DMA^2: Pixel-space Distribution Matching with Adversarial and Anchor Losses

Anonymous authors
Paper under double-blind review



Figure 1: Qualitative results of four-step text-to-image generation of DMA^2 (512×512).

Abstract

Distribution matching distillation (DMD) provides a general framework for few-step diffusion generation, but its modern text-to-image instantiations have been developed primarily around latent diffusion. It therefore overlooks key properties and design opportunities of native RGB. We revisit two DMD interfaces for pixel-space teachers. On the teacher-matching side, diagnostics show low-noise RGB matching is dominated by a local-texture cue, motivating a fixed high-noise matching band. On the real-data side, native clean-RGB outputs allow guidance from an external visual representation without traversing a decoder or sharing the heavy fake-score critic. DINO-Adv removes this critic from the adversarial gradient path and supplies local parametric patch guidance. For distribution-level guidance, we introduce AF-Loss, a parameter-free auxiliary semantic distribution-field objective designed for text-to-image DMD. It operates on detached rolling real and generated supports in the shared DINOv2 space while preserving prompt-conditioned teacher supervision. AF-Loss adds no learnable parameters or inference-time computation. Together these designs form DMA^2. Across DPG-Bench, GenEval, VQAScore, and COCO30K, the four-step DMA^2 student performs better than the 25-step teacher and evaluated few-step distillers.

1 Introduction

Recently, pixel diffusion models (Yu et al., 2026; Ma et al., 2026) advance text-to-image generation by modeling images end-to-end in RGB, avoiding the reconstruction ceiling and decoding artifacts of a VAE. However, recent pixel diffusion still requires many sequential denoiser evaluations, making

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

generation slow and costly. Few-step distillation, represented by distribution matching distillation (DMD) (Yin et al., 2024b;a), is the standard remedy: it uses a learned fake score to optimize a distribution-matching objective toward teacher outputs, instead of regressing the sampling trajectory step by step, which compounds error as steps are removed (Luo et al., 2023; Liu et al., 2024). DMD is general, but its modern text-to-image recipe has been developed primarily around latent diffusion. Direct RGB generation changes both supervision interfaces: fine texture is explicit during teacher matching, while clean outputs expose pretrained visual representations for real-data guidance.

We analyze how the general DMD recipe interacts with native RGB. For teacher matching, the standard recipe samples matching levels broadly, without accounting for how information is represented in uncompressed RGB. Low-noise inputs already determine coarse image content, so teacher–critic differences can be dominated by fine texture rather than the distributional factors DMD should transfer. Our diagnostics indicate that higher-noise matching is more consistent with structural supervision. For real-data guidance, DMD2 adds a GAN loss against real images (Yin et al., 2024a). Its conventional Feature-GAN design attaches the adversarial head to the 560M fake-score network. For a native-RGB student, this coupling is avoidable: clean outputs allow guidance through a pretrained visual representation without traversing a decoder or sharing the fake-score critic.

Motivated by this evidence, we design DMA^2, which specializes the teacher-matching and real-data-guidance interfaces of DMD for native-RGB generation. On the teacher-matching side, frequency controls, gradient visualizations, and a targeted matching-band intervention provide converging evidence for a near-clean local-texture bias in native-RGB matching (Section 3.1). These diagnostics motivate a fixed high-noise matching band that improves the model's performance without additional trainable modules or online scheduling. On the real-data side, we exploit direct access to clean RGB through a dual-granularity, critic-decoupled guidance design. DINO-Adv removes the fake-score critic from the adversarial gradient path and uses an 86M-parameter frozen DINOv2 (Oquab et al., 2023) encoder with only 1.58M learnable parameters in its heads, reducing same-pipeline step time from 1.82 to 0.72 seconds (2.5$\times$ faster). This local adversary, however, does not define how each generated sample should move relative to the empirical real and generated distributions in semantic feature space. We therefore introduce Anchor-Field Loss (AF-Loss), a parameter-free auxiliary semantic distribution-field objective designed for text-to-image DMD. AF-Loss acts alongside the prompt-conditioned teacher objective, maintains detached rolling supports across minibatches, and shares the frozen DINOv2 encoder with DINO-Adv. It complements local adversarial guidance without adding learnable parameters or inference-time computation. With both sides addressed, the four-step student performs better than the 25-step DeCo (Ma et al., 2026) teacher and existing few-step distillers on all reported metrics; the same recipe transfers to a second backbone (PixelGen-XXL; Appendix B). The main contributions of this work are:

- We propose DMA^2, a few-step native-RGB DMD framework that specializes both teacher matching and real-data guidance for pixel-space generation. Across DPG-Bench, GenEval, VQAScore, and COCO30K, its four-step student ranks first among the evaluated few-step distillers and performs better than the 25-step teacher on all seven reported metrics; the same recipe transfers to a second pixel-space backbone.
- We specialize both sides of the general DMD recipe for native RGB. On the teacher-matching side, converging diagnostics provide evidence for a near-clean local-texture shortcut and motivate **DM-Band**, which improves both alignment and image quality. On the real-data side, direct RGB access enables **DINO-Adv** to decouple local adversarial guidance from the 560M learnable fake-score critic, using an 86M-parameter frozen DINOv2 backbone and only 1.58M learnable head parameters, while reducing same-pipeline step time from 1.82 to 0.72 seconds.
- We introduce **AF-Loss**, a parameter-free auxiliary semantic distribution-field objective designed for text-to-image DMD. It supplies distribution-level guidance from detached rolling supports while preserving prompt-conditioned teacher supervision, shares the frozen DINOv2 encoder with DINO-Adv, and adds no learnable parameters or inference-time computation.

2 Related Work

Pixel-space diffusion. A line of work generates in RGB without a frozen VAE, from cascaded pixel models (Saharia et al., 2022) to end-to-end RGB transformers competitive with latent models (Jabri et al., 2023; Hoogeboom et al., 2023; 2025; Crowson et al., 2024). PixelDiT scales dual-level pixel

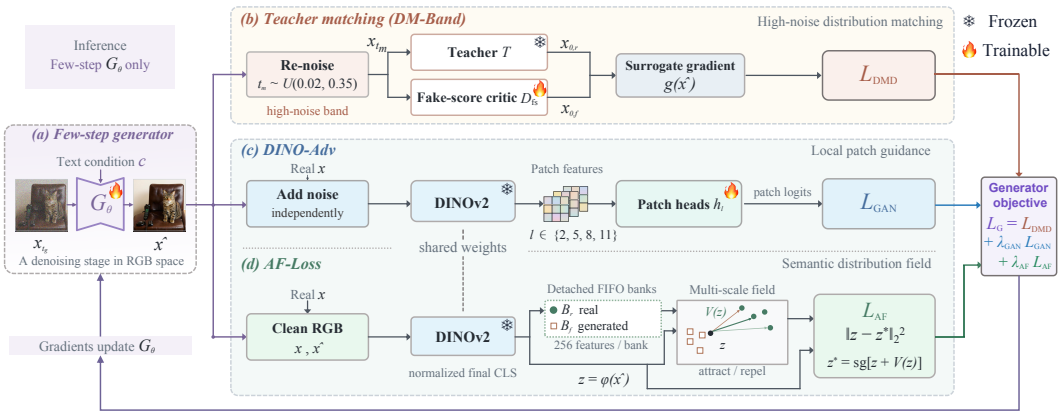

Figure 2: Overview of the proposed DMA^2 framework. DM-Band restricts teacher matching to a diagnostically selected high-noise range. For real-data guidance, DINO-Adv and AF-Loss share a frozen DINOv2 encoder and provide local patch guidance and a parameter-free semantic distribution-field target, respectively. Only G_θ, the fake-score critic, and the adversarial heads are trained; all auxiliary components are removed at inference.

transformers to megapixel generation (Yu et al., 2026), whereas PiD uses conditional pixel diffusion for latent decoding and upsampling (Lu et al., 2026). The paradigm also extends to controllable generation, restoration, depth and geometry, and 3D (Lin et al., 2026; Sun et al., 2026; Xu et al., 2025; Liu et al., 2026a; Yuan et al., 2026; Xu et al., 2026; Gao et al., 2026). Our teacher DeCo (Ma et al., 2026) is a frequency-decoupled member of this family. This literature primarily studies the training of native-RGB text-to-image models; few-step distillation tailored to their representation characteristics remains comparatively underexplored.

DMD-based diffusion distillation. Distribution matching distillation trains a few-step student to match a teacher's output distribution through a learned fake score (Yin et al., 2024b), and DMD2 adds a real-image GAN term with two-time-scale updates for stability (Yin et al., 2024a). Recent baselines Flash, TDM, Decoupled DMD, and DMDR span general acceleration, trajectory matching, CFG–DMD decoupling, and RL post-training, respectively (Chadebec et al., 2025; Luo et al., 2025; Liu et al., 2026b; Jiang et al., 2026). Related score and variational objectives distill in one step (Zhou et al., 2024; Nguyen & Tran, 2024), while consistency and rectified-flow methods regress the sampling trajectory (Luo et al., 2023; Liu et al., 2024). Modern text-to-image distillation has centered on latent teachers; we specialize DMD's matching noise and real-data guidance for native RGB.

Feature-space objectives and field losses. Frozen encoders support projected/DINOv2 discriminators (Sauer et al., 2021; Oquab et al., 2023), representation alignment (Yu et al., 2025), and distributional objectives based on feature statistics or attract–repel fields (Yang et al., 2026; Feng et al., 2026; Deng et al., 2026). Drifting Models use a kernelized field as the standalone objective for class-conditional generation (Deng et al., 2026). AF-Loss is designed as an auxiliary objective for text-to-image DMD: prompt-conditioned teacher supervision remains active, detached rolling supports extend beyond the current minibatch, and the DINOv2 encoder is shared with DINO-Adv.

3 Method

Given a Gaussian noise map ϵ and a conditioning input c, a single few-step generator G_θ produces a clean RGB image $\hat{x}$ in K pixel-space denoising steps (as few as one), with no latent tokenizer or decoder at any point (Figure 2). This native RGB output exposes both the matching behavior and the external-guidance opportunity analyzed in Section 1. On the **teacher-matching side**, converging diagnostics motivate a DMD branch specialized to native RGB through a fixed high-noise **distribution-matching band** (**DM-Band**, Section 3.1). On the **real-data side**, direct access to $\hat{x}$ lets one frozen DINOv2 backbone support dual-granularity, critic-decoupled guidance without traversing a VAE decoder: **DINO-Adv** uses multi-layer spatial patch features (Section 3.2), whereas **AF-Loss**

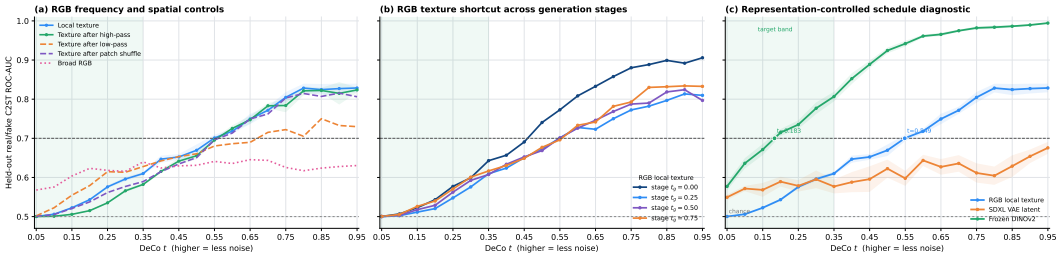

Figure 3: **Separability diagnostic (C2ST).** Each panel plots a held-out classifier's real/fake separability (y-axis: ROC-AUC over 1,536 matched real/generated pairs, 0.5 is chance) against the DeCo matching level t (x-axis; larger t is less noise). The panels vary the setup: *(a)* the RGB input transform (raw texture, high-pass, low-pass, 32×32 patch shuffle); *(b)* the generation stage $t_g \in \{0, .25, .5, .75\}$; *(c)* the input representation (RGB, the SDXL latent, frozen DINOv2). RGB separability changes sharply with t and is high only near clean, whereas the SDXL latent stays flat and DINOv2 stays separable into the high-noise band.

uses the normalized final-layer CLS feature for a parameter-free semantic distribution-field objective (Section 3.3). Only G_θ, the fake-score critic, and the adversarial heads are trained; AF-Loss adds no learnable parameters, and every auxiliary module is discarded at inference.

3.1 Diagnosing native-RGB teacher matching

We distill G_θ with the DMD objective, using the short schedules $\mathcal{S}_4 = \{0, .25, .5, .75, 1\}$ and $\mathcal{S}_1 = \{0, 1\}$ ($t = 0$ noise, $t = 1$ clean, DeCo convention). At a matching time t_m, we noise the student output once to $x_{t_m} = \alpha_{t_m}\hat{x} + \sigma_{t_m}\epsilon$ ($\epsilon \sim \mathcal{N}(0, I)$) and pass this *same* x_{t_m} to both the real score (frozen teacher) and the fake score (a critic initialized from the teacher and trained on noised student samples), which denoise it to clean estimates $x_{0,r}$ and $x_{0,f}$, giving the normalized surrogate gradient

$$g(\hat{x}) = \frac{(\hat{x} - x_{0,r}) - (\hat{x} - x_{0,f})}{\text{mean}\,|\hat{x} - x_{0,r}| + \epsilon}, \quad \mathcal{L}_{\text{DMD}} = \tfrac{1}{2}\|\hat{x} - \text{sg}(\hat{x} - g(\hat{x}))\|_2^2, \tag{1}$$

with sg the stop-gradient. The only choice we revisit is which matching levels $t_m \in [.02, .98]$ (lower is noisier) supervise this gradient.

Analysis: what separates real from generated RGB images. For our native-RGB teacher, the choice of matching level matters. We therefore probe which cues distinguish real from generated RGB images at each level (Figure 3). Using 1,536 matched real/generated pairs, we noise both members of each pair to the same matching level t and perform a classifier two-sample test (C2ST (Lopez-Paz & Oquab, 2017)). We fit a linear real/fake classifier on a diagnostic training split and report ROC-AUC on held-out pairs as the separability score (0.5 is chance; higher t is cleaner). *(a)* Separability varies sharply with the noise level: high near clean and near chance at high noise, so the two sets are easiest to classify at low noise. The cue survives a high-pass and a 32×32 patch shuffle but weakens under a low-pass; together, these controls are most consistent with a high-frequency, spatially local texture cue rather than broad structure. *(b)* Across the four generation stages $t_g \in \{0, .25, .5, .75\}$ the curves nearly coincide, suggesting that the pattern is approximately stage-stable and that one fixed band can serve all stages. *(c)* We compare the classifier's input representation. The latent representation of SDXL remains nearly flat across noise levels, whereas RGB and DINOv2 features vary substantially with noise. RGB texture separability is concentrated near clean images; in contrast, semantic information represented by DINOv2 (Oquab et al., 2023) remains discriminative between real and generated samples into the high-noise range. We use C2ST to localize the noise regions in which real and generated representations differ. This C2ST localizes *where* a real-versus-generated cue exists; the update-level and causal evidence come from the teacher–critic difference Δ (Figure 4) and the band intervention with its latent negative control. Frequency and spatial controls further indicate that the near-clean separable cue is predominantly local and high-frequency; the gradient visualization below illustrates its spatial form, while the matching-band intervention and latent-space negative control further test the resulting native-RGB matching hypothesis.

The high-noise band (DM-Band). To preserve DMD's core goal of constraining structure and semantics, we keep the near-clean texture cue out of the gradient by truncating the matching range and supervising DMD only at high noise. In Figure 3c, for $t \lesssim .35$, the semantics-oriented DINOv2 features distinguish real from generated samples, whereas the texture- and detail-sensitive RGB representation remains weakly separable. Figure 4 visualizes the Eq. 1 terms (the noised input, the teacher and critic residuals, and their difference Δ), each panel scaled to its own maximum: this visualization compares spatial content, not magnitude. The gradient Δ exhibits broad coherent patterns at $t_m = .35$ and sparse fine-scale patterns by $t_m = .55$, placing .35 near the observed transition. Combining the two diagnostics, we draw matching levels from $t_m \sim U(.02, .35)$, applied to the DMD branch only. The band ablation below then tests this targeted intervention, while the latent-space negative control tests whether the intervention is specific to the observed native-RGB matching bias. It is a single hand-set hyperparameter with no stage-dependent schedule, classifier at training time, or online estimation.

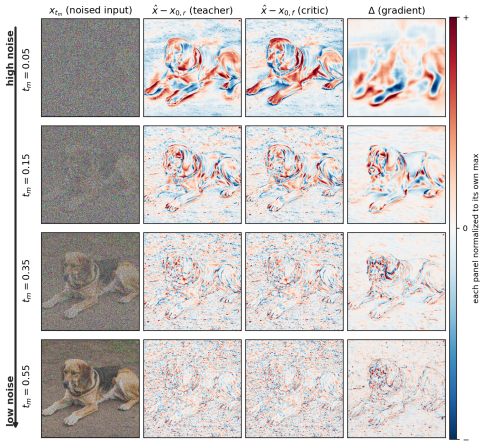

Figure 4: Eq. 1 terms across t_m; Δ exhibits broader coherent patterns at high noise and sparse fine-scale patterns near clean. Each panel is independently normalized for spatial visualization; colors are not comparable in magnitude across cells.

Latent-space negative control. Because the prescription is motivated by the observed near-clean RGB texture bias, it should not be universally beneficial when that representation-level bias is absent. On latent SDXL-DMD2 the same band lowers DPG ($67.64 \rightarrow 63.23$), opposite to native RGB. This is consistent with VAE compression suppressing low-level cues, making DM-Band a native-RGB remedy rather than a universal DMD schedule.

DPG-Bench↑	$U(.02, .98)$	$U(.02, .35)$
Pixel (DeCo)	82.55	**82.96**
Latent (SDXL)	**67.64**	63.23

3.2 Critic-decoupled local adversary (DINO-Adv)

Matching the teacher alone is limited by its imperfect score on real data (Yin et al., 2024a), so we add guidance that learns from real images directly. We exploit a design freedom exposed by native RGB outputs: because the student emits clean RGB natively, real-data guidance can be evaluated directly by an external pretrained model without traversing a VAE decoder or sharing the fake-score critic. DINO-Adv decouples the adversarial gradient path from the co-adapting fake-score critic with 560M learnable parameters and instead uses an 86M-parameter frozen DINOv2 ViT-B/14 backbone with only 1.58M learnable head parameters, reducing per-step training time from 1.82 to 0.72 seconds (2.5× faster) while supplying a stable local realism signal. Specifically, we independently sample a noise level for each real image x and generated image $\hat{x}$, and feed the resulting noisy RGB images into a frozen DINOv2 ViT-B/14; layers $\{2, 5, 8, 11\}$ provide spatial patch features f_l to learnable heads

$$h_l(f_l) = \mathrm{Conv}_{1\times1}^{768\to512} \rightarrow \mathrm{GN}_{32} \rightarrow \mathrm{LeakyReLU} \rightarrow \mathrm{Conv}_{1\times1}^{512\to1}, \tag{2}$$

which output patch logits, trained with the non-saturating softplus objective. The same frozen backbone later supplies the normalized final-layer CLS feature for AF-Loss (Section 3.3), so one representation supports both local parametric patch guidance and a parameter-free semantic distribution-field objective. The default weight is $\lambda_{\mathrm{GAN}} = .01$.

Effect of critic decoupling. We compare the two discriminators head-to-head by porting DMD2's feature GAN unchanged (a head on block-7 of the 16-block critic, the "Feature GAN" baseline) and swapping only that critic-coupled head for the frozen-DINOv2 adversary, keeping the same DMD and schedule. The frozen-DINOv2 adversary comprises an 86M-parameter frozen ViT-B/14 backbone and only 1.58M learnable head parameters, whereas the DMD2 Feature GAN uses 560M learnable

	DPG-Bench↑		GenEval↑		Train cost↓	
Adversary	1	4	1	4	s/step	#learn. (adv. path)
DMD2 Feat. GAN	78.35	82.12	0.6960	0.7709	1.82	560M
Frozen-DINOv2	**81.26**	**82.55**	**0.7723**	**0.8155**	**0.72**	**1.58M**

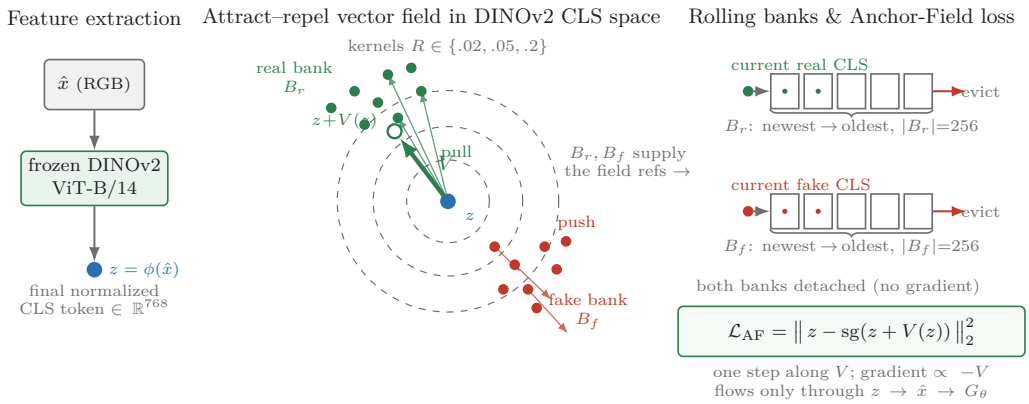

Figure 5: **Anchor-Field Loss (AF-Loss).** The generated image's normalized final-layer DINOv2 CLS feature $z = \phi(\hat{x})$ is attracted toward a rolling bank of real features B_r and repelled from a bank of generated features B_f, giving a displacement $V(z)$. AF-Loss converts this displacement into a stopped target whose gradient flows only through $z \to \hat{x} \to G_\theta$; the field itself has no learnable parameters. The banks are detached FIFO queues capped at 256 features.

adversary parameters. The swap improves both DPG-Bench and GenEval at one and four steps, reaching 0.8155 four-step GenEval, with the gap widening as the trajectory shortens, and it trains about 2.5× faster (0.72 vs. 1.82 s/step).

3.3 Parameter-free semantic distribution anchor-field loss (AF-Loss)

DINO-Adv supplies local parametric patch guidance, but its patch decisions do not specify a distribution-level direction in semantic feature space. A complementary objective for DMD should provide such a direction without replacing the conditional teacher objective, learning another estimator, or modifying inference. We therefore introduce AF-Loss, a *parameter-free semantic distribution anchor-field loss for DMD*. It turns detached empirical real and generated feature supports into a sample-dependent, multi-scale displacement and injects that displacement as a stopped target for the generator. This is distinct from direct feature regression, whose target does not depend on the geometry of both empirical supports, and from standalone field-based generation (Deng et al., 2026), where the field itself defines the generative process.

Concretely, AF-Loss operates on normalized final-layer DINOv2 CLS features extracted directly from generated RGB. Multi-scale kernels attract each generated feature toward real-feature neighborhoods and repel it from generated-feature neighborhoods. DMD continues to provide teacher and prompt guidance, while AF-Loss reuses the same frozen encoder as DINO-Adv and adds no learnable parameters or inference-time computation. For a stable empirical support, current real and student features are prepended to two detached FIFO banks B_r and B_f, each capped at 256 features; the sole trainable path is $z \to \hat{x} \to G_\theta$ (Figure 5).

We measure similarity in the DINOv2 feature space with the Euclidean distance $d(z,u) = \|z-u\|$, normalized by the batch-mean distance $\bar{d}$ for scale invariance. At bandwidth R the affinity is a softmax kernel $K_R(z,u) \propto \exp\big(-d(z,u)/(\bar{d}\,R)\big)$, symmetrized over the batch and self-masked so a sample never interacts with its own copy in B_f. Attracting z toward its real neighbors in B_r and repelling it from its generated neighbors in B_f, kernel-weighted and summed over the bandwidths $R \in \mathcal{R}$, gives an illustrative simplified field $\widetilde{V}(z)$ at $z = \phi(\hat{x})$

$$\widetilde{V}(z) = \sum_{R\in\mathcal{R}} \left[\frac{\sum_{r\in B_r} K_R(z,r)(r-z)}{\sum_{r\in B_r} K_R(z,r) + \epsilon} - \frac{\sum_{f\in B_f} K_R(z,f)(f-z)}{\sum_{f\in B_f} K_R(z,f) + \epsilon} \right]. \tag{3}$$

$\widetilde{V}$ is an illustrative single-softmax simplification; the actual field V used below is the symmetrized (doubly-normalized), cross-coupled, per-scale force-normalized field summed over $\mathcal{R}$ and defined exactly in Appendix D (Algorithm 1). Because V is treated as a fixed target displacement (stop-

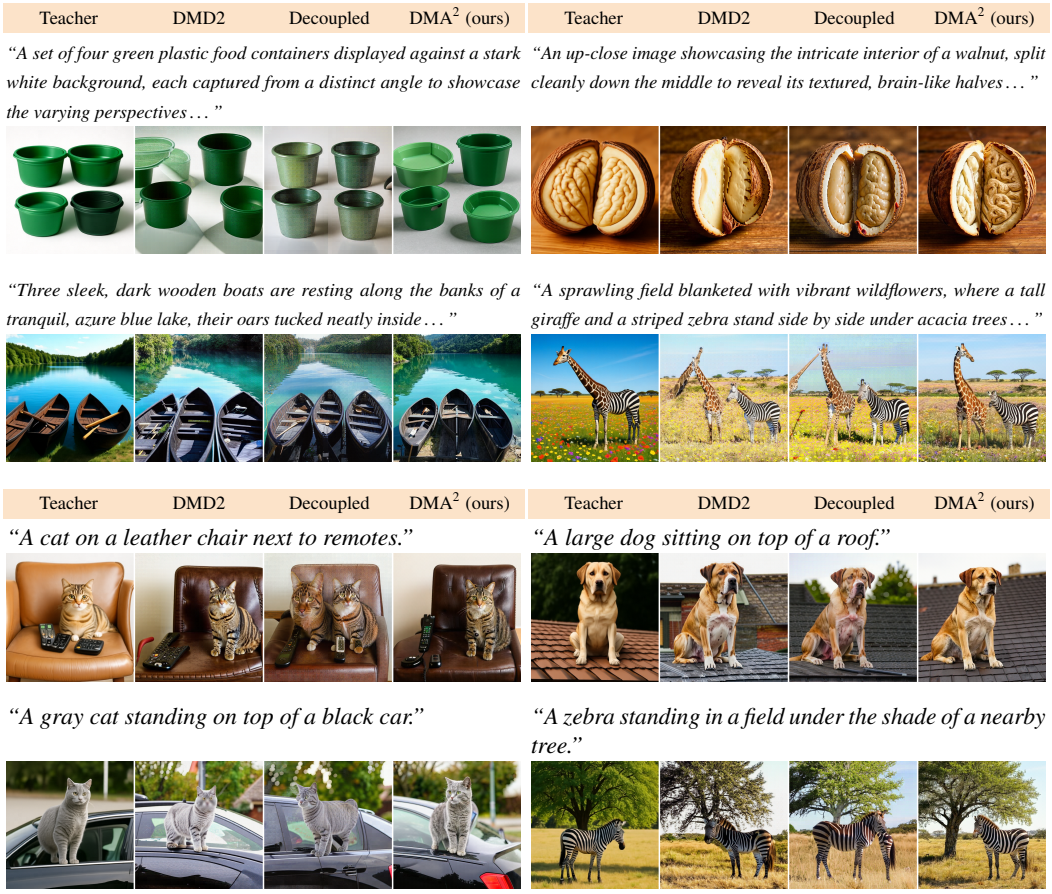

Figure 6: Qualitative comparison at four steps. Top block: long, densely specified DPG-Bench prompts (abbreviated with "…"); bottom block: short COCO captions. DMA^2 preserves object count and composition and recovers fine structure, across both long and short text.

gradient), the loss simply moves z one step along the field,

$$\mathcal{L}_{\mathrm{AF}} = \|\phi(\hat{x}) - \mathrm{sg}(\phi(\hat{x}) + V(\phi(\hat{x})))\|_2^2, \tag{4}$$

so its gradient with respect to z points along $-V(z)$, up to the positive scalar induced by the squared-error reduction, and the signal flows only through $z \to \hat{x} \to G_\theta$. The reference clouds are *detached* first-in-first-out queues: at each step we prepend the current real and student features and truncate to $N_b = 256$,

$$B_r \leftarrow [\,\phi(x)\,;\,B_r\,]_{1:N_b}, \qquad B_f \leftarrow [\,\phi(\hat{x})\,;\,B_f\,]_{1:N_b}, \tag{5}$$

With a global batch of 32, each step admits the newest 32 real and 32 generated tokens and evicts the oldest 32, so every 256-token cloud spans the current batch and the previous seven; the clouds receive no gradient. The balanced anchor uses $|B_r| = |B_f| = N_b = 256$, $\mathcal{R} = \{.02, .05, .2\}$, and $\lambda_{\mathrm{AF}} = .05$; our maximum-DPG variant changes only the Anchor-Field weight to .10 and the radii to $\{.01, .05, .5\}$. No other auxiliary losses are used. Kernel batch symmetrization, the scope of $\bar{d}$, and the loss reduction are detailed in Appendix D. The total generator objective is

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{DMD}} + \lambda_{\mathrm{GAN}}\mathcal{L}_{\mathrm{GAN}} + \lambda_{\mathrm{AF}}\mathcal{L}_{\mathrm{AF}}. \tag{6}$$

4 Experiments

4.1 Experimental setting

Implementation details. We distill a 1.1B-parameter, 25-step DeCo teacher (Ma et al., 2026) at 512×512 resolution. The student is initialized from the teacher and trained for 20k optimizer steps

Table 1: Main comparison. All few-step methods are re-implemented on the same teacher; bold/underline mark the best/second-best students at each step count.

Model	NFE	GenEval↑	DPG↑	VQA↑	IS↑	CLIP↑	Rec.↑	NIQE↓
DeCo (Ma et al., 2026) (official)	25	0.8216	81.580	0.7014	35.58	0.3200	0.3110	4.064
Flash (Chadebec et al., 2025)	4	0.7814	76.703	0.6900	29.94	0.3178	0.1486	5.688
DMD2 (Yin et al., 2024a)	4	0.7709	82.115	0.7058	<u>36.76</u>	0.3170	0.4349	4.168
TDM (Luo et al., 2025)	4	0.7924	82.367	0.7071	36.48	0.3165	<u>0.4447</u>	3.871
Decoupled (Liu et al., 2026b)	4	<u>0.8100</u>	83.095	<u>0.7111</u>	36.51	0.3195	0.3516	3.953
DMDR (Jiang et al., 2026)	4	0.8063	<u>83.160</u>	0.7023	34.22	<u>0.3198</u>	0.4414	<u>3.710</u>
DMA^2 (ours)	4	**0.8232**	**83.692**	**0.7146**	**41.01**	**0.3203**	**0.4701**	**2.925**
Flash (Chadebec et al., 2025)	1	0.6104	68.438	0.6517	28.74	0.3147	0.0490	5.103
DMD2 (Yin et al., 2024a)	1	0.6960	78.352	0.6972	34.43	0.3196	<u>0.2834</u>	<u>3.489</u>
TDM (Luo et al., 2025)	1	0.7129	79.267	0.7043	34.87	0.3222	0.2785	3.672
Decoupled (Liu et al., 2026b)	1	<u>0.7826</u>	80.792	0.7044	<u>36.55</u>	<u>0.3236</u>	0.2639	3.562
DMDR (Jiang et al., 2026)	1	0.7512	<u>82.395</u>	<u>0.7076</u>	34.15	0.3229	0.2598	3.973
DMA^2 (ours)	1	**0.7914**	**82.963**	**0.7155**	**38.99**	**0.3243**	**0.3447**	**3.409**

on eight GPUs with batch size 32. Optimization uses AdamW without weight decay, at learning rates 2×10^{-6} (generator), 5×10^{-6} (fake-score network), and 2×10^{-4} (adversarial heads), with the fake-score network updated five times per generator step; classifier-free guidance is 4.0. The canonical DMA^2 student combines the fixed high-noise band $U(.02, .35)$ with dual-granularity real-data guidance over one frozen DINOv2 encoder: local parametric DINO-Adv at weight 0.01 and parameter-free semantic distribution-field AF-Loss at weight 0.05, using radii $\{.02, .05, .2\}$ and 256-token real/generated banks. The four-step generator denoises at $t \in \{0, .25, .5, .75\}$.

Dataset and benchmark. Training uses 200k image–caption pairs from the BLIP3-o long-caption split (Chen et al., 2025), which pairs open-domain images with detailed Qwen-generated captions. We evaluate prompt following with DPG-Bench (Hu et al., 2024) (1,065 readable 2-by-2 grids, four images per prompt, 4,260 images total) and compositionality on GenEval (Ghosh et al., 2023) (553 prompts, four samples each, 2,212 images). Prompt–image alignment additionally uses VQAScore (Lin et al., 2024) on the full 1,600-prompt GenAI-Bench (Li et al., 2024), with one generated image per prompt. Fidelity and no-reference quality use COCO30K with 30,000 readable 512^2 COCO samples, reporting IS, CLIP similarity, recall, and NIQE.

Compared methods. We compare against some few-step methods, each re-implemented on the same pixel-space DeCo teacher and trained under an identical 20k-step budget and shared protocol: Flash (Chadebec et al., 2025), DMD2 (Yin et al., 2024a), TDM (Luo et al., 2025), Decoupled DMD (Liu et al., 2026b), and DMDR (Jiang et al., 2026). All compared methods are trained with three independent seeds, and report the mean. Objective-specific hyperparameters follow the corresponding papers, while the data, optimizer budget, and evaluation protocol are held fixed.

4.2 Comparison with existing methods

Table 1 compares DMA^2 with five distillation baselines and the official 25-step teacher. The four-step model ranks first among the evaluated distillers on all seven metrics: it reaches 83.692 DPG-Bench and 0.8232 GenEval, ahead of the strongest corresponding baselines (DMDR at 83.160 DPG-Bench and Decoupled at 0.8100 GenEval), and raises VQAScore to 0.7146. On COCO30K it also obtains the strongest IS (41.01), recall (0.4701), and NIQE (2.925). The same advantage persists at one step, indicating that the gains are not tied to a single trajectory length.

The efficiency improvement is measured under a common training pipeline, hardware, and global batch size: DMA^2 takes 0.72 s/step, compared with 0.90 for DMDR and 1.82 for DMD2 and Decoupled DMD, making it about $2.5\times$ faster than the latter two (Appendix Table 3). Figure 6 further

Table 2: Component ablation. Bold/underline mark the best/second-best results.

Module			Metric					
DM-Band	DINO-Adv	AF-Loss	GenEval↑	DPG↑	VQA↑	IS↑	Rec.↑	NIQE↓
			0.8037	81.115	0.6930	34.24	0.2802	5.174
	✓		0.8155	82.550	0.7045	39.38	0.4297	3.337
✓	✓		<u>0.8193</u>	82.959	0.7077	<u>40.21</u>	0.4318	<u>3.034</u>
		✓	0.8090	83.194	0.7079	39.46	0.3161	3.574
	✓	✓	0.8178	<u>83.427</u>	<u>0.7113</u>	39.52	<u>0.4324</u>	3.307
✓	✓	✓	**0.8232**	**83.692**	**0.7146**	**41.01**	**0.4701**	**2.925**

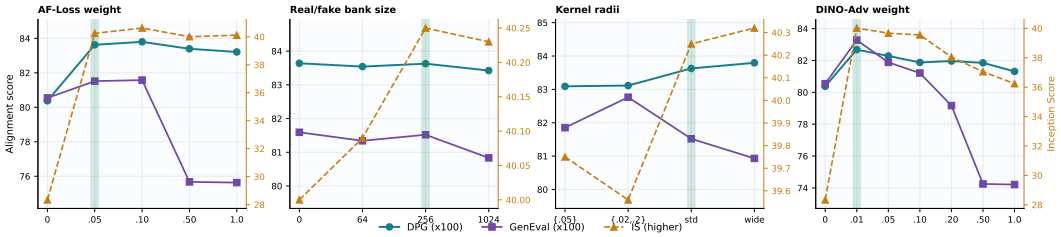

Figure 7: Hyperparameter sensitivity in DPG, GenEval, and IS; green bands mark defaults. These are OFAT sweeps around an anchor that is not the full model, so their default and zero-weight points are not directly comparable to the Table 2 component ablation (anchor protocol in Appendix D).

shows improved object count, composition, and fine structure on both long and short prompts. The recipe also transfers to PixelGen-XXL, which performs better than its teacher (Appendix B).

4.3 Ablation Studies

We study the contribution of each signal in Table 2 and sensitivity to the four main hyperparameters in Figure 7. Appendix C analyzes matching-band selection and the diagnostic that motivates it. All ablations use four-step students under the main-comparison protocol.

Component contribution. Table 2 separates DM-Band from the two RGB-native real-data objectives. DINO-Adv raises DPG-Bench from 81.1 to 82.6 and IS from 34.2 to 39.4, while AF-Loss alone gives the strongest single-objective DPG-Bench (83.2). Their combination without DM-Band reaches 83.427 DPG-Bench and 0.7113 VQAScore, showing that AF-Loss complements the local adversary. Replacing AF-Loss with direct DINOv2-CLS regression degrades six of seven metrics, including 1.12 DPG and 1.61 IS (Appendix D); thus the gain is not explained by frozen features alone. DM-Band is complementary: on the complete real-data branch it raises GenEval from 0.8178 to 0.8232 and IS from 39.52 to 41.01, while lowering NIQE from 3.307 to 2.925.

Hyperparameter ablation. Figure 7 reports one-at-a-time sweeps in DPG, GenEval, and IS. Increasing λ_{AF} from zero to .05 yields the largest improvement, whereas stronger weights sharply reduce GenEval. A bank size of 256 gives the highest IS while leaving DPG and GenEval stable; kernel-radius variants are comparatively insensitive, so we retain $\{.02, .05, .2\}$. Finally, $\lambda_{\mathrm{GAN}} = .01$ gives the best joint result among the tested adversarial weights.

5 Conclusion

We presented DMA^2, which specializes DMD's teacher matching and real-data guidance for native-RGB generation. Converging diagnostics identify a near-clean local-texture bias that motivates DM-Band. Native RGB then enables direct external supervision: DINO-Adv decouples local patch guidance from the fake-score critic, while AF-Loss adds parameter-free auxiliary semantic distribution-field guidance, operating on detached rolling supports alongside prompt-conditioned teacher supervision with no learned estimator or inference-time cost. Across four benchmarks, the four-step student exceeds its teacher on all seven metrics and ranks first among the evaluated distillers.

AI USE STATEMENT

Generative AI tools were used only for copy-editing (grammar and phrasing); all research content, code, experiments, and results are the authors' own.

REFERENCES

Clément Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. BLIP3-o: A family of fully open unified multimodal models—architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *International Conference on Machine Learning (ICML)*, 2024.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Lan Feng, Wuyang Li, Eloi Zablocki, Matthieu Cord, and Alexandre Alahi. Representation distribution matching for one-step visual generation. *arXiv preprint arXiv:2607.02375*, 2026.

Sensen Gao, Zhaoqing Wang, Qihang Cao, Dongdong Yu, Changhu Wang, and Jia-Wang Bian. Pixworld: Unifying 3d scene generation and reconstruction in pixel space. *arXiv preprint arXiv:2607.05373*, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 2023.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning (ICML)*, 2023.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 fid on imagenet512 with pixel-space diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Allan Jabri, David J. Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In *International Conference on Machine Learning (ICML)*, 2023.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Changsheng Lu, Zhen Li, Bo Zhang, Mengmeng Wang, Steven Hoi, Peng Gao, and Harry Yang. Distribution matching distillation meets reinforcement learning. In *European Conference on Computer Vision (ECCV)*, 2026.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

Xin Lin, Haodong Li, Zhifei Zhang, Yutong Yang, Haitian Zheng, Juanxi Tian, Zhe Lin, and Truong Nguyen. Pixelcontrol: Fine-grained condition fidelity in text-to-image diffusion. *arXiv preprint arXiv:2608.15705*, 2026.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*. Springer, 2024.

Bingde Liu, Wu Ran, Jinglei Zhang, Huanhuan Yuan, and Chao Ma. Efficient and high-quality depth estimation via pixel-space diffusion with linear attention. In *European Conference on Computer Vision*. Springer, 2026a.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Steven HOI, et al. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. In *International Conference on Learning Representations*, volume 2026, 2026b.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The twelfth international conference on learning representations*, 2024.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations (ICLR)*, 2017.

Yifan Lu, Qi Wu, Jay Zhangjie Wu, Zian Wang, Huan Ling, Sanja Fidler, and Xuanchi Ren. PiD: Fast and high-resolution latent decoding with pixel diffusion. *arXiv preprint arXiv:2605.23902*, 2026.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Thuan Hoang Nguyen and Anh Tran. SwiftBrush: One-step text-to-image diffusion model with variational score distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In *Advances in Neural Information Processing Systems*, 2021.

Lingchen Sun, Rongyuan Wu, Xiangtao Kong, Jixin Zhao, Qiaosi Yi, Yujing Sun, Shuaizheng Liu, Zhengqiang Zhang, and Lei Zhang. Pixrestore: Unified image restoration via pixel diffusion transformer. *arXiv preprint arXiv:2608.16793*, 2026.

Gangwei Xu, Haotong Lin, Hongcheng Luo, Xianqi Wang, Jingfeng Yao, Lianghui Zhu, Yuechuan Pu, Cheng Chi, Haiyang Sun, Bing Wang, et al. Pixel-perfect depth with semantics-prompted diffusion transformers. *Advances in Neural Information Processing Systems*, 2025.

Haofei Xu, Rundi Wu, Philipp Henzler, Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Marc Pollefeys, Andreas Geiger, Federico Tombari, and Michael Niemeyer. Pointdit: Pixel-space diffusion for monocular geometry estimation. *arXiv preprint arXiv:2607.02515*, 2026.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zhiyuan Yuan, Guanying Chen, Lingteng Qiu, Ruimao Zhang, Shuguang Cui, and Xiaochun Cao. Pxdepth: Pixel-space modeling for structure preserving monocular depth estimation. *arXiv preprint arXiv:2608.16984*, 2026.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024.

A Implementation and Reproducibility Details

Optimization. Training uses eight data-parallel ranks, per-rank batch size 4, and 20k optimizer steps. The generator, fake-score network, and DINOv2 heads use separate optimizers. The generator-side update ratio is 5. Checkpoints are recorded at 5k increments. The DINOv2 backbone remains in evaluation mode and is never updated.

Timing protocol and parameter accounting. One *step* denotes a single generator optimizer update, so "20k steps" counts generator updates; the fake-score critic is updated five times per generator step (update ratio 5). Reported seconds-per-step are end-to-end wall-clock over one generator step *including* its five critic updates and all loss terms, measured on eight A100-80GB GPUs in bf16 autocast mixed precision after 50 warm-up steps, averaged over 200 steps with device synchronization. The 1.58M figure is the learnable size of the adversarial heads only: the 560M fake-score critic is trained by every method, so the speedup does not come from fewer trainable parameters but from removing the adversarial gradient path through that heavy critic. DMD2's feature-GAN back-propagates its discriminator through the 560M critic, whereas DINO-Adv attaches a 1.58M head to a frozen 86M DINOv2. Here "frozen" means we do not compute or apply gradients to the encoder weights; the input-gradient path through $\phi(\hat{x})$ to the image, and hence to G_θ, is retained, as the generator update requires.

Same-pipeline training speed. Table 3 reports end-to-end optimizer-step time measured under the same training pipeline, hardware, and global batch size. DMD2 and Decoupled each require 1.82 seconds per step, DMDR requires 0.90 seconds, and DMA^2 requires 0.72 seconds per step.

Table 3: Same-pipeline training speed. All methods use the same hardware, global batch size, and end-to-end step-timing protocol. Lower is better.

Method	Time (s/step)↓
DMD2 (Yin et al., 2024a)	1.82
Decoupled DMD (Liu et al., 2026b)	1.82
DMDR (Jiang et al., 2026)	0.90
DMA^2 (ours)	**0.72**

Overlap diagnostic. We use 1,536 matched real/generated pairs and five grouped splits. DINOv2 layers $\{2, 5, 8, 11\}$ are mean-pooled and concatenated. Separability is measured in matching-time increments of .05 across early, middle, and late generator checkpoints and stages $t_g \in \{0, .25, .5, .75\}$, showing that the noise-dependent separability pattern is essentially shared across generation stages (Section 3.1).

Exact evaluation. DPG acceptance requires 4,260 generated images, 1,065 readable 2-by-2 grids, and zero skipped grids. GenAI-Bench acceptance requires 1,600 readable images, one for each prompt, scored with VQAScore. COCO30K acceptance requires 30,000 readable 512×512 images and zero corrupt files. GenEval acceptance requires $553 \times 4 = 2,212$ images scored with the original DeCo evaluator. The result registry closes an endpoint only when all required metrics and the exact source revision are recorded.

B Extension to another backbone: PixelGen

To test whether the recipe transfers beyond DeCo, we apply the same DMA^2 distillation to a PixelGen-XXL teacher and evaluate few-step students against it. Table 4 reports the same alignment and quality metrics as the main comparison. The PixelGen-XXL results show the same trend: the four-step student performs better than its 25-step teacher on every reported metric (e.g., DPG $79.360 \rightarrow 81.708$, GenEval $0.7965 \rightarrow 0.8112$, IS $37.24 \rightarrow 40.46$), and the one-step student still stays ahead on dense-prompt following, VQAScore, and CLIP, providing evidence that the native-RGB matching band, DINOv2 adversary, and Anchor-Field loss are not tied to a single backbone.

Table 4: Extension to a PixelGen-XXL teacher: DMA^2 few-step students vs. the original PixelGen teacher. Metrics as in Table 1; ↑/↓ give direction, **bold** is best and underline second best per column. The four-step student uses the strictly-rerun schedule-matched replica.

Model	NFE	GenEval↑	DPG↑	VQA↑	IS↑	CLIP↑	Rec.↑	NIQE↓
PixelGen-XXL teacher	25	0.7965	79.360	0.6804	37.24	0.3141	0.3328	4.064
DMA^2-PixelGen (ours)	4	**0.8112**	**81.708**	0.6897	**40.46**	0.3184	**0.4892**	**3.088**
DMA^2-PixelGen (ours)	1	0.8072	80.898	**0.6961**	39.48	**0.3232**	0.3872	3.275

C Noise-band selection and separability

This appendix supports the fixed high-noise matching band of Section 3.1 with the separability diagnostic that motivates it and the full band-position sweep.

Matching-band position. Table 5 moves the DMD matching range across noise while holding everything else fixed. The tight high-noise band $U(.02, .35)$ gives the best prompt following (82.959 DPG-Bench) and compositionality (0.8193 GenEval), leading on nearly every alignment, compositional, and no-reference quality metric. Narrowing the cap to $U(.02, .20)$ or widening it to $U(.02, .70)$ both erode these (82.720/82.426 DPG-Bench, 0.8095/0.8113 GenEval); raising the lower edge with $U(.15, .70)$ erodes them further (82.038 DPG-Bench, 0.7924 GenEval); and pushing supervision fully into the low-noise regime $U(.60, .90)$ collapses them (80.330 DPG-Bench, 0.7341 GenEval; color-attribute binding $0.735 \rightarrow 0.538$). This pattern supports selecting the high-noise band for alignment and compositionality.

Table 5: Noise selection: matching-band position, width, and per-stage nesting. Four-step DINO-Adv students (20k, λ_{GAN}=.01, AF-Loss disabled) differ only in the DMD matching range; ↑/↓ give direction and bold is best per row. The high-noise band $U(.02, .35)$ leads on nearly every alignment metric.

DMD matching range	$U(.02, .20)$ *narrow*	$U(.02, .35)$ *best*	$U(.02, .70)$ *one-sided*	$U(.15, .70)$ *wide*	$U(.60, .90)$ *low-noise*	$U(.02, .98)$ *full range*	$U(.02, \{.25, .50, .75, .98\})$ *per-stage*
GenEval↑	0.8095	**0.8193**	0.8113	0.7924	0.7341	0.8155	0.8164
DPG-Bench↑	82.720	**82.959**	82.426	82.038	80.330	82.550	82.018
CLIP↑	0.3201	**0.3225**	0.3184	0.3213	0.3194	0.3214	0.3211
IS↑	38.79	**40.21**	39.78	39.30	36.21	39.38	38.71
Recall↑	0.4182	**0.4318**	0.4198	0.4062	0.3773	0.4297	0.4027
NIQE↓	3.257	**3.034**	3.296	3.424	3.636	3.337	3.596

Band width: the cap sits at .35. Fixing the lower edge at the numerical floor .02 and sweeping only the *upper* edge b of a uniform band $U(.02, b)$ traces out the cap directly (Table 5). DPG rises from 82.720 at $b = .20$ to a peak of 82.959 at $b = .35$, then declines as the band admits more low-noise, texture-shortcut mass (82.426 at $b = .70$, 82.550 at $b = .98$, i.e. nearly the full range). The best tested cap coincides with the $t_m = .35$ transition suggested by the representation diagnostic (Section 3.1), and widening past it degrades prompt following.

Single-step supervision. Representation supervision becomes decisive at one step. The full anchor recipe scores 82.963 DPG-Bench and 0.7914 GenEval, whereas the exact one-step no-GAN retrain reaches only 78.980 DPG-Bench and 0.7590 GenEval. The frozen-DINOv2 signals thus recover most of the alignment that is lost when the trajectory collapses to a single step, though the balanced four-step model still leads on most reported metrics.

D Anchor-Field ablations

AF-Loss implementation details. Algorithm 1 gives the exact field used in our runs; the text below states its gradient-affecting choices. **Reference clouds:** the current all-gathered real/generated features are prepended to the rolling banks and truncated to N_b=256, so the pool spans the current and most recent batches; the banks are updated *after* the generator step. **Scale $\bar{d}$:** a single detached

Algorithm 1 AF-Loss: one generator step (auxiliary to the DMD/GAN updates)

Require: generator G_θ; frozen encoder ϕ; rolling banks B_r, B_f; radii $\mathcal{R}=\{.02, .05, .2\}$; pool $N_b=256$; weight λ_{AF}

$\hat{x} \leftarrow G_\theta(\epsilon, c)$; $z \leftarrow \phi(\hat{x})$ (query, keeps grad); $z_r \leftarrow \phi(x)$ ▷ x: real image

all-gather (no grad): $\text{cur}_r \leftarrow z_r$, $\text{cur}_f \leftarrow z$ ▷ local rows placed first

$P \leftarrow [\text{cur}_r; B_r]_{1:N_b}$, $Q \leftarrow [\text{cur}_f; B_f]_{1:N_b}$ ▷ current batch is included in the field

$T \leftarrow [\,Q\,;\,P\,]$ (negatives then positives); $D_{iu} \leftarrow \|z_i - T_u\|_2$

$\bar{d} \leftarrow \text{mean}(D)$; $\tilde{D} \leftarrow D/\bar{d}$ ▷ $\bar{d}$ detached

self-mask: $\tilde{D}_{ii} \mathrel{+}= c$ (large) on the $|Q|$ negative columns ▷ z_i vs its own copy

$V \leftarrow 0$

for $R \in \mathcal{R}$ **do**

$A \leftarrow \sqrt{\text{softmax}_{\text{targets}}(-\tilde{D}/R) \odot \text{softmax}_{\text{queries}}(-\tilde{D}/R)}$ ▷ symmetrized affinity

$A^- \leftarrow A_{:,\le|Q|}$, $A^+ \leftarrow A_{:,>|Q|}$; $C \leftarrow \left[\, -A^- \odot \sum_u A^+ \;\middle|\; A^+ \odot \sum_u A^- \,\right]$ ▷ cross-coupled

$F \leftarrow C\,T - (\sum_u C) \odot z$; $V \mathrel{+}= F / \text{RMS}(F)$ ▷ per-R force norm

end for

$\mathcal{L}_{\text{AF}} \leftarrow \frac{1}{B}\sum_i \|z_i - \text{sg}(z_i + V_i)\|_2^2$; add $\lambda_{\text{AF}}\nabla_\theta\mathcal{L}_{\text{AF}}$ to DMD/GAN grads; update G_θ ▷ grad only

$z \to \hat{x} \to G_\theta$

$B_r \leftarrow [\text{cur}_r; B_r]_{1:N_b}$, $B_f \leftarrow [\text{cur}_f; B_f]_{1:N_b}$ ▷ detached FIFO update, for next step

scalar, the mean of the full query×target L_2 distance matrix. **Symmetrized kernel / self-mask:** at each scale the affinity is the elementwise geometric mean of a softmax over targets and a softmax over queries of $-d(z_i, u)/(\bar{d}R)$ (doubly normalized); each z_i is masked from its own copy among the negatives, the attraction and repulsion coefficients are cross-coupled, and each scale's displacement is RMS-normalized (the RMS is taken over the entire query×feature matrix, per rank) and then summed over $\mathcal{R}=\{.02, .05, .2\}$. **Reduction:** $\mathcal{L}_{\text{AF}} = \frac{1}{B}\sum_{i=1}^{B}\|z_i - \text{sg}(z_i + V(z_i))\|_2^2$, the batch mean of the per-sample squared L_2. T he code matches Algorithm 1.

Figure 7 protocol. Each panel is a single-factor sweep around a fixed four-step, 20k-step anchor with the high-noise band $U(.02, .35)$: the AF-Loss weight/bank/radii panels keep DINO-Adv off and AF-Loss on (bank 256, radii $\{.02, .05, .2\}$, λ_{AF}=.05 except the swept factor), while the DINO-Adv-weight panel keeps AF-Loss off. Each point is a single 20k run rather than a three-seed mean, so it selects defaults but is not directly comparable to the full-model numbers in Tables 1–2.

Why this field rather than another feature-space objective? To test whether the gain comes simply from applying a frozen-DINOv2 penalty, and whether any distribution-matching objective in the same space would do, we compare AF-Loss against two alternatives on the identical normalized last-layer CLS features, encoder, four-step protocol, and loss weight (.05): (i) direct perceptual matching (regression to real CLS features), and (ii) a matched multi-scale *DINOv2-CLS MMD* that pulls the generated feature distribution toward the same 256-feature real bank with the same kernel bandwidths, differing from AF-Loss only in that it minimizes a batch-level distribution distance rather than injecting a sample-dependent attract–repel displacement. All rows use the AF-Loss-only ablation setting in Table 2: DM-Band and DINO-Adv are disabled, and only the feature-space objective is changed. All three add no learnable parameters. Table 6 shows AF-Loss is best on five of the seven metrics; direct regression attains the highest GenEval and MMD the highest recall. Against the stronger MMD baseline it raises DPG by 1.317 (83.194 vs 81.877), IS by 3.69, VQA by 0.0130, and CLIP by 0.0040, approximately matches it on GenEval (0.8090 vs 0.8077), and lowers NIQE by 0.859; MMD attains slightly higher recall (0.3330 vs 0.3161), indicating marginally broader coverage at the cost of alignment and quality. Direct regression gives higher GenEval but is worse on the other six metrics. The gains therefore do not arise from introducing DINOv2 features, nor from distribution matching per se; they depend on the specific sample-dependent attract–repel field.

Which feature. Table 7 ablates which frozen-DINOv2 feature the Anchor Field acts on, holding the field form and weight fixed at $\lambda_{\text{AF}} = .05$. The normalized final-layer CLS feature, our default, gives the best DPG.

Attract vs. repel. Table 8 isolates the two halves of the field, keeping the feature and weight fixed: attracting the student toward the real bank B_r, repelling it from the fake bank B_f, and both together (our default). Attraction toward the real bank supplies most of the gain, repulsion from the fake bank

Objective	GenEval↑	DPG↑	VQA↑	IS↑	CLIP↑	Rec.↑	NIQE↓
AF-Loss (CLS field, ours)	0.8090	**83.194**	**0.7079**	**39.46**	**0.3246**	0.3161	**3.574**
Direct DINOv2-CLS perceptual	**0.8156**	82.077	0.7028	37.85	0.3217	0.3146	4.675
DINOv2-CLS MMD (distribution matching)	0.8077	81.877	0.6949	35.77	0.3206	**0.3330**	4.433

Table 6: AF-Loss versus two alternative feature-space objectives on the same normalized last-layer CLS features, encoder, four-step protocol, and loss weight (.05): direct DINOv2-CLS perceptual matching, and a matched multi-scale DINOv2-CLS MMD (distribution matching) against the same 256-feature real bank. Only the feature-space objective changes; **bold** is best per column.

AF feature	DPG↑
DINOv2 $\{2, 5, 8, 11\}$ spatial patch features (per-layer mean-pool, concat)	81.970
DINOv2 $\{2, 5\}$ spatial patch-feature mean+std	81.430
DINOv2 $\{2, 5\}$ spatial patch-feature quantile	81.760
DINOv2 normalized final-layer CLS feature (ours)	**83.194**

Table 7: Anchor-Field feature ablation. All variants share the same field form at $\lambda_{\mathrm{AF}} = .05$; the normalized final-layer CLS feature is our default.

alone is markedly weaker, and the full attract–repel field is best; we therefore keep both terms as our default.

Anchor-Field term	DPG↑
Attract only (real bank B_r)	82.910
Repel only (fake bank B_f)	80.170
Attract + repel (ours)	83.194

Table 8: Anchor-Field attract/repel ablation, at the default feature and $\lambda_{\mathrm{AF}} = .05$.

E Additional Qualitative Comparisons

Figure 8 shows eight further four-step comparisons under matched prompts, complementing Figure 6. As in the main text, DMA[2] keeps object count, anatomy, and composition correct (two rabbits, a single rider on one horse, the dog beside the penguin, the two sheep) while the critic-coupled baselines more often duplicate or distort subjects.

F Limitations

Two limitations remain. First, DINO-Adv and AF-Loss inherit the inductive biases of the frozen DINOv2 representation. Although its multi-layer patch features and final-layer CLS feature provide complementary local and semantic guidance, they may not fully capture text-conditioned spatial relations or domain-specific visual attributes. Second, our evaluation emphasizes established automated benchmarks and qualitative comparisons; broader human-preference, fairness, and safety evaluations remain future work.

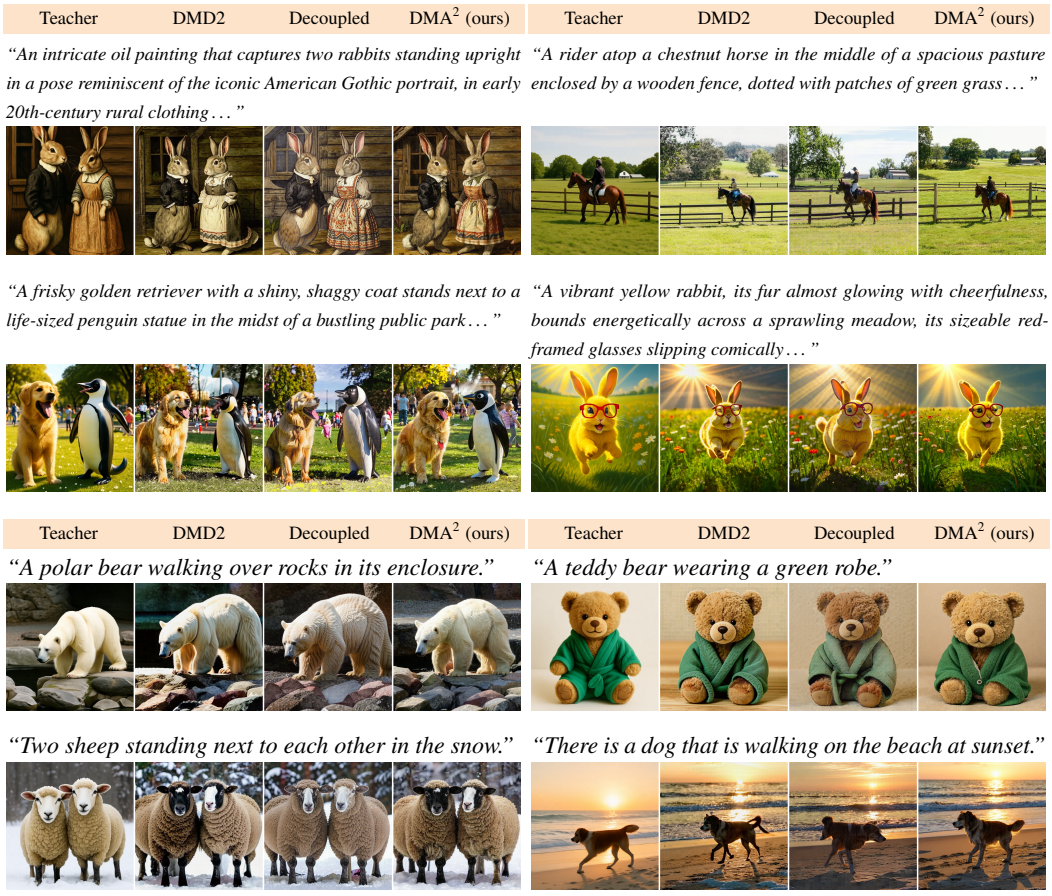

Figure 8: Additional four-step qualitative comparisons under identical prompts (columns: 25-step teacher, DMD2, Decoupled, and DMA^2). Top block: long DPG-Bench prompts (abbreviated with "…"); bottom block: short COCO captions.