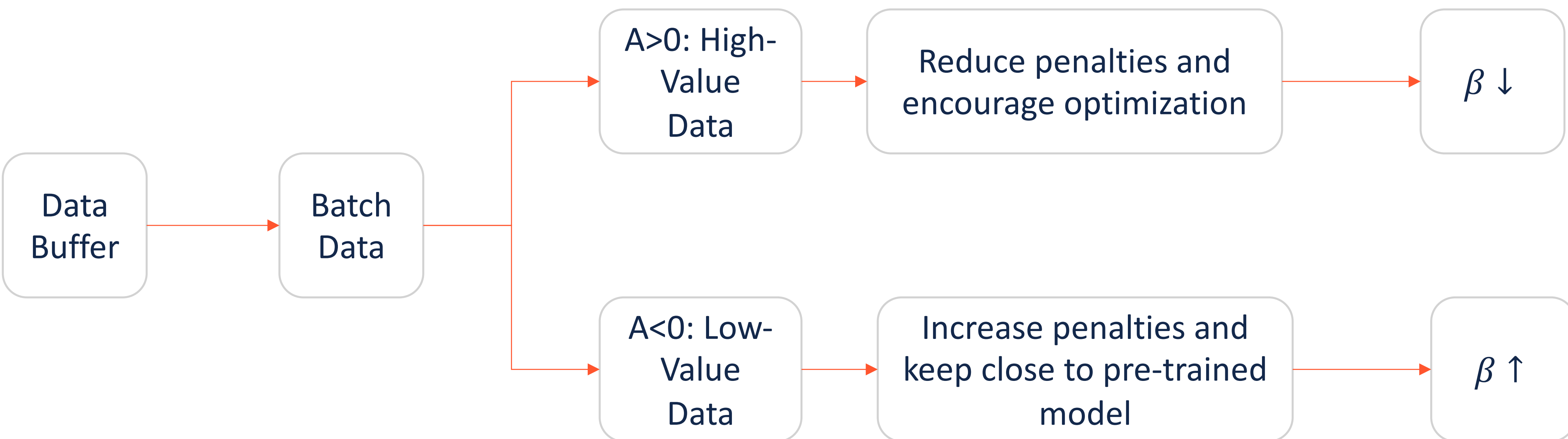
Figure 3: **Reward-Diversity/Divergence Trade-off.** Left: policy optimization trajectories (using a same seed) of different methods throughout training, with transparency indicating progression from early (transparent) to convergent (solid) to final (star) checkpoints. Each point is a learned policy from different iterations. Center and right: final reward and diversity/divergence across methods.

ADRPO for LLM & Reasoning Models

$$\mathcal{L}_{\text{ADRPO}}(\theta) = \mathcal{L}_{\text{RL}}(\theta) + (\beta_0 - A) \cdot \mathcal{L}_{\text{D}}(\theta)$$

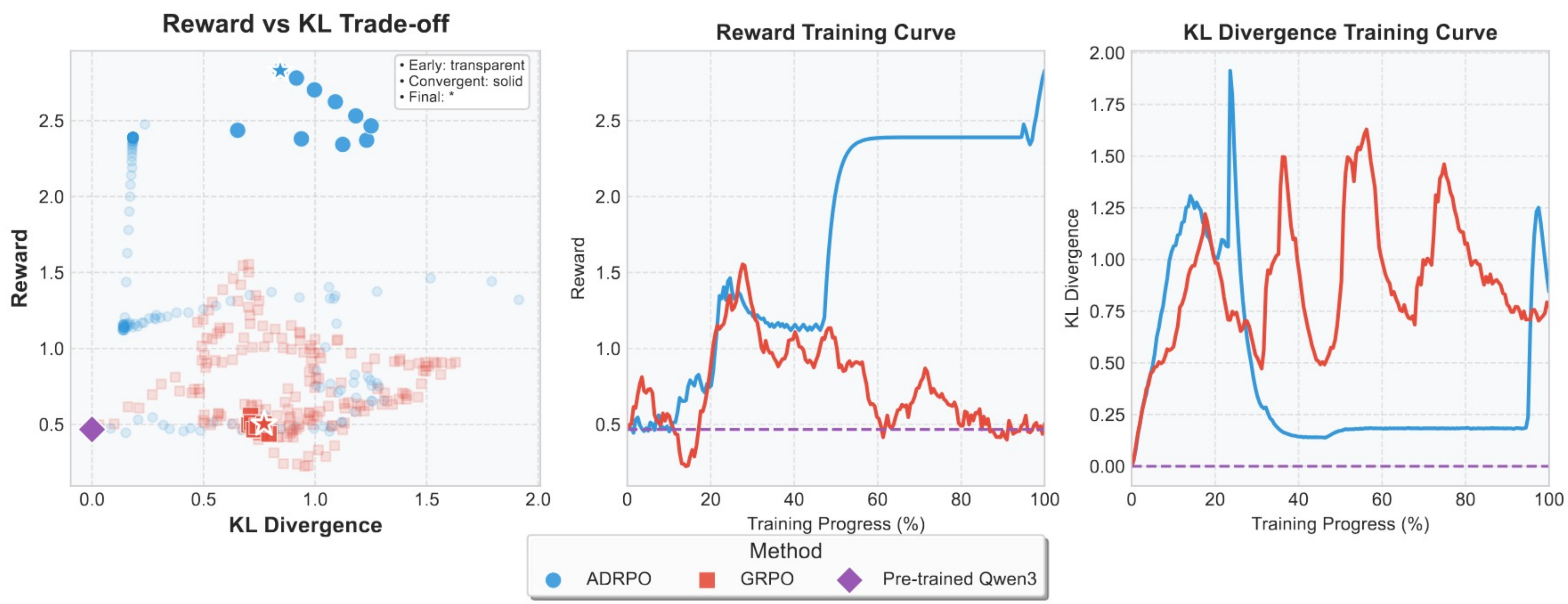
$$\mathcal{L}_{\text{ADRPO-GRPO}}(\theta) = \mathcal{L}_{\text{PG}}(\theta) + (\beta_0 - A_{\text{GRPO}}) \cdot D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right) \tag{7}$$

Here, $\mathcal{L}_{\text{PG}}(\theta)$ is the clipped policy gradient term [32] (i.e., $-\min(A{*}\text{ratio}, A{*}\text{clip}(\text{ratio}, 1-\epsilon, 1+\epsilon))$ and ratio $= \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}$), D_{KL} is the KL divergence, and β_0 is a baseline regularization. The term $(\beta_0 - A_{\text{GRPO}})$ acts as an adaptive coefficient, decreasing for good samples ($\boldsymbol{A}_{\text{GRPO}} > 0$) to promote exploitation and increasing for poor samples ($\boldsymbol{A}_{\text{GRPO}} < 0$) to enforce conservative exploration, allowing ADRPO-GRPO to achieve a better exploration-exploitation trade-off (See Fig. 4).



LLM Post-Training (Improving GRPO)

Base Model: Qwen2/3
Reward Model: RM-Gemma-2B



$$\mathcal{L}_{\text{ADRPO-GRPO}}(\theta) = \mathcal{L}_{\text{PG}}(\theta) + (\beta_0 - A_{\text{GRPO}}) \cdot D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right)$$

Multi-Modal Reasoning Model

Base Model: Qwen2.5-Omni
Reward Model: **Verifiable Rewards**

Table 2: Multi-modal audio reasoning results on MMAU benchmark [29].

Method	Sound (%)	Music (%)	Speech (%)	Total (%)
Qwen2.5-Omni (base)	72.37	64.37	69.07	68.6
GRPO	77.18	70.66	74.77	74.2
Gemini 2.5 Pro	75.08	68.26	71.47	71.6
GPT-4o Audio	64.56	56.29	66.67	62.5
ADRPO (Ours)	**81.98**	**70.06**	**75.98**	**76.0**

Table 3: Ablation on advantage clipping ranges.

Clipping Range	$A_{\min}$	$A_{\max}$	Sound (%)	Music (%)	Speech (%)	Total (%)
$1 \times \beta_0$ (recommended)	-0.04	0.04	81.98	70.06	75.98	**76.0**
$2 \times \beta_0$	-0.08	0.08	84.08	69.46	73.57	75.7
$0.5 \times \beta_0$	-0.02	0.02	82.58	71.26	74.47	76.1
GRPO (baseline)	-	-	77.18	70.66	74.77	74.2

Thanks