LANDMARK-ANCHORED RECTIFIED FLOW: REDUCING VELOCITY-REGRESSION VARIANCE WITH TRAINING-FREE CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Rectified flow training regresses a velocity field onto its targets, and for small t the irreducible error of this regression approaches the full variance of the training distribution. For any partition of the training set, a measurable share of this variance, 16–63% on the datasets we study, is inter-group, and it encodes only which region of the data a trajectory targets. This share can be computed from the dataset, so the amount by which conditioning lowers the irreducible error is known before training. We propose Landmark-anchored Rectified Flow (LRF), which conditions the flow on a per-group landmark, a training image that represents the group, to lower the irreducible error by its inter-group component. Selecting the landmark requires no second generative model, and the uniform draw over groups makes the trained model an explicit mixture over the landmarks. At matched parameters and training budget, LRF improves FID over a plain rectified flow by 16.3% on ImageNet-256 and 24.4% on CIFAR-10, and reaches the final FID of the baseline with 51–54% fewer training iterations. The explicit mixture gives each sample a known conditioning landmark, and groups can be reweighted without retraining and examined for memorization against held-out data. Code will be released upon acceptance.

1 INTRODUCTION

Diffusion models and rectified flows (Ho et al., 2020; Song et al., 2020; Liu et al., 2022; Lipman et al., 2022) produce samples of high quality but have two practical costs. First, high sample fidelity requires long training schedules and substantial computational budgets, even at modest resolutions. Second, the trained model does not identify the training data a sample depends on, and its output distribution cannot be changed without retraining.

In this work, we trace the first cost to the training objective. For small t, the interpolant carries little information about the data, and the irreducible error of the velocity regression is high (Figure 1 and Section 3.1). We show that, for any partition of the training set into groups, this irreducible error contains an inter-group component. This component accounts for 16–63% of the total variance on the datasets we study, and it depends only on which group a sample belongs to. It can therefore be supplied to the network as input.

Several methods (Casanova et al., 2021; Blattmann et al., 2022; Sheynin et al., 2022; Li et al., 2024) condition generation on training instances or on representations of them, so that part of the data structure enters the network as input rather than being learned (Section 2). In each case the conditioning signal is produced at generation time by a second learned model or by a heuristic. However, these methods do not quantify, before training, how much their conditioning signal lowers the irreducible error. Adaloglou et al. (2025) condition diffusion models on the index of a cluster of pretrained features, which identifies a group of training images but contains none of their content (Table 1).

Our method, Landmark-anchored Rectified Flow (LRF), takes the conditioning signal directly from the training set. We partition the training set into groups of similar items and represent each group by one of its members, which we call its landmark. A rectified flow shared across groups then models the variation within groups given their landmarks, and at generation time a landmark is drawn

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.

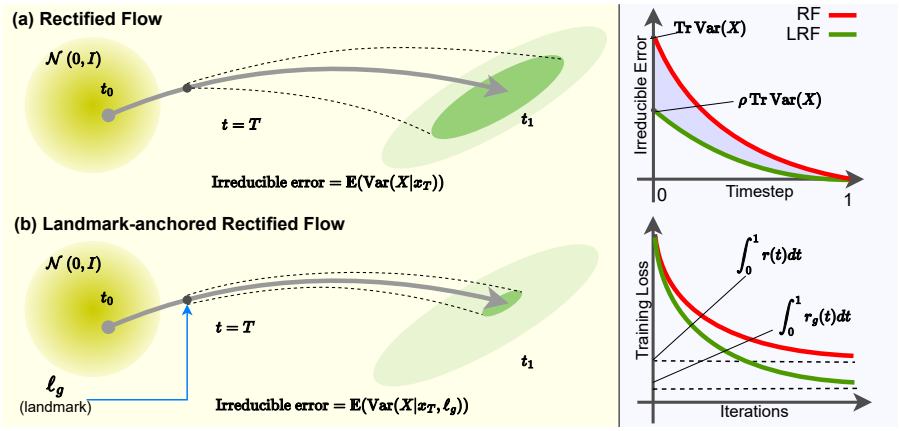

Figure 1: Irreducible error of the velocity regression with and without landmark conditioning. (a) Given the state x_T, the target may lie anywhere in the shaded region. (b) The landmark ℓ_g restricts the target to its group. Right: a schematic of the irreducible error along the trajectory and of the training loss along iterations in each case.

uniformly (Figure 2). Unlike the methods above, the conditioning signal is one stored training image per disjoint group. Selecting it therefore involves no second generative model, retrieval procedure, or learned per-group parameters, and the grouping uses pretrained image features (Section 3.5). Section 3.2 quantifies how much LRF lowers the irreducible error. At matched parameters and training budget, LRF attains 16.3% lower FID than a plain rectified flow on ImageNet-256, 24.4% on CIFAR-10, 2.4% on CelebA-256, and 5.7 against 18.1 on MNIST. It also reaches the final FID of the baseline with 51–54% fewer training iterations on CIFAR-10 and CelebA. Moreover, the share of the irreducible error at $t = 0$ that a partition accounts for is computed from the grouped data before training (Section 3.4), and Section 5.1 compares that share with the realized gain.

The construction also gives the trained model an explicit form, which is a uniform mixture over the landmarks. Each sample is therefore generated from one landmark, the components of the mixture can be enumerated and examined for memorization against held-out data, and a component can be reweighted without retraining.

Contributions.

1. We propose Landmark-anchored Rectified Flow, which reduces the variance of the velocity regression by conditioning the flow on the landmark of the target's group, where the landmark is a training image that represents the group.
2. We show that conditioning rectified flow on a partition of the training set lowers the irreducible error of the velocity regression by its inter-group component, which can be measured from the grouped data before training.
3. We conduct a controlled comparison of four conditioning signals under matched architecture, data, and training budget, and show that LRF reduces clean-FID by 16.3% on ImageNet-256 relative to a plain rectified flow and matches a learned group embedding. However, LRF achieves this without learning any per-group parameters.
4. LRF enables reweighting groups at generation time and identifying the training image each sample is conditioned on without retraining, and it allows the memorization of each group to be measured separately.

2 Related Work

Iterative generative models. Diffusion models (Ho et al., 2020; Song et al., 2020; Karras et al., 2022) and their flow-based successors, rectified flow (Liu et al., 2022) and flow matching (Lipman et al., 2022), train a network to transport noise to data along simple reference paths, and the framework remains under active development (Lee et al., 2024; Zhang et al., 2025). LRF follows this

Table 1: Generative models conditioned on training data, compared by conditioning signal and sampler.

Method	Conditioning signal	Sampler	Attribution
IC-GAN (Casanova et al., 2021)	instance features	uniform over instances	to features
kNN-Diffusion (Sheynin et al., 2022)	retrieved neighbors	retrieval keys	to retrieved set
Cluster-conditioned (Adaloglou et al., 2025)	cluster index	draw over clusters	to a cluster
RCG (Li et al., 2024)	learned code	second generator	none
LRF (ours)	training item	fixed landmark table	to its landmark

framework and changes only what the network models: the variation within a group given its landmark, instead of the full data distribution. A separate line of work improves training efficiency by changing the loss rather than the inputs of the network: representation alignment adds a term that aligns the internal features of the denoiser with those of a pretrained encoder (Yu et al., 2024; Tian et al., 2026; Leng et al., 2025), which accelerates optimization without changing the sampler and could be combined with LRF. Other work lowers the variance of the regression target by averaging it over reference samples (Xu et al., 2023; Yang et al., 2026) or by conditioning on a latent variable (Guo & Schwing, 2025), and Xu et al. (2026) estimate the optimal value of the diffusion loss.

Data-conditioned generation. Several methods condition generation on training data or on representations of it. Instance-Conditioned GAN (Casanova et al., 2021) conditions the generator on the features of a training instance and models the dataset as overlapping kNN neighborhoods, whose overlaps reweight dense regions at generation time. Retrieval-augmented diffusion (Blattmann et al., 2022) and kNN-Diffusion (Sheynin et al., 2022) condition on neighbors retrieved from an external database and require a distribution over retrieval keys at generation time. RCG (Li et al., 2024) conditions on a self-supervised representation and trains a second generative model to sample it.

Discrete structure and mixtures. Other work splits generative capacity across discrete components, through learned or fixed mixture weights (Hoang et al., 2017; Ghosh et al., 2018) or through cluster indices derived from the data, as in ClusterGAN (Mukherjee et al., 2019) and Self-Conditioned GANs (Liu et al., 2020). Closest to our setting, Adaloglou et al. (2025) condition diffusion models on clusters of pretrained features and report state-of-the-art label-free FID, improved training sample efficiency, and a bound on the useful number of clusters. Their bound selects the number of clusters, whereas the residual variance ratio measures the share of the irreducible error that a given partition accounts for. Our experiments include this signal as a baseline, the Cluster model (Section 4), which isolates what the content of the landmark adds beyond the identity of the group. VQ-VAE and VQGAN (Van Den Oord et al., 2017; Esser et al., 2021) factor generation through a learned discrete code and require a second model as a prior over codes, whereas our partition is computed once from the data and each discrete symbol is a data point. Recently, Wang et al. (2026) build a Gaussian-mixture source from a coreset of the training set and train a correction flow from it.

Semi-parametric methods. Semi-parametric models place a component on the training data, from kernel density estimation to Exemplar VAE (Norouzi et al., 2020), which transforms a drawn exemplar stochastically and trains with a leave-one-out objective so that the transformation does not collapse onto the exemplar. LRF excludes each landmark from the regression targets of its own group for the same reason (Section 5), and differs in placing one component on each group rather than on each training point, with one flow shared across groups and K landmarks retained in place of the full dataset. The term "landmark" follows its use in manifold learning, where a small set of real points represents a dataset (De Silva & Tenenbaum, 2004).

Memorization and attribution. Generative models can replicate their training data (Carlini et al., 2023; Somepalli et al., 2023; Kim et al., 2025; Rao & Moyer, 2026), more often on small training sets and under conditioning (Gu et al., 2023), and several works characterize when diffusion models memorize rather than generalize (Kadkhodaie et al., 2024; Hakemi et al., 2025; Ye et al., 2026; Zhang et al., 2026). Tracing a sample to the training data that influenced it ordinarily requires post-hoc estimation (Georgiev et al., 2023). Because LRF stores landmarks, replication is measured per group against held-out data, and each sample is attributed to the landmark it is conditioned on.

3 LANDMARK-ANCHORED RECTIFIED FLOW

In this section we define the irreducible error of velocity regression, show that conditioning on a group's landmark lowers it by an inter-group component, and give the sampling procedure, the ratio that quantifies this decrease before training, and the implementation details.

3.1 THE IRREDUCIBLE ERROR OF VELOCITY REGRESSION

Rectified flow trains a velocity field to transport noise to data along straight interpolation paths (Liu et al., 2022; Lipman et al., 2022). Let $X \in \mathbb{R}^d$ be a sample with $\mathbb{E}\|X\|^2 < \infty$, let $z_0 \sim \mathcal{N}(0, I_d)$ be independent of X, and for $t \in [0, 1)$ set

$$x_t = (1-t)\, z_0 + t\, X, \qquad u = X - z_0,$$

so that x_t is the interpolant and u is the regression target. Training minimizes

$$\mathcal{L}(v) = \mathbb{E}_{t\sim\mathcal{U}[0,1]}\, \mathbb{E}\big\|u - v(x_t, t)\big\|^2$$

over measurable $v : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$. For fixed t the inner expectation is a squared-error regression of u on x_t, whose minimizer over measurable functions is the conditional expectation $v^*(x, t) = \mathbb{E}[u \mid x_t = x]$. Writing $v(x_t, t) = v^*(x_t, t) + (v(x_t, t) - v^*(x_t, t))$ and taking conditional expectations, the cross term vanishes and

$$\mathbb{E}\big\|u - v(x_t, t)\big\|^2 = \underbrace{\mathbb{E}\big[\,\mathrm{Tr}\,\mathrm{Var}(u \mid x_t)\big]}_{=:\, r(t)} + \mathbb{E}\big\|v(x_t, t) - v^*(x_t, t)\big\|^2. \tag{1}$$

The second term measures the distance of the model from the optimal predictor and vanishes at $v = v^*$, whereas the first is independent of v, and we call $r(t) = \mathbb{E}[\mathrm{Tr}\,\mathrm{Var}(u \mid x_t)]$ the irreducible error of the velocity regression at time t. No predictor attains a loss below $r(t)$ (Xu et al., 2026), and since the training objective averages over t drawn uniformly, the total training loss is bounded below by $\int_0^1 r(t)\, dt$.

Since $x_0 = z_0$ is independent of X and $u = X - z_0$, we have $r(0) = \mathrm{Tr}\,\mathrm{Var}(X)$, so at $t = 0$ the irreducible error equals the full variance of the data distribution. The irreducible error depends only on the joint distribution of the interpolant and the target, so it can be lowered either by changing the coupling between noise and data, as in reflow (Liu et al., 2022), or by giving the regression additional information about the target. LRF does the latter.

3.2 LANDMARK CONDITIONING

Let $x_1, \ldots, x_N$ be the training set in the space the flow operates in, which is either pixels or the latent space of a frozen autoencoder. We take the data distribution to be the empirical one, so X is a uniform draw from the training set and all expectations are over this draw and z_0. We fix a group size G, set $K = N/G$, and partition the items into disjoint groups $S_1, \ldots, S_K$ of G members by balanced nearest-neighbor pairing in a feature embedding $\varphi(x)$ (Section 3.5). Each group contributes one landmark $\ell_k \in S_k$, its medoid, the member minimizing the summed squared distance to the other members. We write g for the group of a sample. Each landmark is excluded from the regression targets of its group (Section 3.5), and the results below hold for these targets, with S_k replaced by $S_k \setminus \{\ell_k\}$.

The velocity network receives the landmark of the target's group, the landmark itself in additional input channels and its pretrained features $\psi(\ell_g)$ through feature-wise linear modulation (FiLM) (Perez et al., 2018), which applies a per-channel scale and shift to the intermediate activations. Since the partition is disjoint and the landmarks are distinct, conditioning on ℓ_g and conditioning on g are equivalent, and the irreducible error of the conditioned regression is $r_g(t) = \mathbb{E}[\mathrm{Tr}\,\mathrm{Var}(u \mid x_t, g)]$.

Proposition 1 (Landmark conditioning lowers the irreducible error)**.** *For any partition of the training set into groups, $r_g(t) \le r(t)$ for every t, and at $t = 0$ the gap is the inter-group variance,*

$$r(0) - r_g(0) \;=\; \mathrm{Tr}\,\mathrm{Var}\big(\mathbb{E}[X \mid g]\big).$$

The proposition follows from the law of total variance applied to the velocity target, and the proof is in Appendix A.1. The identity holds for any distribution of X with a group assignment, and we state it for the training set because the partition is defined there. At $t = 0$, landmark conditioning

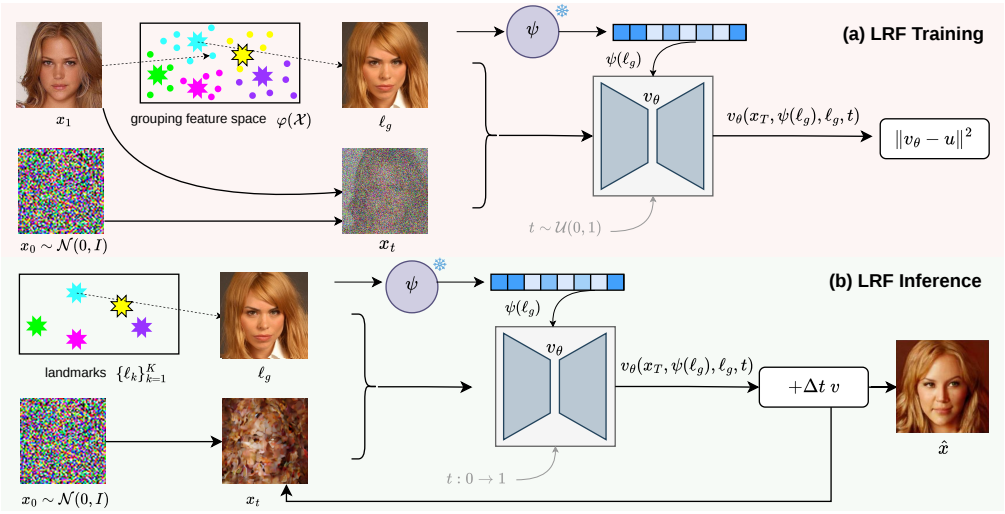

Figure 2: Training and inference in Landmark-anchored Rectified Flow. (a) During training, the landmark of the target's group conditions the velocity regression: the landmark itself enters the network as additional input channels, and its features $\psi(\ell_g)$ enter through feature-wise linear modulation (FiLM). (b) During inference, a group is drawn from the landmark table and the flow is integrated conditioned on its landmark. Here u is the velocity target and Δt is the step size.

lowers the irreducible error by the share of variance that the partition explains, and the remaining error is the intra-group variance. The landmark and the group index yield the same irreducible error but differ in what the network must learn: the landmark provides the content of the group directly, whereas an index requires the network to store an association for each group, a difference that the Cluster model of Section 4 isolates.

Remark 1 (The decrease along the trajectory). For every t, the difference $r(t) - r_g(t)$ equals $\mathbb{E}\,\|v_g^*(x_t, g) - v^*(x_t)\|^2$, the squared difference between the optimal prediction given the group and the optimal prediction given the interpolant alone (Appendix A.1). At $t = 0$ it equals the inter-group term, and it tends to zero as $t \to 1$, since the group is a function of the sample and the interpolant determines the sample in that limit.

Proposition 1 concerns optimal predictors. For networks of a given capacity, setting the conditioning weights to zero turns the LRF network into a plain rectified flow network with the same architecture, so the minimum loss of LRF is no higher than that of the plain flow, as in the residual-learning argument of He et al. (2016). This holds for feature modulation and for input-channel concatenation (Figure 2), and the two networks have identical parameter counts.

3.3 Sampling

Sampling draws a group index k uniformly from $\{1, \ldots, K\}$ and integrates the flow conditioned on ℓ_k. Because the partition is balanced, this draw reproduces the group marginal of the data distribution, with no learned or estimated component.

Proposition 2 (Sampler marginal law)**.** *Let p_k be the data distribution conditioned on group k, so that $p = \frac{1}{K}\sum_k p_k$ for a balanced partition, let q_k be the law of the intra-group model conditioned on ℓ_k, and let w be a probability vector over groups. Write W_1 for the Wasserstein-1 distance in the training space and D for the diameter of the training set. Then*

$$W_1\Big(\textstyle\sum_k w_k q_k,\ p\Big) \;\le\; \underbrace{\textstyle\sum_k w_k\, W_1(q_k, p_k)}_{\textit{intra-group model}} \;+\; \underbrace{\tfrac{D}{2}\textstyle\sum_k |w_k - \tfrac{1}{K}|}_{\textit{group weights}},$$

and the uniform draw over a balanced partition sets the second term to zero.

Proof. Add and subtract $\sum_k w_k p_k$ and apply the triangle inequality. Mixing couplings of the pairs (q_k, p_k) with weights w_k gives a coupling of the two mixtures, which bounds the first term by

$\sum_k w_k W_1(q_k, p_k)$. For the second, by the Kantorovich–Rubinstein duality it suffices to bound $\sum_k (w_k - \frac{1}{K}) \int f \, dp_k$ over 1-Lipschitz f. The coefficients sum to zero, so f can be replaced by $f - c$, where c is the midpoint of the range of f on the training set. There $|f - c| \le D/2$, which gives $\frac{D}{2} \sum_k |w_k - \frac{1}{K}|$. □

3.4 The residual variance ratio

The gap in Proposition 1 is measurable from the data alone. Write μ_k for the mean of group S_k and define

$$\rho \;=\; \frac{\frac{1}{N} \sum_k \sum_{i \in S_k} \|x_i - \mu_k\|^2}{\frac{1}{N} \sum_i \|x_i - \bar{x}\|^2}, \tag{2}$$

the fraction of total variance that the partition leaves unexplained, which lies in $[0, 1]$ by the law of total variance. Computing it requires one feature pass and one pairing, before any flow training. In our experiments ρ is computed over all training items, including the landmarks.

Corollary 1. $r(0) - r_g(0) \;=\; (1 - \rho)\ \mathrm{Tr}\,\mathrm{Var}(X)$.

Proof. By the law of total variance, the inter-group variance equals $\mathrm{Tr}\,\mathrm{Var}(X)$ minus the intra-group variance, which is $\rho\,\mathrm{Tr}\,\mathrm{Var}(X)$ by equation 2. Proposition 1 identifies this quantity with $r(0) - r_g(0)$. □

Section 5.1 compares ρ with the realized quality gain.

3.5 Implementation details

We build the partition once, before training, by greedy minimum-distance matching applied hierarchically. At each level we match the closest available pair until every item is paired, represent each pair by its centroid, and repeat this $\log_2 G$ times, which requires G to be a power of two and keeps the groups balanced at every level. On MNIST, CIFAR-10, and CelebA-256, all pairwise distances in the grouping embedding are sorted globally at each level. On ImageNet-256 the candidate pairs are restricted to the 32 nearest neighbors of each item, retrieved with a nearest-neighbor index, and any items left unmatched are paired recursively, which reduces the cost of a level from quadratic in N to $O(N \log N)$ after the neighbor search. The construction is deterministic given the embedding. Proposition 1 holds for any partition and Proposition 2 for any balanced one.

We use a balanced partition for three practical reasons. It avoids groups so small that the flow has little to model beyond the landmark itself, it keeps the modeling problems of different groups comparable, and it lets the held-out groups used to calibrate the memorization measurement match the training groups in size.

The grouping embedding φ concatenates features of each image from the supervised ImageNet ResNet-18 of torchvision (`IMAGENET1K_V1`), standardized per run, with a PCA of its training-space representation, mixed by a weight λ. Pixel datasets use $\lambda = 0.5$, and for the latent datasets the PCA block is dropped ($\lambda = 0$), a choice made empirically in Section 5. Each landmark is excluded from the regression targets of its own group (Section 5), and landmark conditioning is dropped with probability 0.1 during training. In the DiT-B/2 model of Section 4.3, the landmark is patch-embedded together with the state and its features enter through adaptive layer normalization. Within an experiment all models share the same backbone and the same number of backbone parameters, trained with Adam at identical settings. Sampling uses 64 Euler steps. Full hyperparameters and the run protocol are in Appendix A.2.

4 Experiments

We compare four models across four datasets and on two backbones, and report sample quality, diversity, and the rate at which samples reproduce training images. The four models in each comparison are trained in the same run and share the data subset, the partition, and the reference statistics: LRF, a plain rectified flow (RF), an RCG-style model that conditions on a 64-d PCA projection of the item's pretrained features and samples this code at generation time with a small flow trained jointly,

and a Cluster model that conditions on the group index of the same partition through a learned embedding of the same dimension as the landmark features while keeping our uniform draw, so that the closest published conditioning signal (Adaloglou et al., 2025) is evaluated with our sampler. LRF and RF are parameter-identical, the RCG-style model has 0.7% more parameters and the Cluster model 1.0%, 3.8%, and 3.6% more on MNIST, CIFAR-10, and CelebA-256 (ImageNet-256 counts are in Section 4.2). A fixed code per group is the Cluster embedding kept at its initial values. RF and the Cluster model regress on all training images, and LRF on all images other than the landmarks.

We train three seeds per dataset, for 40k iterations on MNIST, 100k on CIFAR-10, and 60k on CelebA-256, with the same budget for all models of a comparison. ImageNet-256 uses 1,279,104 images in SD-VAE latents and 79,944 groups, with three seeds, 300k iterations per model, and 50,000 samples per model for evaluation.

Sample quality is reported under clean-FID (Parmar et al., 2022) and FD-DINOv2 (Stein et al., 2023), computed on generated samples against the full training set, and diversity under precision and recall (Kynkäänniemi et al., 2019) and density and coverage (Naeem et al., 2020). Against held-out images, the gap between LRF and RF changes by at most 2.2 percentage points (Appendix A.4). A sample is counted as memorized when its nearest training neighbor in feature space is closer than a threshold calibrated on held-out data. Numbers below are means with standard deviations over three seeds, except in Table 4, which reports a single seed.

4.1 Sample quality on MNIST, CIFAR-10, and CelebA-256

At matched parameters and budget, LRF attains lower clean-FID than RF on all three datasets (Table 2), with 5.7 against 18.1 on MNIST, 38.1 against 50.4 on CIFAR-10, a difference of 24.4%, and 22.2 against 22.8 on CelebA-256, a difference of 2.4%. FD-DINOv2 agrees on MNIST, with 115 against 260, and on CIFAR-10, with 1041 against 1168. On CelebA-256 RF is ahead under FD-DINOv2, with 646 against 652, and the per-seed results are mixed: RF is ahead on one of three seeds under clean-FID and on two of three under FD-DINOv2. The ordering on MNIST and CIFAR-10 holds on every seed and under both metrics.

On CelebA-256, where RF needs 120k to 140k iterations to plateau, we repeated the comparison at 150k with four models, where LRF attains 22.5 against 23.0 clean-FID, and at 200k with two, and the ordering is unchanged. LRF also reaches the final quality of RF earlier in training. On the two runs where clean-FID was measured throughout training, LRF reaches that level at 39k iterations on CIFAR-10 and at 43k on CelebA at 64px, against 79k and 93k for RF, with crossings interpolated linearly between checkpoints 10k apart and the final level taken as the median of the last three measurements. On CIFAR-10 the gap does not close: the last three measurements of that run give 40.6 for LRF against 47.0. Comparisons of LRF with the Cluster and RCG-style models are dataset-dependent and are analyzed in Section 5.

4.2 Sample quality on ImageNet-256

On ImageNet-256, where $\rho = 0.84$, LRF attains 16.3% lower clean-FID than RF (117.2 against 140.1, Table 3) and 4.3% lower FD-DINOv2 (1436.6 against 1500.7), and it has the highest precision, density, and coverage. The improvement holds on all three seeds, at 17.0%, 15.4%, and 16.6% under clean-FID. LRF and the Cluster and RCG-style models differ by at most 1.4 points of clean-FID, so on ImageNet-256 the presence of a conditioning signal matters more than its form. The Cluster embedding adds 41M parameters at this number of groups, 53.4M in total against 12.4M for LRF, and that cost grows linearly in the number of groups. LRF instead stores the 79,944 landmarks with their features and uses them at generation time.

4.3 Experiments on DiT-B/2

We repeat the matched comparison with a DiT-B/2 backbone of 130.5M parameters, in which the landmark enters through the patch embedding and its features through adaptive layer normalization, on CelebA-HQ at 256px with 27,952 images and 150k iterations per model. LRF attains 5.4% lower clean-FID and 3.3% lower FD-DINOv2 than RF (Table 4). Figure 3 shows samples from this model, with the landmark held fixed along each column.

Table 2: Comparison on MNIST, CIFAR-10, and CelebA-256 at matched parameters and budget, mean and standard deviation over three seeds. Models are trained for 40k (MNIST), 100k (CIFAR-10), and 60k (CelebA-256) iterations and evaluated with 50k, 50k, and 10k samples. Bold: best per column.

Model	MNIST ($\rho = 0.37$)		CIFAR-10 ($\rho = 0.73$)		CelebA-256 ($\rho = 0.70$)	
	clean-FID	FD-DINOv2	clean-FID	FD-DINOv2	clean-FID	FD-DINOv2
LRF (ours)	**5.7** $\pm$ 0.7	**115** $\pm$ 18	**38.1** $\pm$ 1.4	**1041** $\pm$ 7	22.2 $\pm$ 0.6	652 $\pm$ 17
Cluster	15.1 $\pm$ 2.2	274 $\pm$ 15	40.6 $\pm$ 1.4	1077 $\pm$ 14	22.1 $\pm$ 5.0	**623** $\pm$ 49
RCG-style	18.7 $\pm$ 4.3	272 $\pm$ 45	48.7 $\pm$ 2.1	1173 $\pm$ 29	**20.5** $\pm$ 0.7	627 $\pm$ 5
Plain RF	18.1 $\pm$ 3.5	260 $\pm$ 38	50.4 $\pm$ 1.3	1168 $\pm$ 16	22.8 $\pm$ 1.7	646 $\pm$ 8

Table 3: ImageNet-256 at matched backbone parameters and budget, mean and standard deviation over three seeds, 50,000 samples per model. Bold: best per column.

Model	clean-FID $\downarrow$	FD-DINOv2 $\downarrow$	Precision $\uparrow$	Recall $\uparrow$	Density $\uparrow$	Coverage $\uparrow$
LRF (ours)	**117.2** $\pm$ 3.0	1436.6 $\pm$ 14.5	**0.783**	0.077	**0.944**	**0.190**
Cluster	118.6 $\pm$ 4.1	1454.3 $\pm$ 16.3	0.766	0.064	0.894	0.183
RCG-style	117.8 $\pm$ 0.5	**1431.9** $\pm$ 9.9	0.741	**0.095**	0.780	0.181
Plain RF	140.1 $\pm$ 4.9	1500.7 $\pm$ 24.5	0.739	0.039	0.784	0.126

4.4 Memorization

Because LRF stores K landmarks, we measure how often a sample reproduces a training image. On ImageNet-256, the setting in which the models are trained on the full data, the memorization rate of LRF is 0.058 ± 0.006 against 0.094 ± 0.047 for RF, 0.073 ± 0.021 for the Cluster model, and 0.078 ± 0.042 for the RCG-style model, so conditioning on stored training items does not increase memorization at this scale and gives the most stable rate across seeds. On the smaller training sets the rate of LRF is higher than that of RF, at 0.050 ± 0.012 against 0.035 ± 0.006 on CelebA-256, and the effect grows as the number of images per group falls (Section 5.2). The rate is reported per group, so a group with a high rate can be down-weighted or removed from the mixture without retraining (Appendix A.3). On the DiT-B/2 model, 55.2% of the samples conditioned on the 81 landmarks annotated with eyeglasses wear eyeglasses, against 49.7% of the training images in those groups, and removing these groups lowers the rate among all samples from 5.2% to 2.7%.

5 Analysis

5.1 The residual variance ratio and the realized gain

We compare the gain of LRF over RF with ρ over all trained configurations: group sizes $G \in \{4, 16, 64\}$ on MNIST and CelebA-256 and $G \in \{4, 8, 16, 32, 64\}$ on CIFAR-10, with three seeds per configuration on MNIST and CIFAR-10 and one seed for $G = 4$ and $G = 64$ on CelebA-256. Every configuration with $\rho \leq 0.48$, which includes all MNIST settings, reaches the FID of held-out real images (Appendix A.2), so FID cannot separate LRF from real samples. Configurations with ρ near 0.7 improve on RF by 15 to 23% in FID.

Within a dataset, ρ is weakly predictive of the gain. On CIFAR-10 the FID gain peaks at $+15\%$ for $G = 8$ to 16 and is -2.6% at $G = 4$, consistent with the optimal cluster granularity reported by Adaloglou et al. (2025). On CelebA-256, LRF with groups formed on image features alone outperforms the hybrid grouping on two seeds, although its ratio is higher (0.742 against 0.699). The ratio therefore separates datasets but does not rank partitions of the same data.

A random balanced partition leaves 8% of the variance between groups ($\rho = 0.92$), against 27% for the feature partition. At 100k iterations on CIFAR-10, LRF with the random partition attains 50.1 clean-FID and 1185 FD-DINOv2, against 40.2 and 1056 with the feature partition and 53.1 and 1189 for RF (Appendix A.5).

5.2 Conditioning signals

Against RF, LRF improves clean-FID on all three datasets (Table 2) and FD-DINOv2 on MNIST and CIFAR-10, with the CelebA-256 exception reported in Section 4.1. Against the Cluster model,

Table 4: DiT-B/2 on CelebA-HQ at 256px, at matched parameters and budget, single seed, 10,000 samples per model. Bold: best per column.

Model	clean-FID↓	FD-DINOv2↓
LRF (ours)	**8.94**	**266.4**
Plain RF	9.45	275.5

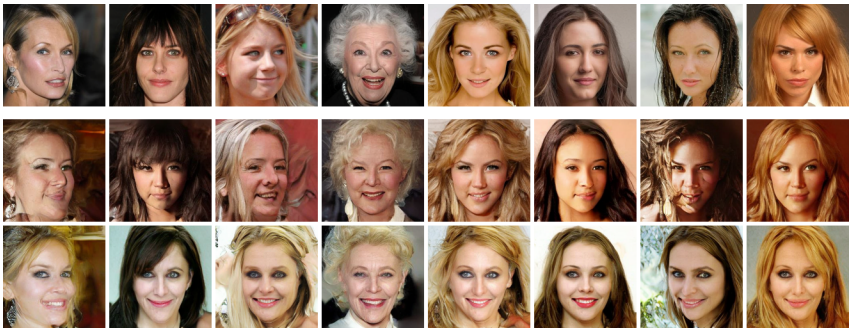

Figure 3: Samples from the DiT-B/2 model on CelebA-HQ. Along a column the landmark is fixed and the noise varies, so the samples share the content of the landmark and differ within the group.

the comparison is dataset-dependent: LRF is ahead on CIFAR-10 on every seed under both metrics, and the Cluster model is ahead on CelebA-256 under both. Against the RCG-style model, LRF is ahead on MNIST and CIFAR-10, and the RCG-style model is ahead on CelebA-256 under both.

Two explanations are possible for the ordering on CelebA-256: faces are homogeneous, or the VAE-latent training space is unfavorable to content conditioning. We separate them with the same faces at 64px in pixel space, with all four models and every other setting fixed, and measure clean-FID every 10k iterations (Section 4.1). In pixel space LRF is ahead of the Cluster model by a consistent margin, as on CIFAR-10, so the ordering on CelebA-256 follows from the latent space rather than from faces. The RCG-style model is ahead during training and matches LRF at the end, so its advantage follows from the homogeneity of the data. Pixel-space grouping also lowered the ratio to 0.549, predicting a larger gain than in the latent setting, and the FID gain was 26 to 36% against 21% in latents.

None of the three interventions in the latent setting, each targeting a different mechanism, separates LRF from the Cluster model. LRF with decoded-landmark feature maps as the conditioning signal is ahead of the Cluster model until 40k iterations and matches it afterwards, image-feature-only grouping improves both models by a similar amount, and a landmark-centered source distribution with σ untuned reduces the FID gain to 5.1% and raises the fraction of memorized samples whose nearest training image is the landmark to 0.12, against 0.05 expected by chance.

Restoring each landmark to the regression targets of its own group raises the fraction of memorized samples whose nearest training image is a landmark from 0.05 to 0.65. Group size affects measured memorization more than the ratio does: at $G = 4$ on CelebA-256 the memorization rate late in training is 14 to 19%, with 0.74 of the memorized samples closest to a training image of their own group, while the pixel-space face run at a comparable ratio and $G = 16$ stays at the calibrated base rate.

6 Conclusion

We identified an inter-group component in the irreducible error of the velocity regression, measured its size before training, and lowered it by conditioning the flow on a landmark drawn from the training set. At matched parameters and budget this improves clean-FID on four datasets and FD-DINOv2 on three of them, and reaches the final quality of RF earlier in training, and the generated distribution is a finite mixture over the landmarks, which gives each sample a known conditioning landmark and allows groups to be reweighted and examined for memorization. Limitations are discussed in Appendix A.6.

AI USE STATEMENT

In this work, we used AI tools to proofread and improve the clarity and presentation of the manuscript. AI tools were also used for debugging and making modifications to the code used in our experiments. All other aspects of the work, including the research ideas, methodology, experimental design, analysis, and conclusions, were conducted without the use of AI.

REPRODUCIBILITY STATEMENT

The construction is specified in Section 3, with the partition procedure, the grouping features, and the conditioning inputs in Section 3.5. Proofs are in Section 3 and Appendix A.1, with their assumptions stated in the propositions themselves. The training and evaluation protocol, including dataset sizes, budgets, seeds, sample counts, and the reference statistics used for every metric, is in Section 4 and Appendix A.2. Code will be released upon acceptance.

REFERENCES

Nikolaos Adaloglou, Tim Kaiser, Felix Michels, and Markus Kollmann. Rethinking cluster-conditioned diffusion models for label-free image synthesis. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 3603–3613, 2025.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

Vin De Silva and Joshua B Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, technical report, Stanford University, 2004.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.

Arnab Ghosh, Viveka Kulharia, Vinay P Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8513–8521, 2018.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

Pengsheng Guo and Alexander G Schwing. Variational rectified flow matching. *arXiv preprint arXiv:2502.09616*, 2025.

Shahin Hakemi, Naveed Akhtar, Ghulam Mubashar Hassan, and Ajmal Mian. Deeper diffusion models amplify bias. *arXiv preprint arXiv:2505.17560*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. Multi-generator generative adversarial nets. *arXiv preprint arXiv:1708.02556*, 2017.

Zahra Kadkhodaie, Florentin Guth, Eero Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, volume 2024, pp. 46543–46567, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Juyeop Kim, Songkuk Kim, and Jong-Seok Lee. How diffusion models memorize. *arXiv preprint arXiv:2509.25705*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *Advances in neural information processing systems*, 37:63082–63109, 2024.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272. IEEE, 2025.

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37: 125441–125468, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14286–14295, 2020.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan. Clustergan: Latent space clustering in generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4610–4617, 2019.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pp. 7176–7185. PMLR, 2020.

Sajad Norouzi, David Fleet, and Mohammad Norouzi. Exemplar vae: Linking generative models, nearest neighbor retrieval, and data augmentation. *Advances in Neural Information Processing Systems*, 33:8753–8764, 2020.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11400–11410. IEEE, 2022.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Mingxing Rao and Daniel Moyer. Generalization and memorization in rectified flow. In *European Conference on Computer Vision*, pp. 537–554. Springer, 2026.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

George Stein, Jesse 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. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Yuchuan Tian, Hanting Chen, Mengyu Zheng, Yuchen Liang, Chao Xu, and Yunhe Wang. U-repa: Aligning diffusion u-nets to vits. *Advances in Neural Information Processing Systems*, 38:11003–11024, 2026.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Xiao Wang, Zihua She, and Jianxi Su. Coreset-induced conditional velocity flow matching. *arXiv preprint arXiv:2605.12951*, 2026.

Yilun Xu, Shangyuan Tong, and Tommi Jaakkola. Stable target field for reduced variance score estimation in diffusion models. *arXiv preprint arXiv:2302.00670*, 2023.

Yixian Xu, Shengjie Luo, Liwei Wang, Di He, and Chang Liu. Diagnosing and improving diffusion models by estimating the optimal loss value. In *International Conference on Learning Representations*, volume 2026, pp. 134165–134197, 2026.

Donglin Yang, Yongxing Zhang, Xin Yu, Liang Hou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Renjie Liao. Stable velocity: A variance perspective on flow matching. *arXiv preprint arXiv:2602.05435*, 2026.

Zeqi Ye, Qijie Zhu, Molei Tao, and Minshuo Chen. Provable separations between memorization and generalization in diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 143375–143425, 2026.

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. *arXiv preprint arXiv:2410.06940*, 2024.

Yichi Zhang, Yici Yan, Alex Schwing, and Zhizhen Zhao. Towards hierarchical rectified flow. In *International Conference on Learning Representations*, volume 2025, pp. 41537–41558, 2025.

Zekai Zhang, Xiao Li, Xiang Li, Lianghe Shi, Meng Wu, Molei Tao, and Qing Qu. Generalization of diffusion models arises with a balanced representation space. In *International Conference on Learning Representations*, volume 2026, pp. 101444–101475, 2026.

A Appendix

A.1 Proof of Proposition 1

Fix t and abbreviate $v^*(x) = \mathbb{E}[u \mid x_t = x]$ and $v_g^*(x, k) = \mathbb{E}[u \mid x_t = x, g = k]$. Both minimize their respective L^2 errors, so $r(t) = \mathbb{E}\|u - v^*(x_t)\|^2$ and $r_g(t) = \mathbb{E}\|u - v_g^*(x_t, g)\|^2$. By the tower property, $\mathbb{E}[v_g^*(x_t, g) \mid x_t] = v^*(x_t)$, and the L^2 Pythagorean identity gives

$$r(t) - r_g(t) \;=\; \mathbb{E}\left\|v_g^*(x_t, g) - v^*(x_t)\right\|^2 \;\geq\; 0,$$

which proves the first claim.

At $t = 0$ we have $x_0 = z_0$, which is independent of (X, g). Hence $v^*(z_0) = \mathbb{E}[X] - z_0$ and $v_g^*(z_0, k) = \mathbb{E}[X \mid g = k] - z_0$, so the gap above equals $\mathbb{E}\,\|\mathbb{E}[X \mid g] - \mathbb{E}[X]\|^2 = \operatorname{Tr}\operatorname{Var}(\mathbb{E}[X \mid g])$, the inter-group variance. □

Remark 1. The identity is the display above. For the limit, $x_t \to X$ almost surely as $t \to 1$, and g is a function of X. On the empirical distribution the posterior over (X, g) given x_t concentrates on the true pair, so $v_g^*(x_t, g) - v^*(x_t) \to 0$ pointwise, and the gap vanishes by dominated convergence (all quantities are bounded on the finite support). □

A.2 Run protocol and hyperparameters

All models of a configuration are trained in a single run with shared data, partition, references, and RNG stream. Budgets: 40k/100k/60k iterations (MNIST/CIFAR/CelebA), evaluations every 1–2.5k iterations, three seeds per dataset, means with standard deviations reported. CelebA is repeated at 150k with four models and at 200k with two. Architecture: a shared U-Net backbone (base width 64 on MNIST and CIFAR and 96 on CelebA, with 3.35M, 3.45M, and 7.22M parameters), the landmark as extra input channels plus FiLM features, landmark dropout 0.1, Adam at a constant learning rate of 10^{-3} without EMA, batch 128, and 64 Euler steps at sampling without guidance. The DiT-B/2 models use EMA weights. CelebA-256 trains in SD-VAE latents ($4 \times 32 \times 32$), and metrics are computed on decoded images. Clean-FID (Parmar et al., 2022) and FD-DINOv2 (Stein et al., 2023) use 50k generated samples on MNIST and CIFAR-10 and 10k on CelebA-256 against the full training reference. The FID values in Section 5 are computed during training on the features of the grouping backbone projected to 64 dimensions, against held-out images. Memorization is measured with nearest-neighbor thresholds calibrated on held-out data. The runs on MNIST, CIFAR-10, and CelebA fit a single consumer GPU (RTX 3090), and per-run wall-clock ranges from twenty minutes (MNIST) to ≈1 day (CelebA at 150k).

A.3 Deletion and reweighting

We present the deletion and reweighting experiment of Section 4.4 on the DiT-B/2 model trained on CelebA-HQ. We report the rate of eyeglasses among samples and clean-FID when the groups whose landmark wears eyeglasses are removed from the mixture or given a weight of 0.30, without retraining (Table A.1), and the rate among the training images and the samples of each group (Table A.2).

A.4 Held-out evaluation

LRF conditions on training images, so a gain measured against the training images, the reference of Table 2, could reflect a closer fit to those images rather than to the data distribution. We therefore also compute clean-FID and FD-DINOv2 of the models in Table 2 against held-out images, from the same 10,000 samples per model (Table A.3). Against held-out images, the relative clean-FID gap between LRF and RF is 64.1% on MNIST, 23.6% on CIFAR-10, and 1.8% on CelebA-256, compared with 66.3%, 24.2%, and 1.9% against the training images. The FD-DINOv2 gaps change by at most 0.5 points, and the ordering of the models is the same under both references.

Table A.1: Removing or up-weighting the groups whose landmark wears eyeglasses, without retraining (DiT-B/2, CelebA-HQ). D is the 81 of 1,747 groups whose landmark is annotated with eyeglasses. Removing D lowers the eyeglasses rate and the FID vs. images w/o eyeglasses, and raises the FID vs. all images. Eyeglasses are detected with a classifier trained on the CelebA-HQ labels (held-out AUC 0.989).

			FID vs. training images	
Edit	Weight of D (%)	With eyeglasses (%)	w/o eyeglasses	all
None	4.6	5.2 ± 0.2	9.64 ± 0.06	8.98 ± 0.08
Remove D	0	2.7 ± 0.2	9.05 ± 0.12	9.67 ± 0.15
Up-weight D	30	18.4 ± 0.4	17.52 ± 0.09	13.63 ± 0.07

Table A.2: Samples reproduce the eyeglasses rate of their conditioning group. Training-image rates exclude the landmark and use the 1,029 of 1,747 groups whose recomputed partition reproduces the stored landmark, since recomputing the grouping features on a different GPU changes some pairings. Sample rates as in Table A.1.

	With eyeglasses (%)	
Group	Training images	Samples
Landmark w/ eyeglasses (D)	49.7	55.2 ± 1.2
Landmark w/o eyeglasses	3.0	2.7 ± 0.2
All groups	4.9	5.2 ± 0.2

Table A.3: Clean-FID and FD-DINOv2 against the training images and against held-out images, 10,000 samples per model, mean and standard deviation over three seeds. Held-out images are the test splits of MNIST and CIFAR-10 and the CelebA-HQ images outside the training subset of each run.

		Clean-FID		FD-DINOv2	
Dataset	Model	training	held-out	training	held-out
MNIST	LRF (ours)	6.2 ± 0.7	6.9 ± 0.8	116 ± 17	117 ± 17
	Cluster	15.5 ± 2.2	16.3 ± 2.2	277 ± 16	281 ± 16
	RCG-style	19.4 ± 4.3	20.0 ± 4.3	275 ± 47	278 ± 47
	Plain RF	18.6 ± 3.6	19.3 ± 3.6	264 ± 39	268 ± 39
CIFAR-10	LRF (ours)	40.2 ± 1.2	42.3 ± 1.2	1056 ± 4	1068 ± 4
	Cluster	42.4 ± 1.3	44.7 ± 1.3	1090 ± 17	1103 ± 17
	RCG-style	50.7 ± 2.0	53.0 ± 2.0	1188 ± 26	1200 ± 25
	Plain RF	53.1 ± 1.1	55.4 ± 1.0	1189 ± 18	1201 ± 18
CelebA-256	LRF (ours)	22.2 ± 0.6	22.4 ± 0.6	652 ± 17	652 ± 17
	Cluster	22.1 ± 5.0	22.2 ± 5.1	623 ± 49	623 ± 48
	RCG-style	20.5 ± 0.7	20.6 ± 0.8	627 ± 5	628 ± 5
	Plain RF	22.8 ± 1.7	22.9 ± 1.7	646 ± 8	647 ± 8

A.5 Control experiments

We present three controls on CIFAR-10 (Table A.4): LRF with a random balanced partition, LRF with a random group member as landmark, and RF with base width 72. The random partition is trained for 100k and 30k iterations, and the other two controls for 30k iterations, each compared with LRF and RF at the same budget.

Table A.4: Control experiments on CIFAR-10, mean and standard deviation over three seeds, 10,000 samples per model, against the training images. The feature partition has $\rho = 0.73$ and the random partition $\rho = 0.92$. Plain RF with base width 72 has 4.3M parameters, against 3.4M for the other models.

Iterations	Model	Clean-FID	FD-DINOv2
100k	LRF (ours)	40.2 ± 1.2	1056 ± 4
	LRF, random partition	50.1 ± 3.6	1185 ± 21
	Plain RF	53.1 ± 1.1	1189 ± 18
30k	LRF (ours)	50.7 ± 5.5	1169 ± 53
	LRF, random partition	52.4 ± 2.2	1206 ± 33
	LRF, random landmark	48.2 ± 3.2	1166 ± 37
	Plain RF	54.8 ± 3.2	1222 ± 25
	Plain RF, base width 72	53.7 ± 3.2	1203 ± 29

A.6 Limitations

The claims are matched-budget comparisons rather than state-of-the-art results. The ratio separates datasets but does not rank partitions of the same data: the best partition we found has a higher ratio than the one it outperforms (Section 5.1). In VAE-latent space LRF performs on par with the Cluster model, including under three interventions that target different mechanisms, while remaining ahead in pixel space on the same images. Changing the regression target in latent space, for example with a perceptually weighted loss, is left for future work. The construction is compatible with methods that act on the network's internal features, such as representation alignment (Yu et al., 2024; Tian et al., 2026), which apply to LRF unchanged. The grouping uses features of a supervised ImageNet classifier, so on ImageNet-256 the conditioning of LRF, Cluster, and RCG-style carries class information that RF does not receive, and the gain over RF includes its effect. ImageNet-256 is evaluated against the training split only, and training cost is reported in iterations rather than wall-clock time. The U-Net models are trained without EMA at a constant learning rate. A lower irreducible error on the training set also makes the training images easier to reproduce, and the memorization measure flags samples by their proximity to the training set, so it also responds to sample fidelity. The deployed model stores its landmarks, which on CelebA-HQ are photographs of people, and links each sample to one of them.

A.7 Additional figures

Figure A.2 shows samples from the model used in Section 4.3, Figure A.3 shows the training image each sample is conditioned on, and Figure A.1 shows the effect of interpolating the conditioning signal between two landmarks at fixed noise.



Figure A.1: Samples obtained by interpolating the conditioning signal between two landmarks at fixed noise.



Figure A.2: Samples generated by LRF with a DiT-B/2 network on CelebA-HQ.



Figure A.3: Samples generated from the landmark of each group, one group per row. The left column shows the landmarks, and the images to their right are generated while conditioned on them.