REPRESENTATION DISTRIBUTION MATCHING FOR ONE-STEP VISUAL GENERATION

Lan Feng[1] **Wuyang Li**[1] **Éloi Zablocki**[2] **Matthieu Cord**[2,3] **Alexandre Alahi**[1]
[1]EPFL, Switzerland [2]Valeo.ai, France [3]Sorbonne Université, France

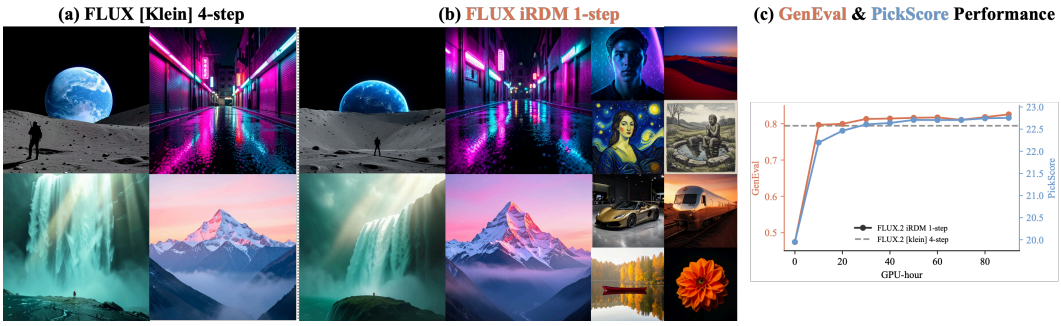

Figure 1: **iRDM post-trains the four-step FLUX.2 [klein] into a one-step generator at matched quality.** (a) Four-step FLUX.2 [klein]. (b) One-step iRDM after post-training with the joint image-text objective.(c) GenEval and PickScore over post-training compute, the one-step model surpassing the four-step version (grey dashed) on both metrics in about 90 H200 GPU-hours.

ABSTRACT

We elucidate the design space of Representation Distribution Matching (RDM), our name for the paradigm that trains a one-step image generator by matching generated and reference feature distributions under frozen pretrained encoders. We identify two design axes, how the distributions are compared and the representations they are compared in, and controlled studies along them yield three findings. First, the classical MMD, which could not train convincing generators a decade ago, becomes a strong and scalable objective once estimated right. Second, the generated batch is then the operative variable, with an optimum above 2048, far beyond customary batch sizes. Third, any single representation can be gamed, driven below the real score while images stay visibly fake, so we match against a balanced battery of encoders and evaluate with $\mathrm{SW}_{r^{14}}$, a Sliced-Wasserstein distance over 14 encoders that is independent of the training loss and resists gaming. Combining the preferred choices yields improved RDM (iRDM): it sets the one-step state of the art on ImageNet at $\mathrm{SW}_{r^{14}}$ 1.30, corroborated by PickScore, a human-preference proxy our objective never optimizes, which prefers it over the prior best one-step generator on 71.2% of matched samples. The same recipe post-trains the four-step FLUX.2 [klein] into a one-step generator, surpassing the four-step version on GenEval, 0.826 to 0.794, and on PickScore, 22.76 to 22.58, in 90 H200 GPU-hours.

1 INTRODUCTION

Generative modeling is fundamentally distribution matching: we want a generator whose output distribution matches the data, and we already judge that match by the distance between their representation distributions, the basis of FID (Heusel et al., 2017). Diffusion and flow models pursue this distributional goal only implicitly, learning to reverse a noising process so that many denoising steps, simulated at inference, carry noise onto the data (Ho et al., 2020; Song et al., 2021; Lipman

et al., 2023). A recent alternative pursues it explicitly and directly, matching the two distributions in the feature space of a frozen pretrained encoder and producing an image in a single network evaluation, with no online teacher, adversary, or trajectory to simulate. We refer to this paradigm as Representation Distribution Matching (RDM).

Several recent one-step generators (Deng et al., 2026; Yang et al., 2026) can be viewed as this paradigm, differing along just two axes. The first is the comparison: which discrepancy scores the gap between the generated and real feature laws, how it is estimated from finite samples, and what reference stands in for each side. The drifting field measures pairwise kernel forces within each batch and reads its real reference off that same batch (Deng et al., 2026), while the Fréchet-distance loss keeps only the first two moments, precomputed over the full dataset (Yang et al., 2026). The second axis is the representations, by which we mean the frozen encoder feature spaces in which the two distributions are compared; here every method has settled on the same default, a few encoders under fixed weights.

Existing methods fix these choices jointly, so it is unclear which of them is responsible for quality. We vary one axis at a time, and the resulting controlled studies overturn several assumptions implicit in current practice.

Start with the comparison. The maximum mean discrepancy was dismissed a decade ago as too weak to train a competitive generator (Li et al., 2015; Dziugaite et al., 2015); it was never too weak, only badly estimated. A good estimate needs a structured feature space and enough samples on each side, and the two sides differ. The reference is fixed in advance and never moves, so we use all of it: the entire 1.28M-image training set is compressed once into a frozen Nyström reference (Chatalic et al., 2022), 4096 landmarks standing in for the attraction at a fraction of the cost (fig. 3). The generated side moves at every step and is drawn fresh, where a larger batch sharpens the estimate but buys fewer updates; the optimum lies above 2048, an order of magnitude past common practice, with gradient caching (Gao et al., 2021) absorbing the memory. Finally, on conditional tasks we match the joint law of caption and image features rather than the image marginal alone, making prompt fidelity part of the objective, which post-trains four-step FLUX.2 into a one-step model at a higher GenEval.

Now the representations. Modern pretrained encoders already provide good spaces in which to measure the distance, so the question is which space, or which combination of spaces, makes a low MMD achievable only by genuinely realistic samples. A single encoder is not enough: the generator overfits whichever one it trains against, beating the real data on that encoder's own score while its samples stay visibly fake. The fix is to rely on none alone. We match across a diverse battery of encoders, and rather than weight them uniformly we keep them in balance by constrained optimization: a proportional Lagrangian controller (Stooke et al., 2020) upweights whichever encoder is hardest to satisfy and downweights whichever the generator is beginning to overfit. The intuition is the weakest-stave rule: just as a bucket holds water only to its shortest stave, a viewer judges an image by its most pronounced artifact (Larson and Chandler, 2010; Wang and Shang, 2006), so the encoder that still objects is the one worth heeding.

Combining the two axes gives improved RDM (iRDM), a simple but effective recipe that generates in a single step at higher quality. We measure it with our new metric $\mathrm{SW}_{r^{14}}$, a relative Sliced-Wasserstein distance averaged over 14 pretrained encoders, with real data scaled to 1. As an evaluation metric the Sliced-Wasserstein distance is harder to game than the Fréchet distance or the MMD (Berthet et al., 2026), and since we never train against it, a gain rules out reward hacking. Post-training pMF-H FD-SIM (Lu et al., 2026; Yang et al., 2026), whose $\mathrm{SW}_{r^{14}}$ result held the previous state of the art 2.05, iRDM reaches a new one-step state of the art at $\mathrm{SW}_{r^{14}}$ 1.30, corroborated by a 71.2% PickScore (Kirstain et al., 2023) win rate, a learned human-preference proxy our objective never optimizes. The recipe carries to text-to-image: applied to FLUX.2 [klein] (Black Forest Labs, 2026), a 4B four-step generator, iRDM post-trains it into a one-step model that surpasses the four-step version on GenEval, 0.826 to 0.794, and on PickScore, 22.76 to 22.58, in 90 H200 GPU-hours. We summarize our contributions as follows.

- **A unifying framework.** We formalize distribution matching into a single paradigm, RDM, that needs no online teacher, and identify the two design axes that govern it, how the distributions are compared and the representations they are compared in. This lets us trace the quality ceiling of a method to a specific design choice rather than its headline idea.

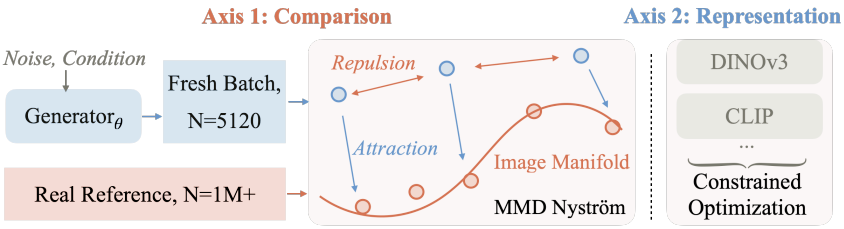

Figure 2: iRDM trains a one-step generator by representation distribution matching alone: no online teacher, no adversary, no trajectory. Each step draws a fresh batch of N samples and embeds it, together with a reference computed once and frozen, under a battery of ten pretrained encoders. In every feature space, generated samples are pulled toward the reference manifold by a Nyström attraction and kept apart by an exact within-batch repulsion (eq. (3)).

- **A simple recipe at the state of the art.** Varying each axis in isolation, we establish what drives quality: the right way to estimate the MMD, an exact within-batch repulsion paired with a Nyström attraction to a frozen full-data reference; large fresh generation batches; a joint image-text objective on text-to-image tasks; and a constrained optimization that keeps a diverse encoder battery in balance. These choices combine into iRDM, which reaches state-of-the-art one-step ImageNet generation at $\mathrm{SW}_{r^{14}}$ 1.30 against the real-data 1, with no online teacher, adversary, or reward model; the same recipe post-trains four-step FLUX.2 [klein] into a one-step model at a higher GenEval than the four-step base.
- **A metric that resists gaming.** We evaluate with $\mathrm{SW}_{r^{14}}$, a Sliced-Wasserstein distance averaged over 14 encoders and real data scoring 1 by construction, an optimal-transport metric independent of the training loss and far harder to game than any single-encoder score.

2 Related Work

One-step and few-step generation. Diffusion and flow models (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Karras et al., 2022; Rombach et al., 2022; Peebles and Xie, 2023) pay an inference cost per denoising step. Step reduction either distills a pretrained teacher or removes it. Distillation matches the teacher's trajectory, score, or moments (Salimans and Ho, 2022; Liu et al., 2023; Luo et al., 2023; Yin et al., 2024b; Zhou et al., 2024; Salimans et al., 2024) or trains against an adversary (Sauer et al., 2024; Yin et al., 2024a); the teacher-free route constrains the model on its own outputs or trajectories (Song et al., 2023; Song and Dhariwal, 2024; Geng et al., 2025b; Lu and Song, 2025; Frans et al., 2025; Geng et al., 2025a; Lu et al., 2026; Geng et al., 2026; Zhou et al., 2025). RDM needs no online teacher and constrains no trajectory: it compares generated samples against a frozen reference directly.

Matching distributions in fixed feature spaces. Casting generation as distribution matching is the GAN program (Goodfellow et al., 2014; Salimans et al., 2016); with fixed kernels it gave moment matching networks (Li et al., 2015; Dziugaite et al., 2015), adversarial kernels (Li et al., 2017; Bińkowski et al., 2018), and sliced Wasserstein generators (Deshpande et al., 2018; Wu et al., 2019). What changed since is the feature space: frozen pretrained encoders, used for perceptual losses (Johnson et al., 2016; Zhang et al., 2018), discriminator features (Sauer et al., 2021; Kumari et al., 2022), and alignment targets (Yu et al., 2025), now support direct feature-distribution matching, as the drifting field (Deng et al., 2026), the Fréchet-distance loss (Yang et al., 2026), and a concurrent Sinkhorn flow (Han et al., 2026) show. The principle runs implicitly through this lineage; our contribution is to name it, chart its two design axes, and locate prior methods within them.

3 Representation Distribution Matching and its Design Space

A one-step generator g_θ maps a prior $z \sim p_z$ to an image in a single evaluation, with output law p_θ. Given a frozen encoder ϕ that sends an image to a feature $\phi(x) \in \mathbb{R}^D$, RDM aligns the feature laws

of generated and real data (fig. 2),

$$\mathcal{L}(\theta) = \mathcal{D}(\phi_* p_\theta,\ \phi_* p_{\text{data}})\,, \tag{1}$$

where ϕ_* is the pushforward and $\mathcal{D}$ a distance between distributions. Constraining the output distribution rather than a per-sample trajectory makes the generator one-step by construction; the same objective post-trains a few-step sampler by treating its final output as g_θ.

Every instance of eq. (1) is fixed by two choices, the axes of this paper: **the comparison**, set by which discrepancy $\mathcal{D}$ scores the feature laws, which estimator computes it from finite samples, what reference stands in for each side, and which joint law is matched under conditioning (Sections 3.1 and 3.2); and **the representations**, which encoders define the feature spaces and how several are weighted (Section 3.3).

Our decomposition locates prior methods on these axes and attributes each method's ceiling to a specific choice. The Fréchet-distance loss freezes a global data-side reference, the right call, but compresses it to two moments, so matching can saturate while images stay flawed. The drifting field has a sharp pairwise estimator, but it rebuilds its reference from every batch at a cost that confines it to small batches, exactly where a distribution estimate is noisiest. Both train against a few encoders under fixed weights, which Section 3.3 shows is gameable. iRDM is the combination of the preferred choice on each axis.

3.1 The comparison axis: choosing and estimating the discrepancy

A positive definite kernel k on feature space defines

$$\mathrm{MMD}^2(P,Q) \;=\; \mathbb{E}_{x,x'\sim P}\,k(x,x') \;-\; 2\,\mathbb{E}_{x\sim P,\,y\sim Q}\,k(x,y) \;+\; \mathbb{E}_{y,y'\sim Q}\,k(y,y'), \tag{2}$$

which vanishes exactly when $P = Q$ for a characteristic kernel such as the Gaussian (Gretton et al., 2012; Sriperumbudur et al., 2010). We adopt the squared MMD with this Gaussian kernel, $k(x,y) = \exp\!\big(-\|x-y\|_2^2/2\sigma_\phi^2\big)$, on the raw encoder embeddings; the bandwidth σ_ϕ is fixed per encoder by the median heuristic and held at a single scale. What a generator optimizes is a finite-sample estimate, and the estimator sets its cost, its variance, and the blind spots it can exploit.

Write $g_i = \phi(g_\theta(z_i))$ for the features of a generated batch of size B. Of the three terms of eq. (2), the data term is constant in θ and dropped; the cross term attracts generated features toward the data; the generator term repels them from one another, the only force preventing collapse onto the densest modes. The two demands are opposite, so we estimate the terms differently,

$$\widehat{\mathcal{L}}_\phi \;=\; \underbrace{\frac{1}{B^2}\sum_{i,j} k(g_i,g_j)}_{\text{repulsion, exact}} \;-\; \underbrace{\frac{2}{B}\sum_{i}\psi(g_i)^\top \bar{\mu}_\phi}_{\text{attraction, Nyström}}, \tag{3}$$

where ψ is the Nyström feature map and $\bar{\mu}_\phi$ the frozen reference mean embedding it induces over the full training set, both made precise below. Every batch is scored by all encoders in the battery, and we sum $\widehat{\mathcal{L}}_\phi$ over them each step with the adaptive weights of Section 3.3.

An exact repulsion, a frozen attraction. The two terms sum over different sets and we estimate them differently. The repulsion runs only within the batch, where the exact $B \times B$ kernel sum is cheap, so we leave it exact, one matrix per encoder. The attraction instead compares against the full training set: resampling it each step, as the standard two-sample estimator does, injects reference noise that grows as the bandwidth shrinks, so we compute it once and freeze it through a Nyström kernel mean embedding (Chatalic et al., 2022). With m=4096 landmarks ℓ_j placed by k-means on the data features and kernel matrix K_{mm}, $\psi(x) = K_{mm}^{-1/2}\big(k(x,\ell_1),\dots,k(x,\ell_m)\big)^\top$ makes $\psi(x)^\top\psi(y)$ the Nyström approximation of $k(x,y)$, and $\bar{\mu}_\phi = \frac{1}{n}\sum_t \psi(r_t)$ is precomputed once over all $n = 1.28$M training images and frozen. Each step pulls the batch toward this zero-variance summary at cost $\mathcal{O}(Bm)$, negligible next to the encoder forward passes.

Why MMD, and why Nyström. Each alternative discrepancy gives up one of these advantages: the Fréchet distance collapses each side to two moments and can saturate while samples stay flawed;

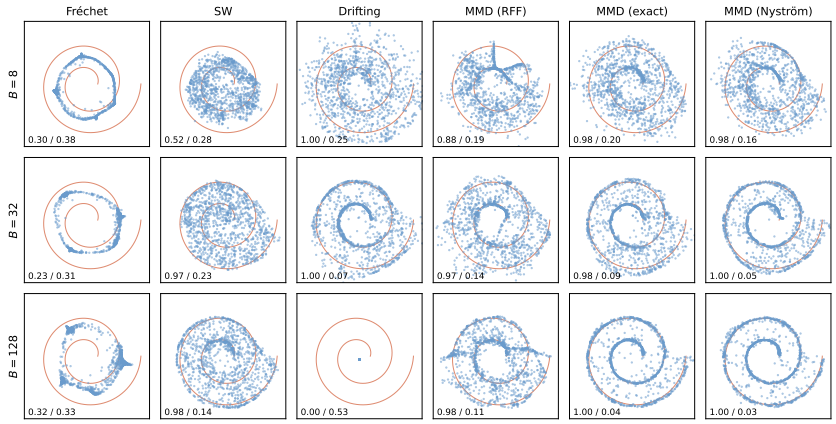

Figure 3: Spiral diagnostic at ambient dimension $D = 64$. Rows sweep the training batch size B, and columns are different methods. Fréchet is the Gaussian 2-Wasserstein on a frozen global mean and covariance; sliced-Wasserstein uses $L = 1000$ resampled projections; drifting is the faithful coupled field, best-effort tuned with a reference bank of 128; and the MMD family uses a multi-scale Gaussian kernel with $m = 512$ features or landmarks per scale, where random features and Nyström match a frozen global reference mean while the exact estimator sees B reference samples per step. Corner numbers are anchor recall / medDist, the median distance to the curve, on which real data floors at 0.033. Fréchet is batch-insensitive and never traces the curve, its two moments cannot encode the manifold; the other distances fail by sampling instead, sliced-Wasserstein collapsing at small B, drifting at large B, and random features with dimension. Nyström is the sharpest in every row and the only distance strong across all regimes.

sliced-Wasserstein relies on sorting within each projection, so it scores the batch only against a resampled batch rather than the full real distribution (Deshpande et al., 2018); and the drifting field is a per-particle normalized form of the same MMD gradient, steadier at small batches but reducing to the plain MMD as the batch grows, its resampled per-batch reference keeping it small-batch (Deng et al., 2026). For the attraction term, Nyström landmarks beat random Fourier features (Rahimi and Recht, 2007): the landmark basis is data-dependent, centered on real points and accurate exactly where generation happens, whereas global cosines spend capacity over an ambient space the manifold barely occupies and leave unresolved directions that a generator under optimization pressure exploits. Theory concurs, with data-dependent bases dominating once the kernel spectrum decays quickly and m of order $\sqrt{n}\,\log n$ landmarks retaining the exact embedding's $n^{-1/2}$ rate (Yang et al., 2012; Chatalic et al., 2022). A controlled study makes both choices concrete.

A controlled study of the estimator. Real encoders place data on a thin manifold in a high-dimensional space, and we isolate this regime with a known target: following Li and He (2025), a two-turn spiral buried in $\mathbb{R}^{64}$ by a fixed orthonormal map, the same MLP generator trained under each objective at a matched budget while the batch sweeps $B \in \{8, 32, 128\}$, scored by anchor recall and medDist, the median distance to the curve, on which real data scores 0.033 (settings in fig. 3).

At the largest batch both MMD estimators, MMD exact and MMD Nyström, lock onto the spiral, while random features stay diffuse, sliced-Wasserstein stays loose, and drifting collapses. As the batch shrinks, MMD exact degrades as its per-batch reference thins, whereas MMD Nyström pulls toward the same frozen reference at every batch size and stays sharpest in every row; sliced-Wasserstein loses recall at the smallest batch and drifting collapses at the largest. MMD Nyström is the only method that fails nowhere.

3.2 The comparison axis: batches and conditioning

The generator side: large, fresh batches. With the data side frozen once over the full training set, the generated distribution is the only quantity still moving, and it moves at every step, so it

must be sampled fresh; estimating it from a stale buffer, as the EMA queue of Yang et al. (2026) does, biases the gradient off-policy. A fresh batch makes its size N the operative variable: a larger N lowers the variance of the estimate but, at a fixed compute budget, buys fewer optimizer steps, trading estimate sharpness against the number of updates. Large fresh batches are normally ruled out by memory, which gradient caching (Gao et al., 2021) removes by accumulating the exact full-batch gradient in chunks at the cost of one chunk. We sweep N at a matched wall-clock budget, scaling the learning rate as $\sqrt{N}$ (Malladi et al., 2022) so every arm sees about one epoch split into more or fewer updates (fig. 4). Quality climbs with N: the trained encoder sharpens while the held-out-dominated panel barely moves, the smallest batch is noise-dominated and regresses despite far more optimizer steps, and the curve then flattens into a broad optimum. We adopt $N{=}5120$ for ImageNet and the larger $N{=}10240$ for the FLUX post-training; exact values are in Appendix C.

Conditional tasks: match the joint, not the marginal. A prompted generator can satisfy the image marginal while drifting from its prompts: realism bought with alignment. We instead match the joint law. With a frozen text encoder τ and coupled features $\Phi(x, c) = \phi(x) \oplus \tau(c)$,

$$\mathcal{L}_{\text{joint}}(\theta) \;=\; \mathcal{D}\big(\Phi_* p_\theta,\; \Phi_* p_{\text{data}}\big), \tag{4}$$

where reference pairs couple each image with its caption and generated pairs couple each output with the prompt that produced it; the estimator is unchanged, landmarks now reference image-text pairs. Under the kernel a generated image is pulled toward reference images whose captions resemble its prompt, so prompt fidelity is part of what is matched. Post-training the four-step FLUX.2 [klein] (Black Forest Labs, 2026) into a one-step model with this objective surpasses the four-step version on GenEval (Ghosh et al., 2023) while also surpassing its PickScore (Section 4.2); the marginal alternative sacrifices alignment with no compensating quality gain (Table 2).

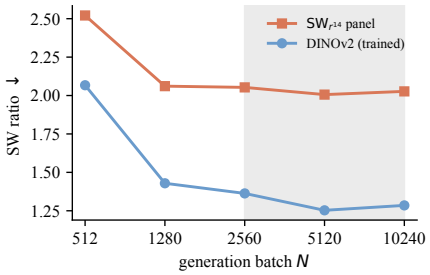

Figure 4: Generation batch size N at a matched wall-clock budget, fine-tuning a single-encoder DINOv2 Nyström-MMD arm; Quality climbs with N to a broad optimum (shaded).

3.3 The representation axis: one encoder is never enough

Feature distances are also how realism is scored: FID and its descendants (Heusel et al., 2017; Bińkowski et al., 2018; Jayasumana et al., 2024; Stein et al., 2023) reduce it to the distributional gap under one pretrained encoder, read as a proxy for human judgment. The proxy is fragile. FID falls under fringe ImageNet-class features with no gain in perceived quality (Kynkäänniemi et al., 2023), and such a distance is directly *optimizable*: a generator can be driven below the score of real validation data while staying visibly fake (Yang et al., 2026). The question this axis turns on: *is there any feature space whose distance, once minimized, yields images humans cannot tell from real?*

Overfitting a single encoder. Below-real scores have so far been shown only on weak proxies, Inception and ConvNeXt, inviting the objection that a sufficiently rich encoder, once satisfied, would force realism. We test the hardest case we can construct: DINOv2, far more semantically structured, on which the base checkpoint starts far from real, $\text{SW}_{\text{dino}} = 1.81$. Matching it alone, $N{=}5120$ for 1000 steps, drives the distance to 1.01, essentially the real-validation floor of 1.00: by DINOv2's account the generator is as close to real as real data. fig. 5 says otherwise. The objective repairs some classes, the lizard becomes hard to tell from a photograph, and leaves others untouched, the typewriter keeps an implausible key layout at that same floor score. The limitation is single-encoder matching itself, not the choice of encoder, and the resolution is not a better encoder but a diverse ensemble.

Constrained optimization against multiple encoders. A single encoder gives only a pseudo-metric, but the combined kernel of a diverse panel is characteristic and vanishes only at the real distribution (Gretton et al., 2012; Sriperumbudur et al., 2010; Schrab et al., 2023); we therefore train

Figure 5: Matching only DINOv2 features drives its distance to the real floor, $\mathrm{SW}_{\text{dino}}$=1.01, yet improves quality unevenly: the lizard (left) becomes indistinguishable from real, the typewriter (right) keeps clear artifacts. A saturated single-encoder distance does not imply realism.

against ten of the fourteen panel encoders (Appendix B), frozen backbones chosen to fail in different ways. The weighting then decides whether this diversity survives: under fixed weights the optimizer drives the aggregate down through whichever encoders are easiest. We instead pose the weighting as a constrained optimization, each encoder required to reach its real-validation floor with its weight the Lagrange multiplier, set by proportional control under a satisfaction gate, the proportional term of the PID-Lagrangian scheme of Stooke et al. (2020). An encoder's excess $e_\phi = s_\phi - b_\phi$ sets its weight: those at or below their floor drop out, while the violators share a fixed budget through a softmax, $\lambda_\phi \propto \exp\big(e_\phi/(\tau\,\bar{e})\big)$, so the representations farthest from real are weighted most; when all are satisfied the weights vanish, a natural anti-overfitting terminal state.

Scaling the multi-representation metric. Evaluation needs the same protection and must not collapse into the training loss. Yang et al. (2026) aggregate a per-encoder ratio over a panel, the Fréchet form FD_{r^k}; we keep that construction but replace the Fréchet distance with the Sliced-Wasserstein (Deshpande et al., 2018; Wu et al., 2019), a proper optimal-transport distance that shares no estimator with the MMD we train against (Berthet et al., 2026). Our metric $\mathrm{SW}_{r^{14}}$ averages the per-encoder ratio over the k encoders,

$$\mathrm{SW}_{r^k} \;=\; \frac{1}{k}\sum_{e=1}^{k} r_e, \qquad r_e \;=\; \frac{\mathrm{SW}\big(\phi_{e*}p_\theta,\ \phi_{e*}p_{\text{train}}\big)}{\mathrm{SW}\big(\phi_{e*}p_{\text{val}},\ \phi_{e*}p_{\text{train}}\big)}. \tag{5}$$

With $k = 14$, real validation data scores 1 by construction, a floor no released generator approaches (Table 1); four of the encoders are held out from training as a generalization check. Appendix D gives a kernel-MMD counterpart MMDr14, and Section 4.1 validates $\mathrm{SW}_{r^{14}}$ against PickScore.

Putting it together: iRDM. Together these choices define iRDM: an exact within-batch repulsion with a Nyström attraction to a reference frozen once over the full data, large fresh generation batches, the joint image-text law on conditional tasks, and a diverse encoder battery balanced by constrained optimization. The reference $\bar{\mu}_\phi$ is precomputed once per encoder; each step then draws a fresh batch, generates in a single evaluation, encodes it under the training encoders with gradient caching, and sums the per-encoder losses of eq. (3) under the proportional Lagrangian weights. Nothing else enters the objective: no online teacher, no adversary, no trajectory.

4 Experiments

Sections 3.1 to 3.3 fixed each design choice with a controlled study in place. The experiments report what remains: the main results, one-step ImageNet generation and text-to-image post-training, and the ablations the studies did not cover.

4.1 One-step ImageNet generation

Setup. On ImageNet-256 (Deng et al., 2009), we post-train the released pMF-H FD-SIM checkpoint (Lu et al., 2026; Yang et al., 2026) for 4000 steps at learning rate 1.6×10^{-6} and batch size N=5120 over the ten training encoders of Appendix B, each a Gaussian kernel at its median-heuristic bandwidth whose attraction is taken against the full 1.28M-image ImageNet training set, compressed once into a 4096-landmark Nyström reference. The ten encoders are kept in balance

Table 1: $\mathrm{SW}_{r^{14}}$, our primary metric, across released ImageNet-256 generators. Per-encoder floor-normalized SW ratio $\mathrm{SW}(\mathrm{gen},\mathrm{train})/\mathrm{SW}(\mathrm{val},\mathrm{train})$, the Sliced-Wasserstein; ≈ 1 matches a fresh real draw, lower = closer). SW is an optimal-transport distance sharing no machinery with the kernel MMD of the training loss, so it cannot be gamed by matching the loss. $\mathrm{SW}_{r^{14}}$ is the arithmetic mean over the 14 encoders (matching MMDR_{14}'s aggregate). Grey rows are one-step (single-NFE) models; * marks an external representation encoder in training. † marks the four encoders held out from training, namely DINOv2, SigLIP (v1), RADIO, and FLUX; $\mathrm{SW}^{\dagger}_{r^4}$ is the same floor-normalized mean restricted to those four, a generalization check.

Model	Inception	ConvNeXt	DINOv2†	MAE	SigLIP2	CLIP	DINOv3	SigLIP (v1)†	PE-Core	RADIO†	WebSSL	AIMv2	DreamSim	FLUX†	$\mathrm{SW}_{r^{14}}\downarrow$	$\mathrm{SW}^{\dagger}_{r^4}\downarrow$
Validation baseline	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Drifting-L*	0.97	1.61	6.12	3.20	8.84	5.83	18.2	7.12	7.89	5.31	5.70	8.45	2.69	1.07	5.93	4.91
iMF-XL	0.96	1.22	5.07	2.96	6.89	5.04	15.2	6.01	7.26	4.19	4.75	7.06	2.62	1.03	5.02	4.08
Open-MAGVIT2-L	1.57	1.39	4.87	2.93	6.33	4.94	6.59	5.33	5.95	4.50	4.66	7.05	2.70	1.41	4.30	4.03
SiT-XL/2	1.29	1.12	4.53	2.68	6.24	4.75	9.94	5.21	6.36	3.87	4.09	6.43	2.32	0.91	4.27	3.63
pMF-H (base)	1.25	0.88	3.91	3.15	6.43	3.93	6.62	4.88	6.69	4.00	4.25	6.87	2.63	1.71	4.09	3.63
DiT-XL/2	1.38	0.99	4.12	2.51	5.50	4.72	8.92	4.91	5.98	3.50	3.82	6.01	2.35	1.03	3.98	3.39
VAR-d30	1.08	1.12	4.18	3.18	5.71	4.51	6.37	5.19	6.52	3.77	3.88	6.28	2.63	0.88	3.95	3.51
JiT-H	1.07	1.46	3.73	2.78	5.52	6.17	5.34	4.84	6.59	3.85	3.75	6.11	2.81	1.13	3.94	3.39
MDTv2-XL/2	0.86	0.98	3.76	2.44	5.53	4.71	9.78	4.94	6.33	3.08	3.59	5.48	2.30	0.79	3.90	3.14
MAR-H	1.00	1.04	4.16	2.39	5.25	4.34	9.05	4.86	6.21	3.23	4.01	6.04	2.20	**0.48**	3.87	3.18
DDT-XL/2*	0.83	0.96	3.74	2.36	5.29	4.60	9.07	4.68	6.18	3.06	3.49	5.44	2.15	0.97	3.77	3.11
SiT-XL/2+REPA*	0.85	1.01	3.63	2.35	5.13	4.33	8.15	4.58	5.99	3.02	3.34	5.29	2.12	0.72	3.61	2.99
REG-XL*	0.79	0.91	3.13	2.03	4.68	3.92	7.09	4.02	5.74	2.60	2.88	4.65	1.83	0.71	3.21	2.62
LightningDiT-XL*	0.89	0.90	3.25	2.04	4.44	3.76	4.90	3.92	5.22	2.80	3.29	5.18	1.99	0.78	3.10	2.69
RAE-XL*	0.75	1.30	2.38	2.11	2.74	3.52	2.76	2.80	4.51	2.39	2.31	3.88	1.51	1.13	2.43	2.18
REPA-E SiT-XL/1*	0.75	1.00	2.79	1.86	3.41	2.83	3.30	3.07	3.89	2.04	2.41	4.13	1.47	0.66	2.40	2.14
pMF-H (FD-SIM)*	**0.67**	**0.67**	1.81	**0.60**	1.86	2.69	2.33	2.63	4.76	2.14	2.31	3.68	**1.24**	1.35	2.05	1.98
iRDM (ours)*	1.27	0.98	**1.35**	0.83	**1.30**	**1.02**	**1.11**	**1.90**	**1.22**	**1.56**	**1.55**	**1.44**	1.32	1.36	**1.30**	**1.54**

Figure 6: PickScore preference, iRDM (orange) against prior generators and a real-photo reference; each bar shows the win rate, mean PickScore below. The FD-SIM bar is matched-noise paired, the others per-class means. iRDM is preferred over every prior generator and is, to our knowledge, the first one-step model to also surpass the held-out real-photo reference. The PickScore ordering agrees with the $\mathrm{SW}_{r^{14}}$ ranking (Table 1), indicating that $\mathrm{SW}_{r^{14}}$ also reflects human preference.

by the proportional Lagrangian controller of Section 3.3 with a satisfaction gate over a fixed budget $\Sigma{=}10$, each encoder's real floor computed on the ImageNet validation set. Evaluation uses two off-objective measures. $\mathrm{SW}_{r^{14}}$ is the Sliced-Wasserstein ratio averaged over the 14-encoder panel, four encoders held out from training, estimated from 16384 samples per set with $M{=}1024$ projections. PickScore (Kirstain et al., 2023) is a learned human-preference model scored against the class prompt; against the pMF-H FD-SIM start we render 4000 class-conditional latents under matched noise with both models and report the paired mean and win rate.

Distributional quality. Table 1 places released ImageNet-256 generators on $\mathrm{SW}_{r^{14}}$; none approaches the real floor of 1, the strongest reaching about 2.05. iRDM sets the state of the art at $\mathrm{SW}_{r^{14}}$ 1.30, below every released generator, and is the best entry on nine of the fourteen encoders and on the aggregate. It cedes five: Inception, ConvNeXt, and MAE to the FD-loss model, which scores below real there by gaming a single space, DreamSim to that same model by a hair, and the held-out FLUX VAE to MAR-H. Appendix D reports the same field under MMDr14, a kernel-MMD panel, which broadly agrees, with some reordering among the mid-field models.

Human preference. An off-objective check agrees. PickScore (Kirstain et al., 2023), a learned human-preference model we never train against, prefers our converged checkpoint to its pMF-H FD-SIM start on 71.2% of matched pairs (20.61→20.96, paired $z{=}30.5$) and to the recent RAE-XL (Zheng et al., 2025) and REPA-E SiT-XL (Leng et al., 2025) on 75.7% and 73.2% of classes (Figure 6); it even prefers our samples to held-out real photographs on 63.6%, to our knowledge the first one-step generator to pass the real-image PickScore.

Table 2: **GenEval and PickScore for one-step FLUX.2 [klein] post-training.** Per-category GenEval and PickScore of the four-step FLUX.2 [klein], the untrained one-step start, a DMD2 (Yin et al., 2024a) baseline, the image-marginal ablation, and the joint one-step iRDM; best per column in bold. PickScore is scored on 500 COCO validation prompts, higher is better. The joint image-text objective lifts the overall GenEval from 0.801 (marginal) to 0.826, surpassing the four-step version overall.

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall	PickScore
FLUX.2 [klein] (4-step)	0.994	0.904	0.791	0.880	0.575	0.623	0.794	22.58
Untrained (1-step)	0.894	0.323	0.603	0.673	0.225	0.128	0.474	19.95
DMD2 (1-step)	**0.997**	0.894	**0.806**	0.864	0.603	0.660	0.804	22.36
iRDM (1-step, marginal)	0.991	0.899	0.763	0.910	0.638	0.608	0.801	22.70
iRDM (1-step)	0.994	**0.924**	0.756	**0.923**	**0.650**	**0.708**	**0.826**	**22.76**

4.2 TEXT-TO-IMAGE POST-TRAINING

Setup. We post-train FLUX.2 [klein] (Black Forest Labs, 2026) from its four-step checkpoint into a one-step model with the joint image-text objective of Section 3.2, at batch size $N{=}10240$ and learning rate 2.83×10^{-6} for 180 steps, about 90 H200 GPU-hours, under the encoder battery and constrained-optimization weighting of Section 4.1. The matching reference is collected once from the four-step teacher and then frozen, so the post-training queries no online teacher: a curated set of about 300K teacher generations, PickScore-ranked COCO renderings (Lin et al., 2014) together with detector-verified GenEval-correct samples, compressed once into a Nyström reference and detailed in Appendix E.1. We evaluate with GenEval (Ghosh et al., 2023) under its standard protocol and PickScore (Kirstain et al., 2023) on 500 COCO validation prompts, and compare against a DMD2 (Yin et al., 2024a) one-step distillation of the same four-step teacher (Appendix E.2).

Results. The one-step model surpasses its four-step start on GenEval overall, 0.826 against 0.794, with the per-category breakdown in Table 2: it matches the four-step version on single-object prompts, exceeds it on two-object, colors, position, and attribute binding, and trails only on counting. On PickScore it reaches 22.76, also above the four-step version's 22.58. The DMD2 baseline reaches 0.804 overall GenEval and 22.36 PickScore, also listed in Table 2; Figure 1(c) traces both metrics over post-training compute.

Joint versus marginal. The joint coupling carries the gain. A marginal variant that drops the caption from the feature, matching the image marginal alone with no SigLIP text concatenation, trails the joint model overall in Table 2, 0.801 against 0.826, and the gap concentrates on the categories that demand image-text alignment, two-object (0.924 against 0.899) and attribute binding (0.708 against 0.608), while single-object, which depends little on coupling, is essentially unchanged. Matching the joint law rather than the image marginal is what makes prompt fidelity part of the objective.

4.3 CONSTRAINED OPTIMIZATION VERSUS UNIFORM WEIGHTING

Setup. We isolate the gated proportional Lagrangian controller of Section 3.3 against uniform weighting: both warm-start from pMF-H and train for 100 steps under one recipe, with only the per-encoder allocation differing. The start is bimodal, the classic encoders already at or below their floor while the modern ones sit far from real, so the aggregate $\mathrm{SW}_{r^{14}}$ of 2.09 is set by the violators.

Table 3: Per-encoder weighting: gated proportional Lagrangian versus uniform, 100 steps from pMF-H on the $\mathrm{SW}_{r^{14}}$ panel (lower is better, real floor = 1). The gated controller edges uniform on the mean and clearly improves the worst encoder, the case the controller targets. Better arm in **bold**.

Aggregate	pMF-H	Gated	Uniform
$\mathrm{SW}_{r^{14}}$	2.09	**1.88**	1.90
max	4.83	**3.49**	4.06

Results. The gated controller pours the budget onto the worst encoder, PE-Core, while gating out the three already at their floor: it edges uniform on the mean, $\mathrm{SW}_{r^{14}}$ 1.88 against 1.90, while decisively improving the worst case, 3.49 against 4.06 from a start of 4.83, nearly twice the cut uniform

Table 4: Training-distance ablation on DINOv2 (cls). The six fine-tuning losses warm-start the same pMF-H checkpoint and fine-tune against a single DINOv2 encoder, flipping only the per-step distance; `baseline` is pMF-H at step 0. Each entry is a floor-normalized ratio (lower = closer to real, $\approx$ 1 matches a fresh real draw) under two neutral distances, a Sliced-Wasserstein ratio (the per-encoder analogue of $\mathrm{SW}_{r^{14}}$) and an RFF-MMD ratio (that of MMDr14). The order `mmdx` $\succ$ `mmd_rff` $\succ$ `mmd_exact` $\succ$ `fd` $\succ$ `sw` $\succ$ `drifting` is identical on both; exact MMD does not beat Nyström, and the SW-trained arm does not win the SW eval. `drifting` is a faithful port of the drifting force field shown at the best of a learning-rate sweep, its native rate regressing the warm-start.

DINOv2 cls ratio (↓)	baseline	mmdx	mmd_rff	mmd_exact	fd	sw	drifting
SW	1.927	**1.420**	1.466	1.492	1.547	1.583	1.746
RFF-MMD	10.393	**4.495**	4.839	5.438	5.798	6.413	8.258

manages (Table 3). Reallocating toward the largest violation is what the controller targets; we make no claim here about perceptual quality, which this aggregate does not directly measure.

4.4 The training distance

Setup. Holding the rest of the recipe fixed, we flip only the per-step distance across six fine-tuning losses, each warm-started from the same pMF-H checkpoint and fine-tuned against a single DINOv2 cls encoder for 100 optimizer steps, sharing AdamW at learning rate 1.6×10^{-6} and a generation batch of 5120. The three kernel arms use an RBF kernel at bandwidth σ=65: `mmdx` is the biased MMD^2 with an exact within-batch term and a Nyström cross-term over 4096 landmarks, `mmd_exact` replaces that cross-term with the exact full generated-to-real pairwise mean, and `mmd_rff` matches a frozen 4096-dimensional random-Fourier-feature mean; `fd` is the Fréchet (Gaussian-moment) distance and `sw` a Sliced-Wasserstein loss with 128 projections. The `drifting` arm is a faithful port of the published coupled force field across radii $\{0.2, 0.05, 0.02\}$, with in-batch generated negatives and real positives on the same features, run time-matched to the other arms (about 60 steps at its generation batch of 8192); we sweep its learning rate and report the gentlest, 1×10^{-6}, since its native 4×10^{-4}, tuned for from-scratch training, regresses the warm-start. We score every arm and the untrained baseline with two neutral third-party distances on the same features, a Sliced-Wasserstein ratio from 16384 samples per set with M=1024 projections and an RFF-MMD ratio from 50000 samples with 4096 random Fourier features, so each arm is read on at least one distance it did not train on; the `sw` and `mmd_rff` arms, which each optimize one of the two eval distances, are judged on the other (table 4).

Results. One ranking holds across both, `mmdx` $\succ$ `mmd_rff` $\succ$ `mmd_exact` $\succ$ `fd` $\succ$ `sw` $\succ$ `drifting`, well above the untrained baseline: the three kernel-MMD estimators fill the top, moment matching follows, the Sliced-Wasserstein loss next, and a faithful port of the drifting force field is weakest even at the best of a learning-rate sweep. As an objective the Nyström MMD moves the feature distribution closest to real, while optimal transport, an excellent judge, is among the least effective losses. Two controls confirm the reading: the exact full-pairwise MMD does not beat its Nyström approximation, the low-rank cross-term being a smoother gradient, and training on a distance buys no advantage on that same distance, the Sliced-Wasserstein arm being beaten on the Sliced-Wasserstein eval by every kernel-MMD arm and even by moment matching. This is why iRDM trains with the MMD-Nyström signal yet is evaluated with the independent Sliced-Wasserstein distance.

5 Conclusion

We have treated representation distribution matching, the principle behind a recent line of teacher-free one-step generators, as a design space rather than a collection of methods. Two axes fix every instance, how the generated and real feature distributions are compared and which representations they are compared in, and varying one at a time turns each into a preferred design with a mechanism behind it. On the comparison axis the classical MMD becomes a strong objective once estimated

right, an exact within-batch repulsion paired with a Nyström attraction toward a reference frozen once in advance, fed by large fresh generation batches and, on conditional tasks, by matching the joint image-text law rather than the image marginal. On the representation axis no single encoder is enough, since any one can be driven below the real score while samples stay visibly fake, so we match against a diverse battery of encoders held in balance by constrained optimization. Combining these choices gives iRDM, which sets the one-step state of the art on ImageNet at $\mathrm{SW}_{r^{14}}$ 1.30 and post-trains the four-step FLUX.2 [klein] into a one-step model that surpasses it on GenEval and PickScore, and we report the remaining gap with $\mathrm{SW}_{r^{14}}$, a Sliced-Wasserstein distance over the panel that shares no machinery with the training loss.

A gap to real remains: at $\mathrm{SW}_{r^{14}}$ 1.30 against a floor of 1, the best one-step generator is still measurably short of a fresh real draw, and narrowing it is the natural next target. The design space leaves room to do so, through multi-scale kernels, learned or task-specific encoder panels, and richer conditional couplings, and the same recipe, a frozen reference matched by a single network evaluation, should transfer to modalities beyond images wherever a pretrained encoder supplies the feature space.

ACKNOWLEDGMENTS

We thank Jiawei Yang for helpful discussions. Additional acknowledgments will be added in the camera-ready version.

REFERENCES

Quentin Berthet, Yu-Han Wu, Clément Crepy, Romuald Elie, Klaus Greff, and Michael Eli Sander. MIND: monge inception distance for generative models evaluation. *CoRR*, abs/2605.06797, 2026.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Black Forest Labs. FLUX.1. `https://github.com/black-forest-labs/flux`, 2024. Official inference repository for FLUX.1 models.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. `https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence`, 2026. Model weights: `https://huggingface.co/collections/black-forest-labs/flux2`.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, 2025.

Antoine Chatalic, Nicolas Schreuder, Lorenzo Rosasco, and Alessandro Rudi. Nyström kernel mean embeddings. In *International Conference on Machine Learning*, 2022.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

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

Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 258–267, 2015.

Ali Falahati, Elliot Creager, Gautam Kamath, and Shubhankar Mohapatra. DriftXpress: Faster drifting models via projected RKHS fields, 2026. arXiv preprint arXiv:2605.12183.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, and Saining Xie. Scaling language-free visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander T. Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, 2023.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2023a.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.repl4nlp-1.31. URL `https://aclanthology.org/2021.repl4nlp-1.31/`.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. MDTv2: Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023b.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, 2025a.

Zhengyang Geng, Ashwini Pokle, Weijian Luo, Justin Lin, and Zico Kolter. Consistency models made easy. In *International Conference on Learning Representations*, 2025b.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J. Candès. One-step generative modeling via Wasserstein gradient flows, 2026. arXiv preprint arXiv:2605.11755.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. RADIOv2.5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for GAN training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of ImageNet classes in Fréchet inception distance. In *International Conference on Learning Representations*, 2023.

Eric C. Larson and Damon M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2017.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise, 2025. arXiv preprint arXiv:2511.13720.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems*, 2024.

Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, 2015.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, 2025.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows, 2026. arXiv preprint arXiv:2601.22158.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-Instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-MAGVIT2: An open-source project toward democratizing auto-regressive visual generation, 2024. arXiv preprint arXiv:2409.04410.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, 2022.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, 2016.

Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. In *Advances in Neural Information Processing Systems*, 2024.

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

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, 2024.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, 2022.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In *International Conference on Machine Learning*, 2020.

Marilyn Strathern. 'improving ratings': Audit in the British University system. *European Review*, 5(3):305–321, 1997.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zhou Wang and Xinli Shang. Spatial pooling strategies for perceptual image quality assessment. In *IEEE International Conference on Image Processing*, pages 2945–2948, 2006.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier than you think. In *Advances in Neural Information Processing Systems*, 2025.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced Wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

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

Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems*, 2012.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. In *International Conference on Machine Learning*, 2025.

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 *International Conference on Machine Learning*, 2024.

A Extended related work

Scalable kernel estimators. Random Fourier features linearize kernel sums with data-independent bases (Rahimi and Recht, 2007); Nyström methods use data-dependent landmarks instead and dominate whenever the kernel spectrum decays quickly (Yang et al., 2012). For the kernel mean embedding, the single point in the RKHS that summarizes a distribution (Muandet et al., 2017), Nyström compression retains the full estimation rate of the exact embedding with far fewer landmarks than data points (Chatalic et al., 2022). Concurrently, DriftXpress accelerates drifting models by projecting the kernel field onto a low-rank RKHS with landmark approximations (Falahati et al., 2026). Our use differs in target and in symmetry: DriftXpress approximates the full per-batch drifting update for speed, while we compress only the stationary data side, into a frozen global attraction target over 1.28M images, to remove reference noise, and keep the moving within-batch repulsion exact.

Metric gaming and multi-encoder evaluation. Single-encoder distances such as FID (Heusel et al., 2017), KID (Bińkowski et al., 2018), CMMD (Jayasumana et al., 2024), and feature-space precision and recall (Kynkäänniemi et al., 2019) inherit the blind spots of their encoder: the scores move with resizing details (Parmar et al., 2022), fall through fringe ImageNet-class features with no quality gain (Kynkäänniemi et al., 2023), and re-rank models when the encoder is swapped (Stein et al., 2023). Yang et al. (2026) push this to the limit, driving a trained generator below the score of the real validation set; we show the failure is single-encoder matching itself rather than any weak encoder. When the proxy is learned, the same phenomenon is studied as reward hacking (Strathern, 1997; Skalse et al., 2022; Gao et al., 2023a), ensembling the proxies mitigates it (Coste et al., 2024), and Lagrangian methods control it in constrained reinforcement learning (Stooke et al., 2020). In our setting every proxy is frozen, so gaming pressure concentrates on the encoder weighting, which is exactly what the proportional Lagrangian controller regulates; evaluation aggregates 14 encoders, 4 held out from training, into $\mathrm{SW}_{r^{14}}$, a Sliced-Wasserstein distance independent of the loss.

Post-training text-to-image models. Few-step text-to-image systems are typically distilled from a multi-step teacher, adversarially (Sauer et al., 2024) or by score-based distribution matching (Yin et al., 2024a), and then steered toward human taste by optimizing learned preference rewards. The reward models are trained from human choices (Xu et al., 2023; Kirstain et al., 2023) and optimized either by policy gradients (Black et al., 2024) or by direct preference objectives (Wallace et al., 2024); all inherit the gameability of a single learned scorer, the same axis our multi-encoder battery addresses for frozen proxies. For our text-to-image result the four-step FLUX.2 [klein] (Black Forest Labs, 2026) generates a reference set in advance, and the joint image-text objective of Section 3.2 matches the one-step model against it with no online teacher and no reward model, evaluated by GenEval (Ghosh et al., 2023) and PickScore.

Table 5: The fourteen-encoder panel. Each backbone is frozen at its released weights and $\phi(x)$ is its pooled image embedding, taken at the listed input resolution. Pool: CLS class token, AVG mean over patch or spatial tokens, ATTN attention-pooling head. Ten encoders supervise training; four are held out for evaluation only.

Encoder	Checkpoint	Architecture	Input	Pool	D
Training panel (10)					
Inception-v3 (Szegedy et al., 2016)	FID Inception-v3	CNN	299	avg	2048
ConvNeXt V2-B (Liu et al., 2022)	`convnextv2_base.fcmae_ft_in22k_in1k`	CNN	224	avg	1024
MAE (He et al., 2022)	`vit_large_patch16_224.mae`	ViT-L/16	224	avg	1024
CLIP (Radford et al., 2021)	`vit_large_patch14_clip_224.openai`	ViT-L/14	256	cls	1024
DINOv3-L (Siméoni et al., 2026)	`vit_large_patch16_dinov3.lvd1689m`	ViT-L/16	224	cls	1024
PE-Core-L (Bolya et al., 2025)	`vit_pe_core_large_patch14_336.fb`	ViT-L/14	224	attn	1024
SigLIP2-So400m (Tschannen et al., 2025)	`vit_so400m_patch16_siglip_256.v2_webli`	ViT-So400m/16	224	attn	1152
AIMv2-H (Fini et al., 2025)	`aimv2_huge_patch14_224.apple_pt`	ViT-H/14	224	avg	1536
Web-SSL DINO 1B (Fan et al., 2025)	`webssl-dino1b-full2b-224`	ViT-1B	224	cls	1536
DreamSim (Fu et al., 2023)	DINO + CLIP + OpenCLIP ensemble	ViT ens.	224	cls	1792
Held out (4)					
DINOv2 (Oquab et al., 2024)	`vit_large_patch14_dinov2.lvd142m`	ViT-L/14	256	cls	1024
SigLIP (v1) (Zhai et al., 2023)	`vit_so400m_patch14_siglip_384.webli`	ViT-So400m/14	384	attn	1152
C-RADIOv3-L (Ranzinger et al., 2024; Heinrich et al., 2025)	NVIDIA C-RADIOv3-L	ViT-L, multi-teacher	256	summary	3072
FLUX VAE (Black Forest Labs, 2024)	FLUX.1 VAE, 4×4 patch-mean	VAE	256	patch-mean	1024

Table 6: Generation batch size N at a matched wall-clock budget ($\approx$ 6000 s each), fine-tuning a single-encoder DINOv2 Nyström-MMD arm; entries are Sliced-Wasserstein ratios (lower is closer to real). The smallest batch regresses above the untrained base despite the most optimizer steps; the optimum is broad, with N=10240 only marginally worse than N=5120.

Batch N	lr	DINOv2 ↓	$\mathrm{SW}_{r^{14}}$ ↓
untrained base	n/a	1.927	2.085
512	5.1×10^{-7}	2.067	2.521
1280	8.0×10^{-7}	1.429	2.061
2560	1.1×10^{-6}	1.363	2.053
5120	1.6×10^{-6}	**1.253**	**2.006**
10240	2.3×10^{-6}	1.285	2.027

The evaluation landscape. The generators placed on $\mathrm{SW}_{r^{14}}$ in Table 1 span the current families: latent diffusion transformers (Peebles and Xie, 2023; Ma et al., 2024; Gao et al., 2023b; Wang et al., 2026; Yao et al., 2025), representation-aligned variants that inject external encoder features during training (Yu et al., 2025; Wu et al., 2025; Zheng et al., 2025), autoregressive and masked-token models (Tian et al., 2024; Li et al., 2024; Luo et al., 2024), pixel-space transformers (Li and He, 2025), and the one-step MeanFlow, drifting, and FD-loss lines (Lu et al., 2026; Geng et al., 2026; Deng et al., 2026; Yang et al., 2026). The models that use an external representation encoder in training, the starred rows of the table, populate its strongest entries, consistent with the premise that representation supervision is the operative ingredient; RDM makes that ingredient explicit and studies it in isolation.

B ENCODER PANEL

MMDr14 takes the arithmetic mean over the 14 encoders of table 5, each frozen at its released weights and read out as a single pooled image embedding $\phi(x)$ at the listed input resolution, with no feature normalization. The panel deliberately spans training paradigms, supervised classification, self-supervised distillation and masked reconstruction, language supervision, multi-teacher agglomeration, multimodal autoregression, human similarity tuning, and a generative autoencoder, so the representations fail in different ways. Ten supervise training; the four held out for evaluation only are DINOv2, SigLIP (v1), C-RADIOv3-L, and the FLUX VAE.

C BATCH-SIZE SWEEP

Table 6 tabulates the sweep plotted in fig. 4.

Table 7: MMD-RFF distance ratio (MMDR; lower = closer to real) of released ImageNet-256 generators and our iRDM across 14 vision encoders. $\overline{\text{MMDR}}_{14}$ is the arithmetic mean over the 14 encoders. The *validation baseline* is MMDR = 1 by definition (real held-out data); parentheses give the raw $\text{mmd}^2(\text{val}, \text{train}) \times 10^3$ normaliser. Grey rows are one-step (single-NFE) models. * marks an external representation encoder in training (REPA/RAE-style alignment, FD-loss, or drift-loss on encoder features). Strongest at bottom.

Model	Inception	ConvNeXt	DINOv2	MAE	SigLIP2	CLIP	DINOv3	SigLIP	PE-Core	RADIO	WebSSL	AIMv2	DreamSim	FLUX	$\overline{\text{MMDR}}_{14}$ ↓
Validation baseline	1.00 (0.321)	1.00 (0.535)	1.00 (0.0455)	1.00 (0.787)	1.00 (0.103)	1.00 (0.600)	1.00 (0.0805)	1.00 (0.420)	1.00 (0.565)	1.00 (0.156)	1.00 (0.0363)	1.00 (0.181)	1.00 (0.209)	1.00 (0.346)	1.00 (0.313)
Drifting-L* (Deng et al., 2026)	0.80	3.61	136	21.1	157	53.0	845	64.2	92.0	52.5	133	128	11.8	3.98	122
iMF-XL (Geng et al., 2026)	0.87	2.08	91.0	17.7	98.1	40.2	594	46.4	79.0	35.0	92.9	93.1	10.8	3.20	86.1
SiT-XL/2 (Ma et al., 2024)	1.73	1.77	75.3	14.3	79.0	35.1	258	35.3	60.9	29.5	68.0	76.7	7.56	2.08	53.2
Open-MAGVIT2-L (Luo et al., 2024)	2.79	2.72	85.9	16.5	84.0	36.1	114	37.6	52.9	38.4	90.6	96.0	11.3	5.55	48.2
MDTv2-XL/2 (Gao et al., 2023b)	0.63	1.23	50.0	11.8	60.5	34.8	254	31.9	59.9	19.0	51.3	56.8	7.82	1.69	45.8
MAR-H (Li et al., 2024)	0.79	1.19	61.5	11.0	56.5	28.7	219	30.1	57.4	20.7	65.0	68.1	7.18	0.37	44.8
DiT-XL/2 (Peebles and Xie, 2023)	2.11	1.35	59.9	12.7	62.2	34.5	204	31.9	53.4	24.1	57.4	67.9	8.00	1.96	44.4
pMF-H (base) (Lu et al., 2026)	1.78	0.91	54.6	17.8	87.5	22.4	115	30.5	65.7	31.5	70.6	94.4	10.9	8.91	43.7
DDT-XL/2* (Wang et al., 2026)	0.46	1.20	49.6	11.1	57.1	33.0	213	28.5	57.2	18.3	48.8	55.2	6.61	1.55	41.5
VAR-d30 (Tian et al., 2024)	1.13	1.73	63.8	18.7	67.4	30.3	108	34.7	61.1	29.7	62.4	75.9	11.4	1.34	40.6
JiT-H (Li and He, 2025)	1.26	3.06	48.1	14.2	66.0	51.1	74.4	31.6	64.1	30.4	60.7	73.3	12.3	2.56	38.1
SiT-XL/2+REPA* (Yu et al., 2025)	0.56	1.37	47.2	11.1	55.4	29.3	171	27.7	54.1	18.4	45.3	52.5	6.40	1.37	37.3
REG-XL* (Wu et al., 2025)	0.44	1.06	33.9	8.02	46.4	24.0	127	21.3	49.5	13.3	31.8	39.2	4.74	1.44	28.7
LightningDiT-XL* (Yao et al., 2025)	0.67	0.92	36.8	8.12	41.5	21.3	61.7	19.9	39.9	14.1	41.6	50.3	5.93	1.63	24.6
REPA-E SiT-XL/1* (Leng et al., 2025)	0.34	1.26	26.1	6.19	24.0	11.9	26.9	12.9	21.9	7.97	21.6	31.8	2.64	0.87	14.0
RAE-XL* (Zheng et al., 2025)	0.36	2.34	19.0	7.36	14.9	16.6	18.4	10.4	28.8	11.1	18.5	26.7	3.32	4.12	13.0
pMF-H (FD-SIM)* (Yang et al., 2026)	0.22	0.36	10.3	0.37	6.34	10.5	12.5	9.39	33.2	8.72	18.5	24.5	2.28	6.56	10.3
iRDM (ours)*	1.54	0.98	3.52	0.69	4.76	1.17	1.39	2.69	1.62	3.16	5.31	3.12	1.80	5.83	**2.69**

D KERNEL-MMD EVALUATION

Table 7 reports MMDr14, the training-aligned kernel-MMD cross-check of our primary $\text{SW}_{r^{14}}$ metric, over the full released field on the 14-encoder panel. Each entry is a per-encoder RFF-MMD ratio against real training data, with real validation scoring 1 by construction, and MMDr14 is their arithmetic mean over the 14 encoders. The ordering broadly agrees with $\text{SW}_{r^{14}}$ (Table 1), with some reordering among the mid-field models; because the loss is itself a kernel MMD, a single encoder can be pushed below the real floor here, Inception at 0.22 for the FD-SIM model, which the optimal-transport $\text{SW}_{r^{14}}$ and the held-out split resist.

E TEXT-TO-IMAGE POST-TRAINING DETAILS

E.1 REFERENCE CURATION

The text-to-image objective of Section 4.2 matches the one-step model against a reference collected once from the four-step FLUX.2 [klein] teacher and then frozen, so the teacher is never queried during post-training. The reference concatenates two independently curated blocks of teacher generations, a perception block and a composition block, roughly 300K image-caption pairs in all, each image kept with the caption that produced it for the joint kernel.

Perception block. For each of the 82,783 COCO train2014 images (Lin et al., 2014), one caption apiece, the four-step teacher draws 24 candidates, which PickScore (Kirstain et al., 2023) ranks; we keep the three highest per caption, giving 248,349 pairs at full coverage. Selecting three of twenty-four oversampled draws anchors the reference on high-quality renderings of natural captions, supplying the perceptual side of the match.

Composition block. To pull the model toward verified-correct composition rather than the teacher's average, which fails a large fraction of the harder compositional prompts, we keep only teacher generations a detector certifies as correct. For the 553 GenEval (Ghosh et al., 2023) prompts the teacher is sampled at 150 seeds per prompt, topped up where a prompt has fewer than 100 correct, and every generation is scored by the standard GenEval Mask2Former detector: a sample passes only when all prompt objects are present at the required count, color, and spatial relation. Capping at 100 correct per prompt yields 53,800 verified images covering 551 of the 553 prompts; two position prompts admit no correct teacher sample even at 1000 seeds. Measured per seed over this pool, the

Table 8: DMD2 (Yin et al., 2024a) one-step student over distillation steps, best LAION-prompt configuration, against the four-step FLUX.2 [klein] teacher. GenEval under the standard protocol and PickScore on the 500 COCO validation prompts; the 500-step peak is the baseline reported in Table 2.

Step	GenEval	PickScore
Teacher (4-step)	0.794	22.58
250	0.778	22.17
500 (reported)	**0.804**	22.36
750	0.792	22.36
1000	0.793	22.27

teacher's correctness ranges from 98% on single-object prompts down to 40% on attribute binding and 33% on position, so the filter most reshapes exactly the binding and spatial prompts on which the one-step model later improves.

Joint reference and prompt pool. The two blocks are embedded under the ten training encoders of Appendix B, each image feature concatenated with its caption's frozen SigLIP2 text embedding (Tschannen et al., 2025) and compared under one Gaussian kernel at 0.25 of the median-heuristic bandwidth with the text component weighted at 1; the stack is compressed once into an 8192-landmark Nyström reference per encoder. The generator's conditioning pool mirrors the reference: the $82{,}783$ COCO captions together with the GenEval prompts replicated so the GenEval share of the pool matches that of the reference, about 18%. Aligning the generated and reference prompt distributions keeps the match well-posed, its optimum reached when the two coincide.

E.2 DMD2 BASELINE

The DMD2 (Yin et al., 2024a) baseline distills the same four-step FLUX.2 [klein] teacher into a one-step student. The released DMD2 targets an SD-UNet under ϵ-prediction; we re-implement it for klein's flow-matching parameterization with the method intact: three networks, the one-step generator under training, a trainable critic estimating the student distribution's score, and the frozen four-step teacher, the DMD gradient being the teacher-minus-critic score difference, the critic updated every step and the generator every fifth. The student is initialized by regressing the teacher to a single step along its sampling ODE, after which distillation proceeds. We train at a global batch of 128 at 512^2 on four H200 GPUs with AdamW and no EMA.

Timestep schedule. The one change klein's parameterization forces is the distillation timestep distribution. klein is guidance-distilled and its velocity is faithful only at high noise, the four native sampling nodes; drawing the timestep uniformly reaches a low-noise regime where the teacher collapses to the mode-averaged posterior mean and drives the generator below its initialization. Mapping the uniform draw through klein's signal-to-noise shift with the empirical $\mu \approx 2.03$ that matches the native node spacing, with a learning rate of 5×10^{-7} and a short warmup, turns this divergence into a student that exceeds the teacher on GenEval.

Configuration and result. Distillation quality is set mainly by the prompt distribution. Training on a broader LAION caption pool rather than COCO captions alone lifts GenEval at every checkpoint and softens the peak-then-erode profile of distribution-matching distillation, a 4.7-point collapse becoming a 1.1-point plateau. The reported baseline is the best configuration, the LAION-prompt run at its 500-step peak, GenEval 0.804 and PickScore 22.36 (Table 8); a variant adding a GAN term on real latents was metric-neutral, its discriminator separating real from generated latents so fast that the adversarial gradient vanished against the distribution-matching one, and is not reported. The student peaks in about 10 H200 GPU-hours.

F ONE-STEP TEXT-TO-IMAGE SAMPLES

Figure 7 shows additional single-step iRDM generations from the post-trained four-step FLUX.2 [klein], each a 512×512 image produced in one network evaluation.

G QUALITATIVE COMPARISON

Figure 8 places uncurated iRDM samples beside pMF-H FD-SIM (Yang et al., 2026), the strongest external one-step baseline by $\mathrm{SW}_{r^{14}}$, across five ImageNet-256 classes spanning a bird, an animal coat, a deformable garment, a rigid man-made object, and a natural landscape. Within each method the enlarged image is one sample and the adjacent 2×5 grid holds ten further draws under the same class label, taken as the first released draws without cherry-picking. Per sample the two models are hard to separate by eye, both reaching sharp, on-class images; this is precisely why a distributional metric is needed, as the $\mathrm{SW}_{r^{14}}$ separation of 1.30 against 2.05 in Table 1 is not visible in any single row.

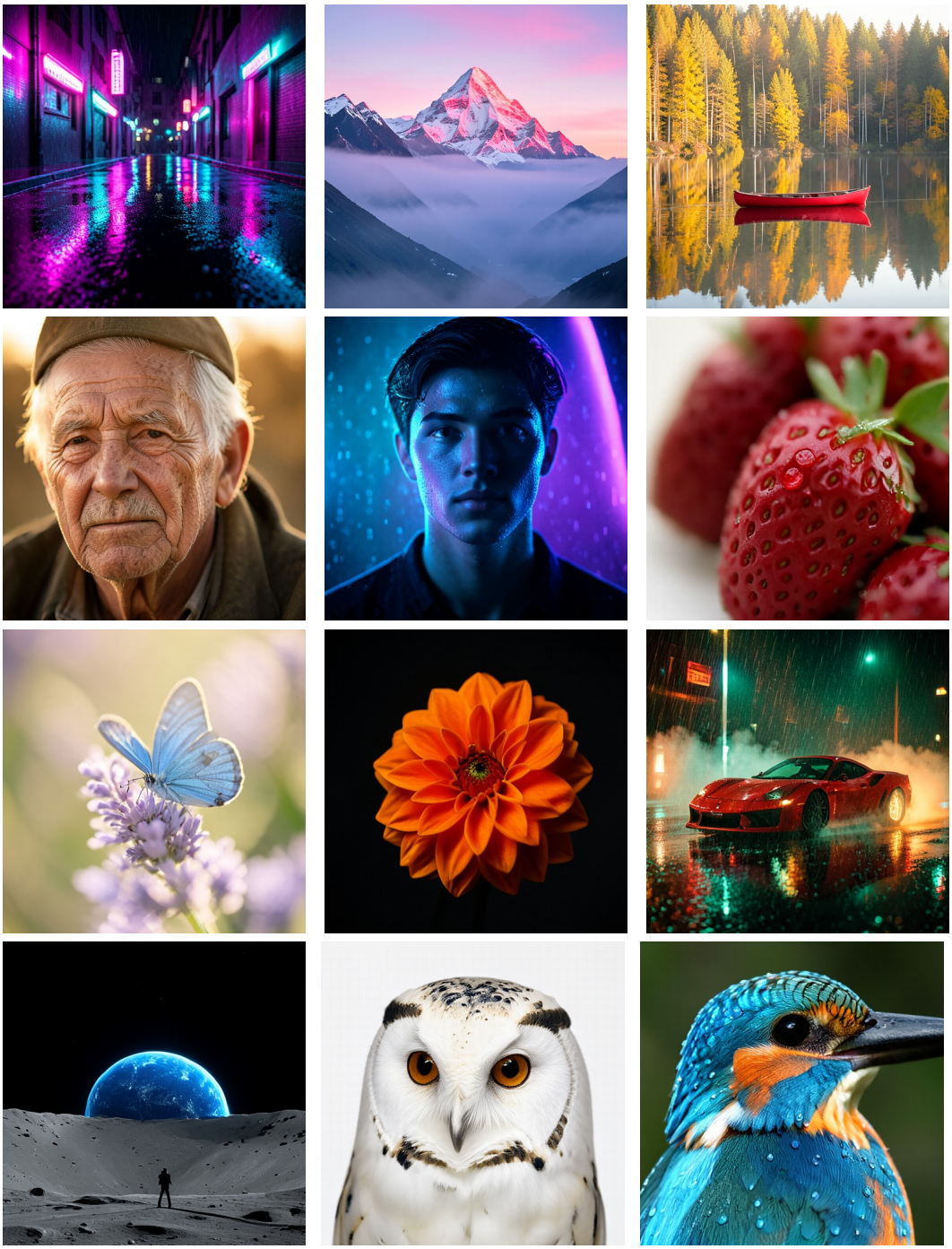

Figure 7: **One-step text-to-image samples from iRDM.** Single-step generations from the post-trained one-step FLUX.2 [klein], 512×512, one network evaluation each.

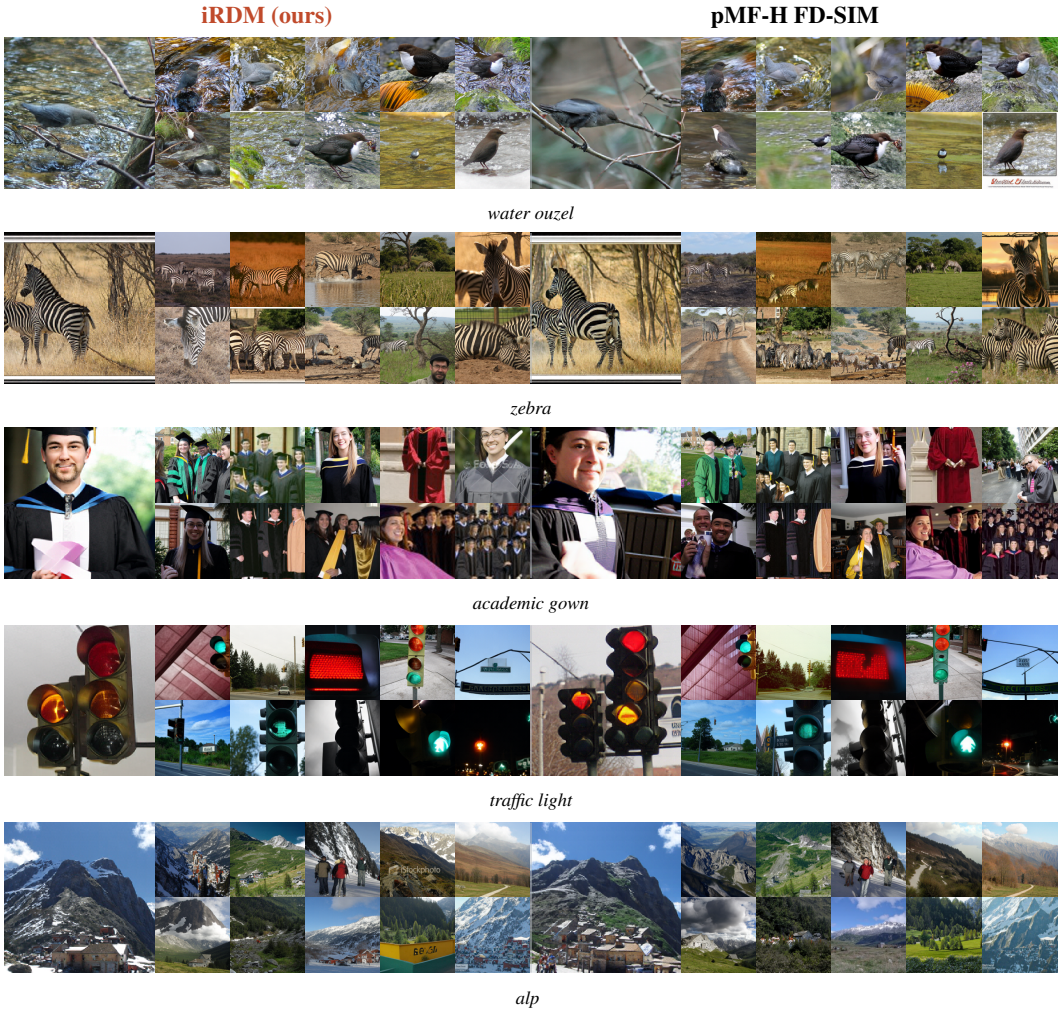

Figure 8: **Uncurated one-step samples** from iRDM and pMF-H FD-SIM (Yang et al., 2026) on five ImageNet-256 classes; column headers name the method. The two are close by eye despite the $\mathrm{SW}_{r^{1.4}}$ gap of 1.30 against 2.05 in Table 1.