Steering Away from Memorization: Reachability-Constrained Reinforcement Learning for Text-to-Image Diffusion

Sathwik Karnik*[1], Juyeop Kim*[2], Sanmi Koyejo[1], Jong-Seok Lee[2], Somil Bansal[1]
[1]Stanford University [2]Yonsei University

NeurIPS 2026

Paper & code

Motivation

Memorization: text-to-image diffusion models can reproduce training images when prompted with their captions.

Memorization Happens Early

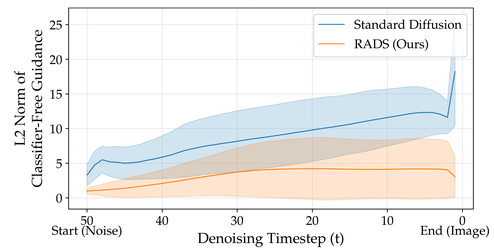


0 steps | 1 step | 2 steps | Memorized

Prompt: "The No Limits Business Woman Podcast"

Text guidance in only the first **2 denoising steps** already reproduces the memorized image, so RADS steers the **caption embedding**.

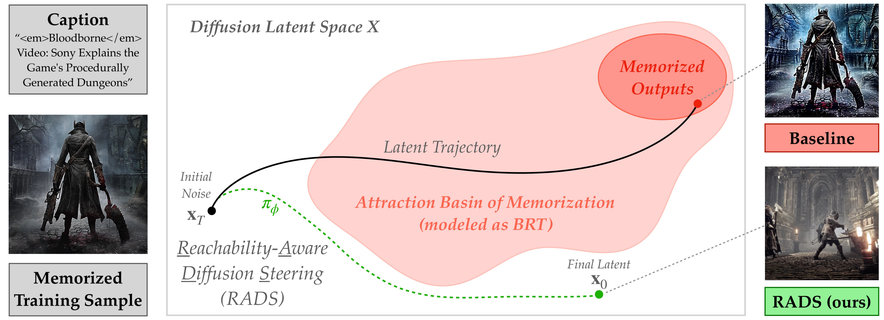
Guidance Norm Signals Memorization



The classifier-free guidance (CFG) norm spikes as standard diffusion falls into the memorization basin (blue); RADS (orange) steers away from this peak.

RADS: Reachability-Aware Diffusion Steering

First application of latent reachability to text-to-image diffusion denoising



Denoising as Control

The noisy latent is the state. The control u_t perturbs the caption embedding in a compact VAE latent space at each step.

$$\mathbf{x}_{T-(t+1)} = f_{\text{DM}}\Big(\mathbf{x}_{T-t},\ \underbrace{\texttt{Dec}(\texttt{Enc}(\mathbf{e}_c) + \mathbf{u}_t)}_{\text{Steered caption embedding } \mathbf{e}'_c},\ T-t\Big)$$

Memorization as Reachability

The policy-conditioned backward reachable tube (BRT) identifies states that reach a failure set under a given policy. We approximate memorization risk using a CFG-norm surrogate ℓ and a learned safety value that approximates the minimum remaining surrogate safety margin.

$$\ell(\mathbf{s}_t, \mathbf{u}_t) = -\tanh\left(\eta \cdot \left(\|\epsilon_\theta(\mathbf{x}_{T-t}, \mathbf{e}'_c) - \epsilon_\theta(\mathbf{x}_{T-t}, \emptyset)\|_2 - \beta\right)\right)$$

$$Q^{\text{safe}}_{\pi_\phi}(s_t, u_t) = \mathbb{E}_{\substack{s_{t+1}\sim P(\cdot|s_t,u_t)\\ u_{t+1}\sim\pi_\phi(\cdot|s_{t+1})}}\left[\min\left\{\ell(s_t,u_t), Q^{\text{safe}}_{\pi_\phi}(s_{t+1},u_{t+1})\right\}\right]$$

Training uses a discounted variant of this recursion.

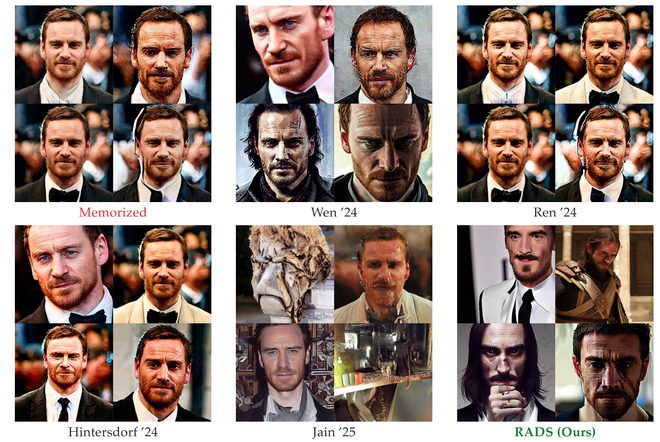
Reachability-Constrained Reinforcement Learning

Reward: CLIP image–prompt alignment. **Constraint:** $\mathbb{E}[Q^{\text{safe}}] \geq \delta$. Trained offline with Lagrangian soft actor-critic; at sampling time π_ψ steers the frozen diffusion model.

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t\sim\mathcal{D},\mathbf{u}_t\sim\pi_\phi}\Big[\underbrace{Q^{\text{task}}_\omega(\mathbf{s}_t,\mathbf{u}_t) + \alpha\mathcal{H}(\pi_\phi(\cdot|\mathbf{s}_t))}_{\text{task value + entropy (SAC)}} + \underbrace{\lambda\cdot(Q^{\text{safe}}_\psi(\mathbf{s}_t,\mathbf{u}_t) - \delta)}_{\text{reachability constraint}}\Big]$$

Diversity Across Seeds

Four random seeds per method. Prior methods return near-identical images or lose quality; **RADS generations vary across seeds.**



Prompt: "Michael Fassbender to Star In Assassin's Creed Movie"

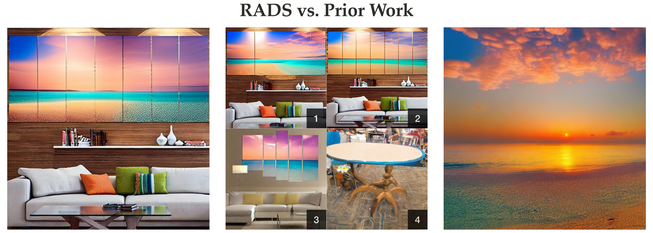
0.49 → 0.16
SSCD across seeds
no mitigation → RADS

0.041
SSCD across prompts
lowest; Jain '25: 0.272

Contributions

1. A **reachability-based formulation** of diffusion denoising as a controlled dynamical system, with noisy latents as states and caption-embedding perturbations as controls
2. **RADS**, a **reachability-constrained RL** algorithm that trains a lightweight steering policy, applied at sampling time without retraining the diffusion model
3. **Experiments** on SD v1.4 and Realistic Vision demonstrate a **stronger empirical Pareto frontier**, with additional evaluations on SD v2.1 and a controlled SDXL memorization setting.

RADS vs. Prior Work



Memorized | Prior work (1–4) | RADS (Ours)

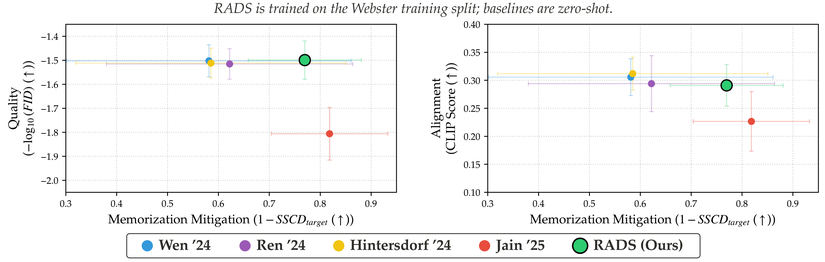
Prompt: "Designart Serene Maldives Seashore at Sunset Oversized Landscape Canvas Art – 4 Panels"
(1) Wen '24 (2) Ren '24 (3) Hintersdorf '24 (4) Jain '25

Memorization Mitigation vs. Quality and Alignment

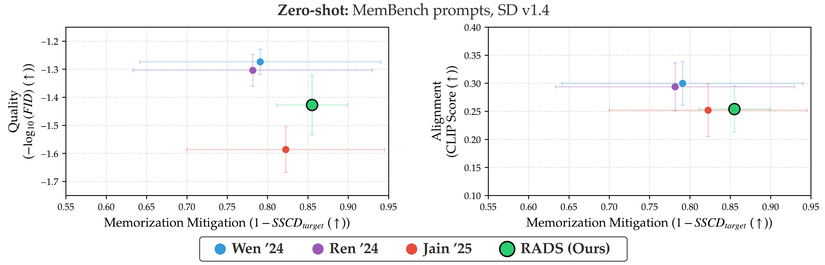
RADS achieves a stronger empirical Pareto frontier than inference-time baselines

In-distribution: Webster et al. prompts, SD v1.4

RADS is trained on the Webster training split; baselines are zero-shot.



Zero-shot: MemBench prompts, SD v1.4



x-axis: 1 – SSCD to the training image (↑, less copying). y-axes: $-\log_{10}$ FID (↑, quality) and CLIP score (↑, alignment).

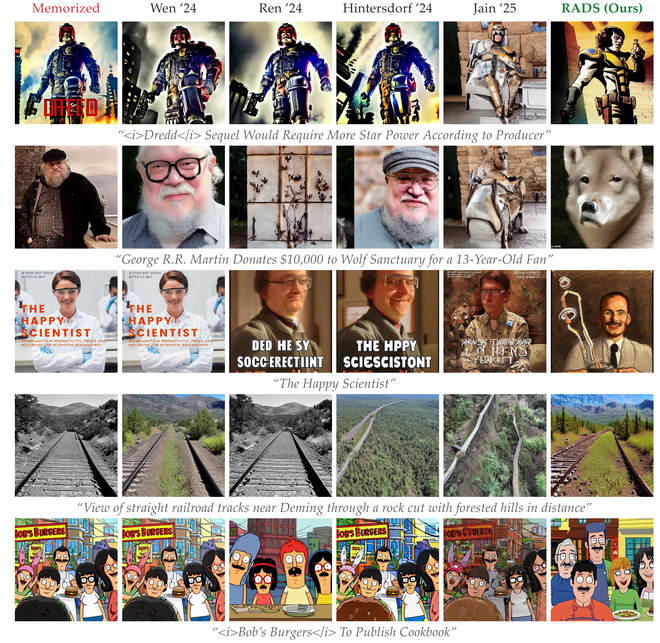
0.64 → 0.23
SSCD to training image
no mitigation → RADS (Webster)

31.6
Lowest mean FID on Webster
no mitigation: 42.1

0.145
zero-shot SSCD (MemBench)
lowest of all methods

0.23 → 0.50
SSCD without the reachability constraint
ablation: constraint removed

More Examples



"Dredd Sequel Would Require More Star Power According to Producer"

"George R.R. Martin Donates $10,000 to Wolf Sanctuary for a 13-Year-Old Fan"

"The Happy Scientist"

"View of straight railroad tracks near Deming through a rock cut with forested hills in distance"

"Bob's Burgers To Publish Cookbook"