DISTRIBUTION MATCHING DISTILLATION MEETS REINFORCEMENT LEARNING

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Zhen Li, Bo Zhang, Mengmeng Wang, Steven Hoi, Peng Gao, Harry Yang

The Hong Kong University of Science and Technology, Alibaba Group, Shanghai AI Laboratory, Zhejiang University of Technology, The Chinese University of Hong Kong

THE CHALLENGE: SLOW DIFFUSION MODEL SAMPLING

>> Diffusion models produce unparalleled quality in visual generation but their iterative sampling process is slow and computationally expensive.

>> The goal is to accelerate sampling speed through model distillation into a generator that requires only a few steps.

>> Distillation approaches can be categorized into trajectory-based and distribution-based methods.

LIMITATION OF DISTRIBUTION MATCHING DISTILLATION (DMD)

- In DMD, the student model aims to match the distribution of the multi-step teacher model.
- This inherently means the student model's performance is capped by the teacher's capabilities.
- Previous solutions using GANs can introduce training instability and require external high-quality image data.

OUR SOLUTION: DMDR FRAMEWORK

>> We propose DMDR: Distribution Matching Distillation meets Reinforcement Learning.

>> DMDR combines DMD with RL concurrently, allowing the student model to surpass the teacher without external image data.

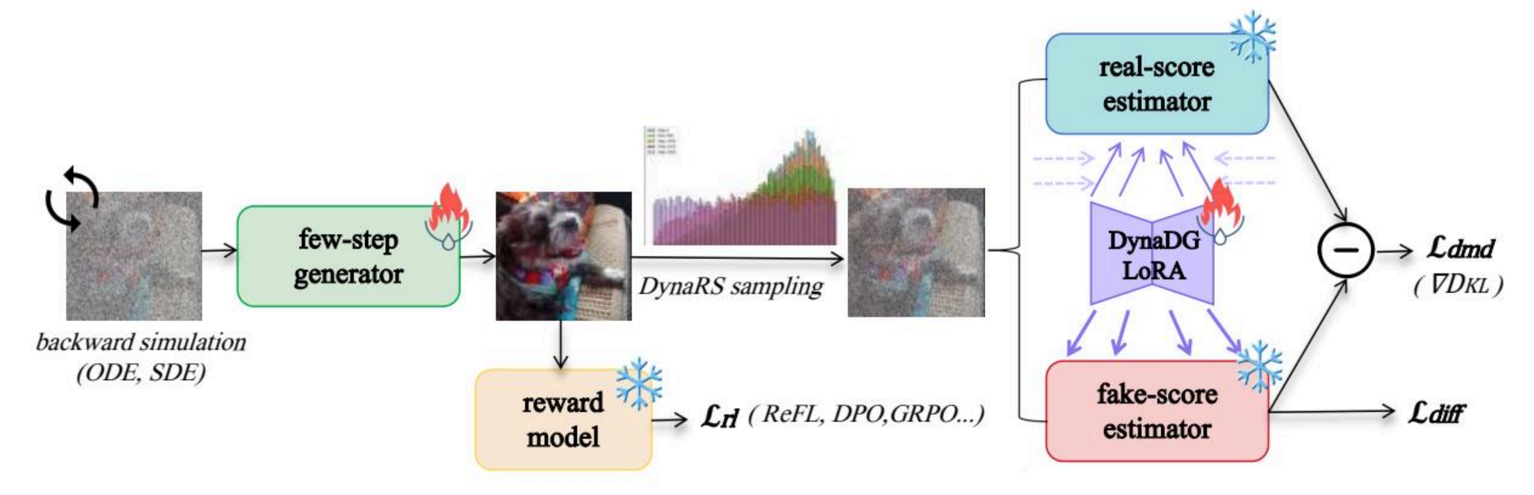
>> The combination is mutually beneficial: RL helps DMD cover high-reward modes, and DMD regularizes RL to prevent reward hacking.

DMDR FRAMEWORK OVERVIEW

>> **DMD Branch:** Optimizes the generator using an implicit distribution matching objective derived from the teacher model.

>> **RL Branch:** Concurrently incorporates reward feedback from a reward model to guide the generator towards preferred attributes.

>> **Dynamic Strategies:** Implements 'DynaDG' and 'DynaRS' to facilitate a more efficient and effective 'cold start' during the initial distillation phase.



PRELIMINARY: DISTRIBUTION MATCHING DISTILLATION (DMD)

- >> DMD compresses a multi-step teacher into a few-step student generator (G) by minimizing the KL divergence between their output distributions at various noise levels.
- >> The gradient for optimizing the generator is expressed as the difference between the score functions of the real (teacher) and fake (student) distributions.

$$\nabla_\theta \mathcal{L}_{dmd} = \mathbb{E}_t[\nabla_\theta \, \mathrm{KL}(p_{\text{fake},t} || p_{\text{real},t})] = -\mathbb{E}_t[\int (s_{\text{real}}(F_t)) - s_{\text{fake}}(F_t)) \frac{dG_\theta(z)}{d\theta} dz]$$

THE SYNERGY OF DMD AND RL

RL Unlocks DMD Performance

- RL provides supervision signals beyond the teacher, guiding the student to surpass it.
- Helps escape the teacher's undesirable modes and mitigates 'zero forcing' by ensuring high-reward modes are covered.

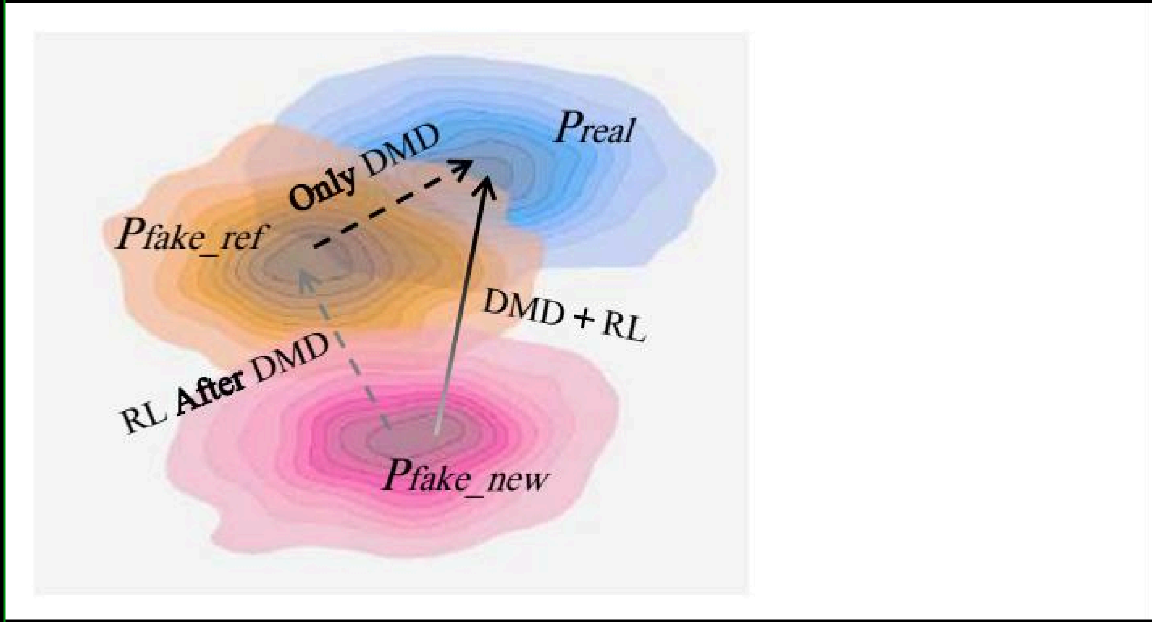
DMD Regularizes RL

- The DMD loss continuously pulls the student's distribution towards the robust teacher distribution, acting as an effective regularizer.
- This mitigates the risk of 'reward hacking' and error accumulation common in RL for generative models.

DMDR LOSS FUNCTION

- >> The final loss function is a straightforward combination of the DMD loss and a plug-and-play loss from the RL branch.
- >> This framework is compatible with various RL algorithms, such as ReFL, DPO, or GRPO.

$$\mathcal{L} = \mathcal{L}_{dmd} + \mathcal{L}_{rl}$$



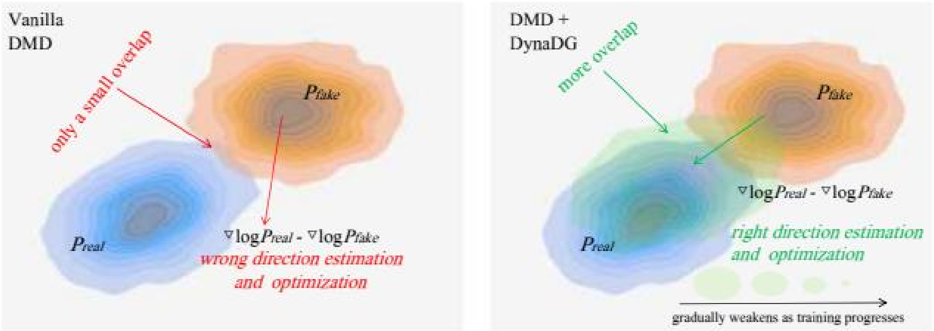
DYNAMIC COLD START STAGE FOR DMDR

01 **Dynamic Distribution Guidance (DynaDG)**

Injects a dynamically scaled LoRA into the real score estimator to create more overlap with the student's nascent distribution, ensuring reliable gradients from the start.

02 **Dynamic Renoise Sampling (DynaRS)**

Initially biases renoise sampling towards higher noise levels to help the generator learn global structures first, then gradually transitions to uniform sampling for finer details.



SYSTEM-LEVEL COMPARISON VS. SOTA METHODS

- >> DMDR-distilled models achieve state-of-the-art results across various base models (SDXL, SD3-Medium, SD3.5-Large).
- >> Our method consistently outperforms other few-step approaches in prompt coherence and aesthetic quality.
- >> DMDR is 'Image-Free', requiring no external real data for training.

Base Model	Method	Step	NFE	CLIP Score↑	Aesthetic Score↑	Pick Score↑	HP Score↑	Image-Free
SDXL-Base	Base (CFG=7.0)	25	50	34.7588	5.6480	22.1085	27.1477	–
SDXL-Base	DMDR (ours)	1	1	35.4835	6.0483	22.5424	31.1442	✓
SDXL-Base	DMDR (ours)	4	4	35.2940	5.9857	22.6268	32.8678	✓
SD3-Medium	Base (CFG=7.0)	25	50	34.9025	5.5942	22.1801	28.4021	–
SD3-Medium	DMDR (ours)	4	4	34.9542	5.8462	22.3578	31.8979	✓
SD3.5-Large	Base (CFG=3.5)	25	50	35.5509	5.7014	22.4856	28.8135	–
SD3.5-Large	DMDR (ours)	4	4	35.8647	6.0284	22.8859	32.4724	✓

QUALITATIVE RESULTS: SURPASSING THE TEACHER

- Images generated by our DMDR-distilled models demonstrate superior quality and prompt coherence compared to both their multi-step teachers and competing few-step distillation methods.
- The improvements are consistent across a variety of complex text prompts.

BENCHMARK EVALUATION: OUTPERFORMING TEACHERS

- >> Our 4-step distilled models consistently outperform their multi-step teachers on the DPG_Bench and GenEval benchmarks.
- >> This quantitatively validates that DMDR successfully unlocks the student model's potential beyond the teacher's limitations.

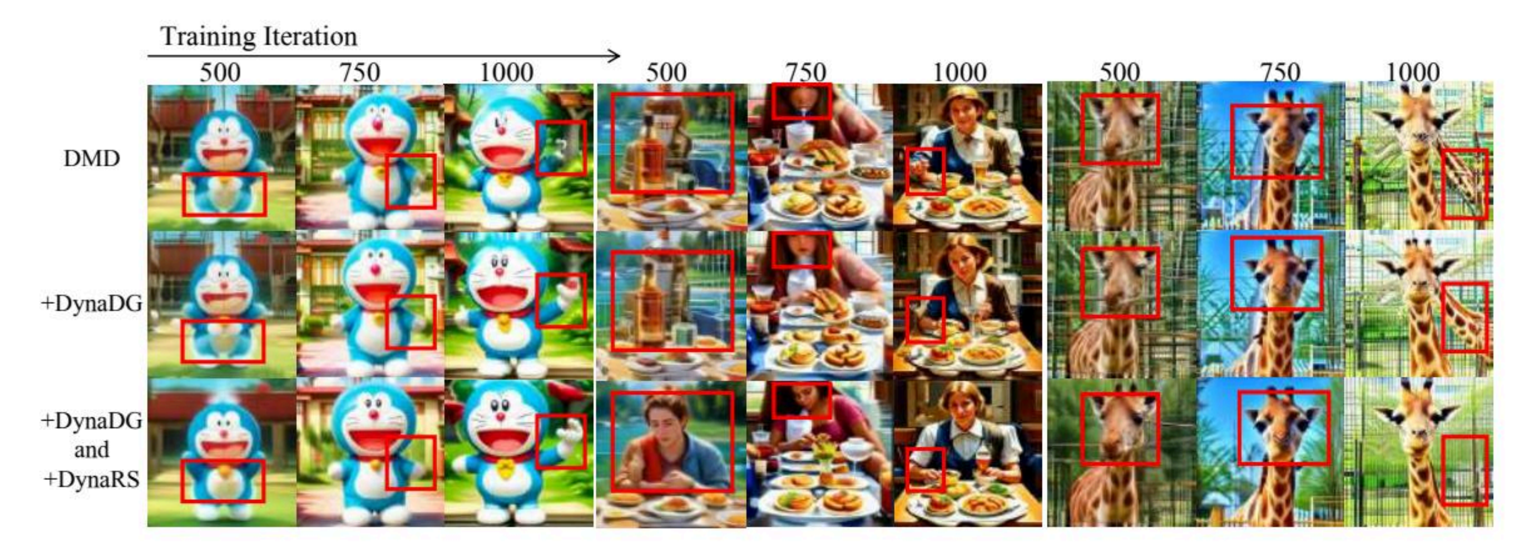
Model	Benchmark	Teacher Score	DMDR Score
SDXL-Base	DPG_Bench Overall	74.65	76.44
SD3-Medium	DPG_Bench Overall	84.08	84.96
SD3.5-Large	DPG_Bench Overall	84.12	85.30
SDXL-Base	GenEval Overall	0.55	0.56
SD3-Medium	GenEval Overall	0.62	0.64
SD3.5-Large	GenEval Overall	0.71	0.72

ABLATION: IMPORTANCE OF DYNAMIC COLD START

>> Both DynaDG and DynaRS significantly improve performance in the initial training phase compared to vanilla DMD.

>> DynaDG provides more reliable gradients by increasing distribution overlap, while DynaRS helps the model learn global structures first.

>> The dynamic, adaptive nature of these strategies is crucial for maximizing performance during the cold start phase.



ABLATION: SYNERGY OF DISTILLATION AND RL

- >> Training with only distillation is capped by the teacher's performance.
- >> Training with only RL can lead to reward hacking and inconsistent improvements.
- >> Combining Distillation + RL consistently achieves superior performance across all metrics, validating our core insight.

Method	CLIP Score	Aesthetic Score	Pick Score	HP Score
init	33.6432	5.6124	21.0489	29.1157
w/ only Distill.	33.6738	5.6248	21.6376	29.1389
w/ only RL (ReFL)	33.1897	5.8841	22.3008	31.2714
w/ Distill. + RL (ReFL)	34.6249	6.1813	22.7578	32.8979
w/ Distill. + RL (DPO)	33.9632	5.9710	21.9865	30.5994
w/ Distill. + RL (GRPO)	34.0055	5.8256	22.0120	30.6248

CONCLUSION

>> We proposed DMDR, a novel framework that synergistically combines Distribution Matching Distillation with Reinforcement Learning.

>> DMDR enables few-step models to surpass their multi-step teachers in an image-free manner.

>> Our dynamic cold start strategies (DynaDG, DynaRS) significantly accelerate and improve the initial distillation.

>> The approach is versatile, demonstrating strong performance across different model architectures and RL algorithms.