WHAT DOES ONE-STEP FLOW MATCHING LEARN WITH ADAPTIVE WEIGHTING?

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step generative models replace iterative sampling with one network evaluation, but training them remains challenging. MeanFlow uses adaptive loss weighting to stabilize training, downweighting large residuals that can reflect valid target variation rather than prediction error. Through improved MeanFlow's velocity-regression formulation, we characterize the selected conditional law, giving necessary and sufficient conditions on radial weights for it to have no greater total variance than the original law at every stationary prediction, accounting for mean shifts. Population flows reveal changes in generated proportions even with convex regression and no neural fitting, while image experiments show that similar FID can conceal substantial shifts in class composition. Our prediction weighting removes target dependence, retaining adaptive scaling and the pointwise conditional mean while better matching training class proportions in our unconditional generation experiments. Comparisons on CIFAR-10 and ImageNet-LT distinguish target preservation from sample quality, establishing weighting as a generative modeling choice rather than merely a stabilization technique.

1 INTRODUCTION

Diffusion and flow models generate high-quality images by progressively transforming noise into data (Karras et al., 2022; Lipman et al., 2023). Their iterative samplers require repeated network evaluations, motivating one-step approaches such as consistency models (Song et al., 2023). MeanFlow offers a conceptually simple formulation based on average velocity, enabling one-call generation without a pretrained teacher, distillation, or a training curriculum (Geng et al., 2025a). Improved MeanFlow (iMF) further strengthens this approach, substantially narrowing the quality gap between one-step and multi-step generation (Geng et al., 2026).

MeanFlow training involves Jacobian-vector products (JVPs), which add computation and can amplify stochastic variation in their velocity inputs (Geng et al., 2026). Loss weighting helps address these optimization difficulties, as it does in consistency training (Geng et al., 2025b; Lu & Song, 2025). MeanFlow's *adaptive weighting* assigns smaller, detached weights to larger residuals and yields substantial sample-quality gains (Geng et al., 2025a). The rule is retained or adapted in iMF (Geng et al., 2026), AlphaFlow (Zhang et al., 2026), and pMF (Lu et al., 2026), and in subsequent extensions (Li et al., 2026; Zaytsev et al., 2026). Across these successful recipes, weighting is mainly assessed through optimization behavior and sample quality, with details sometimes confined to appendices or official code. These criteria do not establish whether it helps learn the original regression target or selects a different one.

We analyze the conditional targets selected by adaptive weighting using the velocity-regression formulation of iMF (Geng et al., 2026). This formulation expresses average-velocity learning as regression on instantaneous velocity targets. The corrected prediction is fixed by the model inputs, while different clean examples supply valid targets at the same noisy state. Their conditional mean defines the marginal velocity of the prescribed probability path (Lipman et al., 2023; Albergo et al., 2025). A large residual can therefore reflect variation that the model must average, not simply prediction error. Giving these targets different weights can change the mean being learned, rather than only the strength of the update. On CIFAR-10-LT100, we find substantial changes in estimated class composition even when FID against the training distribution remains nearly unchanged (Section 5).

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.

We ask what structure this selection imposes on the conditional law and how it affects generation. Our contributions are:

- **Conditional target analysis.** We give necessary and sufficient conditions for radial weights to select a law with no greater total variance than the original law at every stationary prediction. The result accounts for mean shifts and gives the sharp range $0 < p \leq 1$ for inverse-residual weighting.
- **Consequences for generation.** Population flows show that generated proportions can change even under convex regression and without neural fitting. CIFAR-10 experiments show that better FID need not imply closer agreement with training class proportions.
- **Target-preserving weighting.** We introduce *prediction weighting*, replacing the residual norm in the weight with the prediction norm. It retains adaptive scaling while preserving the pointwise conditional-mean target, without an auxiliary network.

Prediction weighting improves FID and class-prior agreement on CIFAR-10. In a guided ImageNet-LT recipe adapted from pMF (Lu et al., 2026), it improves over unweighted training but trails adaptive weighting. The gap narrows substantially with perceptual continuation. These results distinguish target preservation from sample quality and support treating adaptive weighting as a generative modeling choice, not merely an optimization technique.

2 Related Work

From iterative to one-step generation. Diffusion and flow models learn fields for iterative transport (Song et al., 2021; Lipman et al., 2023; Liu et al., 2023). Flow-map learning instead represents finite-time transport directly (Boffi et al., 2025), encompassing consistency models (Song et al., 2023) and MeanFlow's average-velocity parameterization (Geng et al., 2025a). Training can use teacher distillation (Salimans & Ho, 2022) or proceed from scratch. Our analysis uses iMF's velocity-regression formulation (Geng et al., 2026), and our ImageNet study adapts pMF's pixel-space recipe (Lu et al., 2026).

Adaptive objectives for generation. Improved consistency training uses pseudo-Huber loss (Song & Dhariwal, 2024), and ECT interprets its derivative through per-sample weighting and gradient-variance control (Geng et al., 2025b). MeanFlow studies the inverse-residual exponent through FID ablations (Geng et al., 2025a); subsequent recipes retain or adapt the rule, including scale adjustment in AlphaFlow (Zhang et al., 2026) and finite-difference training in Euler Mean Flows (Li et al., 2026). Interval Denoiser's logarithmic loss has the same gradient as $p = 1$ adaptive weighting, alongside residual clipping (Zaytsev et al., 2026). Robust and perceptual objectives also improve rectified-flow training (Lee et al., 2024). These studies emphasize optimization and quality. We examine the selected conditional law.

Weighting and regression targets. Generator Matching characterizes Bregman losses by equality of conditional and marginal training gradients across probability paths and data distributions (Holderrieth et al., 2025). Using the iMF formulation (Geng et al., 2026), we characterize the conditional law selected by adaptive weighting and establish a sharp total-variance guarantee. Related observations include marginal changes under pseudo-Huber in Simple ReFlow (Kim et al., 2025) and endpoint-pair reweighting in Weighted CFM (Calvo-Ordoñez et al., 2025). Target-independent weights can preserve pointwise means (Kim et al., 2025; Billera et al., 2025). Prediction weighting supplies a simple rule without an auxiliary network. This differs from reducing the variance of an estimator of the same mean (Bertrand et al., 2025; Yang et al., 2026).

3 Conditional Target Selection

3.1 Weighted regression and stationary predictions

For a straight flow-matching path, draw data X_0 and independent Gaussian noise X_1, and set

$$X_t = (1-t)X_0 + tX_1, \qquad Y = X_1 - X_0. \tag{1}$$

Generation integrates from noise at $t = 1$ to data at $t = 0$. Conditional flow matching (CFM) regresses Y on the state $Z = (X_t, t)$. The conditional mean $\mu(z) = \mathbb{E}[Y \mid Z = z]$ is the marginal velocity of the prescribed probability path (Lipman et al., 2023; Albergo et al., 2025; Liu et al., 2023).

In iMF, the network $u_\theta(x, r, t)$ predicts average velocity over $[r, t]$, where $0 \le r \le t \le 1$. Writing sg for stop-gradient, we use its instantaneous-velocity regression prediction to analyze weighting (Geng et al., 2026),

$$\widehat{v}_\theta(x, r, t) = u_\theta(x, r, t) + (t - r)\,\mathrm{sg}[D_x u_\theta(x, r, t)\bar{v}_\theta(x, t) + \partial_t u_\theta(x, r, t)]\,, \tag{2}$$

where $\bar{v}_\theta(x, t) = \mathrm{sg}[u_\theta(x, t, t)]$. The bracket is a Jacobian-vector product (JVP) with r held fixed. The model-predicted tangent makes $\widehat{v}_\theta$ a function only of the model inputs, not of the sampled target Y. With the stopped correction fixed, varying u_θ varies $\widehat{v}_\theta$ by the same amount. We therefore analyze weighting in this regression space.

Fix all model inputs $Z = z$ (including r for iMF) and all stopped model quantities, and vary the regression prediction $f \in \mathbb{R}^d$ freely. With a positive detached scalar weight, the expected gradient of the squared regression loss with respect to f is proportional to

$$G_w(f; z) = \mathbb{E}[w(f, Y, z)(f - Y) \mid Z = z]. \tag{3}$$

Class conditioning can leave target variation within each class. A teacher target deterministic given the student's inputs has no such variation.

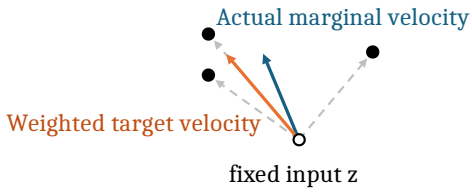
We study the inverse-residual rule used in MeanFlow (Geng et al., 2025a),

$$w_{\mathrm{adapt}}(f, Y) = \left(\|f - Y\|_2^2 + \varepsilon\right)^{-p}, \tag{4}$$

where $\varepsilon > 0$ and $p > 0$. The exponent acts on the squared norm, corresponding to $\|f - Y\|_2^{-2p}$ without the offset. For fixed targets, detached inverse-residual weighting gives, up to a positive constant factor, the same gradient with respect to f as a radial robust loss (Appendix A.2).

Normalizing the weight defines a conditional law $\mathbb{Q}_f$ proportional to $w(f, Y, z)\mathbb{P}(\mathrm{d}Y \mid z)$. At stationarity, $G_w(f_\star; z) = 0$ gives $f_\star = \mathbb{E}_{\mathbb{Q}_{f_\star}}[Y \mid z]$ (Proposition 2). The target values are unchanged, but their relative probabilities now depend on the prediction. Even with equally likely targets, this can shift a unique optimum away from the original mean (Figure 1a).

(a) Mean shift with a unique optimum



(b) A small initial bias can grow

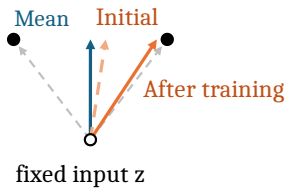


Figure 1: Adaptive weighting changes the relative influence of equally likely velocity targets (black dots) at a fixed input. (a) The unique regression optimum can shift toward the closely spaced pair. (b) A small initial bias can grow during training, even for symmetric targets. Appendix A.3 gives concrete examples of both effects.

Because the weights depend on the prediction, even a symmetric target law can have an unstable mean (Figure 1b). For equally likely $Y = \pm a$, $p = 1$, and $a^2 > \varepsilon$, $f = 0$ is stationary but unstable under the expected update, while $f = \pm\sqrt{a^2 - \varepsilon}$ are stable (Appendix A.3).

3.2 Total-variance characterization

The examples show how weighting can alter the selected mean. Does the selected law also have smaller total variance when measured around its own, potentially shifted mean?

Suppress the fixed conditioning state. Let $\mathbb{P}$ be the law of Y, let $f_\star$ be any stationary prediction under inverse-residual weighting, and define

$$R_\star = \|Y - f_\star\|_2^2, \qquad S_\star = R_\star + \varepsilon, \qquad \frac{\mathrm{d}\mathbb{Q}_\star}{\mathrm{d}\mathbb{P}} = \frac{S_\star^{-p}}{\mathbb{E}_{\mathbb{P}}[S_\star^{-p}]}. \tag{5}$$

Thus $f_\star = \mathbb{E}_{\mathbb{Q}_\star} Y$. Write

$$\mu = \mathbb{E}_{\mathbb{P}} Y, \qquad \delta = f_\star - \mu, \qquad T_{\mathbb{P}} = \operatorname{tr} \operatorname{Cov}_{\mathbb{P}}(Y), \qquad T_{\mathbb{Q}} = \operatorname{tr} \operatorname{Cov}_{\mathbb{Q}_\star}(Y). \tag{6}$$

Theorem 1 (Total variance under adaptive weighting)**.** *Assume $\mathbb{P}$ has a finite second moment and $\varepsilon > 0$. At every stationary prediction, if $0 < p \le 1$,*

$$T_{\mathbb{Q}} \le T_{\mathbb{P}}, \tag{7}$$

with strict inequality whenever $R_\star$ is nonconstant. The range is sharp: for every $p > 1$ and $\varepsilon > 0$, a two-point target law admits a strict local minimum of the expected radial potential for which $T_{\mathbb{Q}} > T_{\mathbb{P}}$.

The subtlety is that residuals and variances are measured around different centers. Define $D_\star = \mathbb{E}_{\mathbb{P}} R_\star - \mathbb{E}_{\mathbb{Q}_\star} R_\star$. Then

$$D_\star = \underbrace{T_{\mathbb{P}} - T_{\mathbb{Q}}}_{\text{decrease in total variance}} + \underbrace{\|\delta\|_2^2}_{\text{squared mean shift}} \tag{8}$$

Downweighting large residuals gives $D_\star \ge 0$, but this alone does not establish $T_{\mathbb{P}} \ge T_{\mathbb{Q}}$: the mean also moves.

At $p = 1$, the proof compares both quantities through the same change of measure:

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}_\star} = \frac{R_\star + \varepsilon}{T_{\mathbb{Q}} + \varepsilon}, \qquad \chi^2(\mathbb{P}\|\mathbb{Q}_\star) = \mathbb{E}_{\mathbb{Q}_\star}\left[\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}_\star} - 1\right)^2\right]. \tag{9}$$

Substitution and the Cauchy–Schwarz mean-shift bound (Chapman & Robbins, 1951) give

$$D_\star = (T_{\mathbb{Q}} + \varepsilon)\chi^2(\mathbb{P}\|\mathbb{Q}_\star), \qquad \|\delta\|_2^2 \le T_{\mathbb{Q}} \chi^2(\mathbb{P}\|\mathbb{Q}_\star). \tag{10}$$

Thus the residual reduction exceeds what can be accounted for by movement of the mean,

$$T_{\mathbb{P}} - T_{\mathbb{Q}} = D_\star - \|\delta\|_2^2 \ge \varepsilon \chi^2(\mathbb{P}\|\mathbb{Q}_\star) \ge 0. \tag{11}$$

This proves the $p = 1$ case. The moment argument for the full $0 < p \le 1$ range and the counterexamples above one are in Appendix A.4 and Proposition 8.

More generally, let s denote the squared residual and let $\eta(s)$ be a positive continuous radial weight. The variance inequality holds for every finitely supported target law and every stationary prediction if and only if

$$\eta(s) \text{ is nonincreasing}, \qquad s\eta(s) \text{ is nondecreasing}. \tag{12}$$

These conditions also imply the guarantee for laws with a finite second moment (Theorem 7). For $\eta(s) = (s + \varepsilon)^{-p}$ with $p > 0$, they hold exactly when $p \le 1$.

Equivalently, for $0 < s_1 < s_2$,

$$1 \le \frac{\eta(s_1)}{\eta(s_2)} \le \frac{s_2}{s_1}. \tag{13}$$

Larger residuals must receive no greater weight, but the relative downweighting cannot exceed their squared-residual ratio. Both restrictions are needed for a uniform variance guarantee.

The variance boundary $p = 1$ differs from the convexity boundary $p = 1/2$, above which the score redescends and the potential is nonconvex (Barron, 2019). The theorem still holds throughout $1/2 < p \le 1$ (Appendix A.2).

In particular, the two stable roots at $p = 1$ coexist with the variance guarantee. Nonincreasing selected variance does not guarantee stability of the original mean.

The theorem concerns the conditional target law, not the final generated distribution. A group's selected probability depends on its average weight given the input (Corollary 5), so its original frequency alone does not determine whether it gains influence. Changing the selected mean changes the CFM velocity. We examine the consequences of integrating that field in Section 5.

4 TARGET-PRESERVING PREDICTION WEIGHTING

A shared positive weight retains the conditional mean while allowing update strength to vary across inputs. This suggests *prediction weighting*, obtained by removing the sampled target from the weighting term. We replace the residual norm with the prediction norm while leaving the squared regression error unchanged. The rule requires no auxiliary network or additional weighting hyperparameters.

Using the same exponent p and offset ε as adaptive weighting, we set

$$w_{\text{pred}}(z) = \big(\text{sg}(\|\widehat{v}_\theta(z)\|_2^2) + \varepsilon\big)^{-p}, \tag{14}$$

where $\widehat{v}_\theta(z)$ is the regression prediction.

At a fixed state, the expected gradient factorizes:

$$G_{\text{pred}}(f; z) = w_{\text{pred}}(z)\big(f - \mu(z)\big). \tag{15}$$

Because the multiplier is shared across targets, the unique pointwise root remains $\mu(z)$ (Kim et al., 2025; Billera et al., 2025; Proposition 3). This is a regression-target guarantee, not a guarantee of exact generation by a finite shared network.

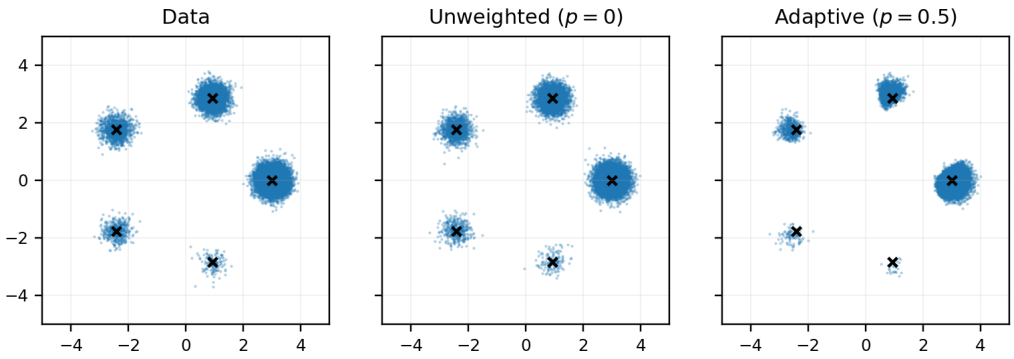
In iMF, both rules weight the same squared error $\|\widehat{v}_\theta - T\|_2^2$, where T is the velocity target, with guidance when present. Class and guidance conditions belong to the fixed state (Appendix B.3). One-call generation uses $x_0 = x_1 - u_\theta(x_1, 0, 1)$, not the corrected prediction $\widehat{v}_\theta$. Diagonal Euler integrates $u_\theta(x, t, t)$.

5 EXPERIMENTS

We compare $p = 0$, 0.5, and 1, extending CIFAR iMF to $p = 1.5$. Architecture, optimizer, and training budget are fixed across weighting rules within each setting.

5.1 POPULATION FLOWS ON A GAUSSIAN MIXTURE

We numerically approximate and integrate population velocity fields for a two-dimensional mixture of five Gaussians with imbalance ratio 100. Both fields start from identical Gaussian particles, without neural fitting.



Component mass (%)	p	1	2	3	4	5
Original law	–	68.594	21.691	6.859	2.169	.686
Unweighted	0	67.700	22.785	6.815	1.940	.760
Adaptive	0.5	85.800	10.770	2.790	.510	.130

Figure 2: Generated marginals from population fields at the finest tested resolution in Table 5. Crosses mark component means, ordered counterclockwise from $(3, 0)$. The table gives prescribed probabilities and nearest-mean assignments of the 20,000 samples shown per field, generated from identical starting particles.

Convex regression can change generated proportions. Adaptive weighting at $p = 0.5$ raises the largest component's mass from the prescribed 68.6% to 85.8% and reduces the smallest from $.686\%$ to $.130\%$ (Figure 2). The pointwise regression potential is convex, so these changes require neither neural fitting nor multiple stable predictions. Finite CFM and iMF models also suppress rare modes (Table 6).

Across the tested numerical resolutions, the largest component retains 84–86% of the generated mass. A sampled-$p = 0$ control follows analytic CFM closely, with a maximum component-mass difference of .025 percentage points at the displayed resolution (Table 5).

Conditional variance alone does not determine mean preservation. At a fixed input, we compare the mixture's conditional target law with its symmetrization. Under the original law, adaptive $p = 0.5$ and 1 shift the computed stationary prediction away from μ. Mixing the law equally with its reflection $Y \mapsto 2\mu - Y$ preserves its mean, covariance, and distribution of $\|Y - \mu\|_2$. Opposite residuals now receive equal weights, so their updates cancel at μ (Appendix B.1). Mean stationarity depends on the association between residual magnitudes and directions, not on variance alone.

5.2 One-Step Generation on a Spiral Distribution

We construct a noisy, deformed spiral from eight groups with geometrically decreasing sample counts (imbalance ratio 80) and embed it in a random two-dimensional subspace of $\mathbb{R}^{16}$. A six-block, width-256 MLP trains for 100,000 updates without group labels. We measure group proportions by Jensen–Shannon divergence (JS) and distributional fit by sliced Wasserstein distance (SW). Figure 3 compares one-call and 100-step Euler samples; Appendix B.2 and Table 7 give construction and quantitative details.

Unweighted one-step learning remains difficult. Even on this simple distribution, unweighted iMF visibly blurs the spiral in one call, whereas FM with 100-step Euler recovers its geometry closely. Prediction weighting improves one-call SW from .0174 to .0096 at $p = 0.5$ under the same training budget. The correct pointwise target alone is therefore insufficient for accurate one-step generation at this budget.

Adaptive weighting suppresses rare groups. The two rarest groups lose mass as p increases. Their one-call mass falls from .0161 at $p = 0$ to .0076 at $p = 0.5$ and .0046 at $p = 1$, against a target mass of .01675. Prediction weighting retains their mass more closely than adaptive weighting. This contrast is consistent with the distinction between selecting conditional targets and rescaling their shared update. The loss of minority modes in these simple distributions also aligns with the observations of Lin et al. (2026).

Good Euler samples need not imply good one-call samples. Adaptive $p = 1$ also degrades diagonal Euler samples. Applying this weight only to nonzero time intervals ($r < t$), while leaving diagonal examples ($r = t$) unweighted, recovers Euler quality close to the unweighted baseline, but one-call SW remains about three times higher (Figure 3 and Table 7). Preserving diagonal regression does not remove the effect of adaptive weighting on off-diagonal regression.

5.3 Unconditional Generation on CIFAR-10(-LT100)

CIFAR-10 provides an unconditional comparison across weighting strength, data balance, and sampling budget. We train FM and iMF (Geng et al., 2026) with a 55.9M-parameter SongUNet on balanced CIFAR-10 and its imbalance-100 split. Repeated divergence in preliminary unweighted iMF runs at MeanFlow's original learning rate of 6×10^{-4} (Geng et al., 2025a) led us to use 10^{-4} for all rules. Table 1 reports matched-reference metrics, with class probabilities estimated by a fixed classifier (Appendix B.4).

Similar FID can accompany different class proportions. On CIFAR-10-LT100, unweighted iMF and adaptive $p = 0.5$ achieve similar matched FIDs (29.85 and 29.16). Yet adaptive weighting reduces the estimated share of the three most frequent training classes from 80.12% to 67.85%, compared with 78.93% in the data (Figure 4). This redistribution occurs even as overall class-prior error improves slightly (Table 1).

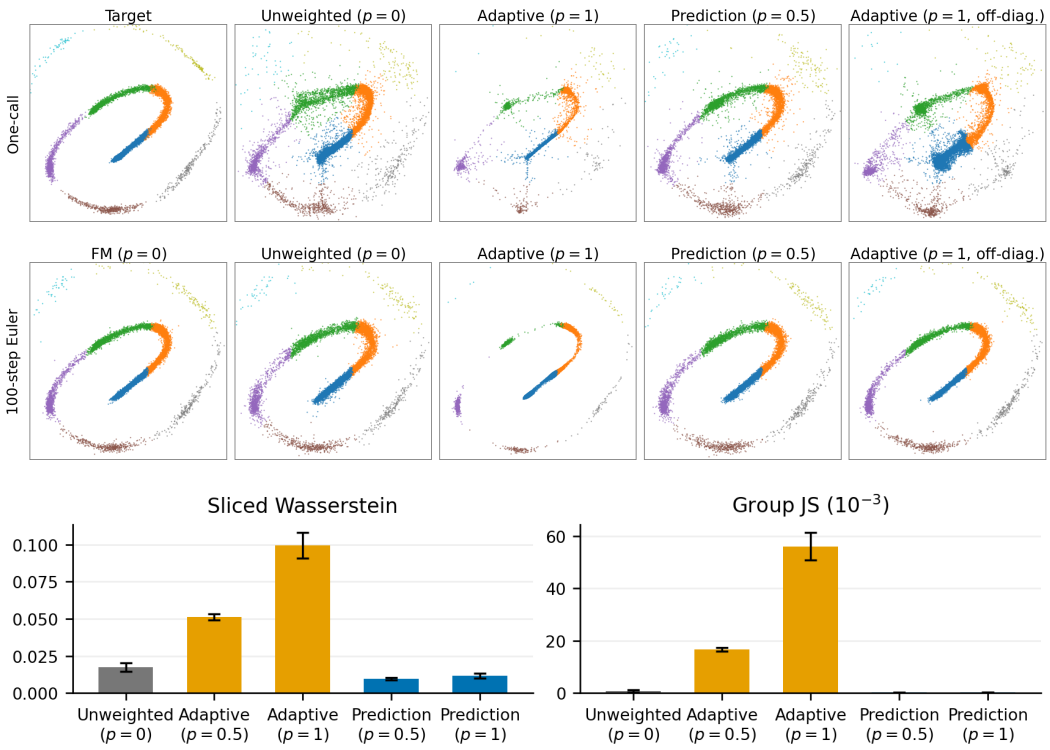


Figure 3: Generation on the spiral distribution. The first row shows the target and one-call iMF samples. The second shows FM and diagonal iMF with 100-step Euler. Colors indicate reference groups. Off-diagonal adaptive weighting uses $p = 1$ for $r < t$ and $p = 0$ for $r = t$. Below, one-call SW and group JS summarize the weighting comparison over three seeds (mean $\pm$ standard deviation).

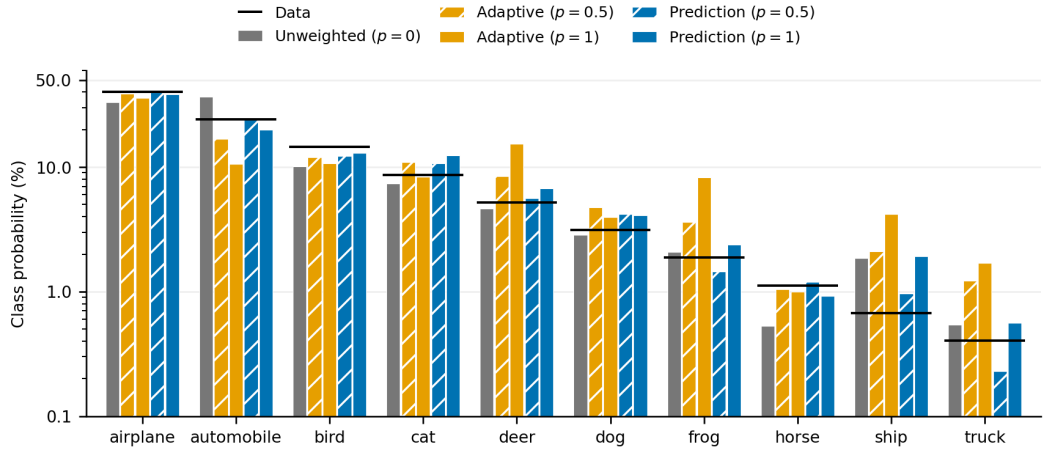


Figure 4: Class composition for one-call iMF after CIFAR-10-LT100 training at 200k updates. The logarithmic axis shows both common and rare classes, with black marks indicating training probabilities. Estimates use 50,000 samples, a fixed classifier, and test-confusion correction.

The change need not favor frequent classes. Increasing adaptive weighting to $p = 1$ worsens both matched FID and prior error on LT training. Its estimated tail mass rises to 6.90%, compared with 2.19% in the data. This is the opposite direction from rare-mode suppression in the Gaussian and spiral examples, so frequency alone does not explain which classes gain influence.

Table 1: CIFAR-10 at 200k updates. Each model generates 50,000 samples, evaluated against its training distribution. FM uses 50-step Euler; iMF uses one call. DINO-FID uses ViT-S/14. Prior error is $100\times$ confusion-corrected class-prior MAE. Lower is better; bold marks the best result within each model family. Cross-reference results are in Tables 8 and 9.

		LT training			BAL training		
Training rule	p	FID	DINO-FID	Prior error	FID	DINO-FID	Prior error
FM, Unweighted	0	55.42	630.87	**4.69**	**8.55**	**184.06**	**1.21**
FM, Adaptive	0.5	68.55	721.80	6.23	9.96	209.81	1.69
FM, Adaptive	1	**49.44**	**617.83**	5.83	11.88	243.95	2.30
iMF, Unweighted	0	29.85	499.09	2.80	16.48	352.76	2.43
iMF, Adaptive	0.5	29.16	494.16	2.23	9.73	210.16	.78
iMF, Adaptive	1	53.05	748.23	4.44	12.31	245.76	1.71
iMF, Adaptive	1.5	57.15	825.78	4.43	18.11	345.79	2.69
iMF, Prediction	0.5	**18.38**	**347.36**	**.78**	8.99	208.61	.68
iMF, Prediction	1	23.22	413.92	1.63	8.35	200.25	**.47**
iMF, Prediction	1.5	26.92	452.53	2.23	**7.92**	**199.64**	.53

Prediction weighting improves fit and prior agreement. Prediction weighting achieves the lowest Inception FID and prior error in both training settings (Table 1). Its $p = 0.5$ model reaches LT FID 18.38 in one call, compared with 29.85 for unweighted iMF and 55.42 for unweighted FM with 50-step Euler. The LT advantage at $p = 0.5$ also holds with four interval calls and diagonal Euler (Table 10). Matched DINO-FID favors prediction over adaptive iMF at each exponent, although unweighted FM has the lowest BAL DINO-FID. Relative to adaptive weighting, its LT precision is lower but recall is higher at $p = 0.5$ and 1 (Table 14).

The variance boundary is not a performance threshold. Adaptive $p = 1.5$ lies outside the variance guarantee (Theorem 1), not beyond a predicted transition in stability or FID. Increasing p from 1 to 1.5 worsens its one-call FID on both datasets, yet improves LT diagonal-Euler FID from 42.03 to 26.97 (Table 11). Prediction weighting improves balanced FID to 7.92 but worsens LT FID. Its target preservation has no $p \leq 1$ restriction.

CIFAR learning curves show similar early FID reduction but a marked late-stage LT FID increase in unweighted training, while prediction weighting remains steadier (Figure 6).

5.4 Class-Conditional Generation on ImageNet-LT

ImageNet-LT tests 256×256 class-conditional generation with guidance, beyond the unconditional CIFAR study. We adapt pMF's recipe (Lu et al., 2026) with a 118M-parameter pixel-space DiT-B/16, retaining its optimizer and classifier-free guidance (CFG) but omitting the auxiliary velocity head and using a single EMA. The initial 300k updates use velocity regression alone. Requested label frequencies are balanced or LT-prior, evaluated against matching references at guidance scale 8.5. Settings developed with adaptive weighting are shared without separate retuning (Appendix B.5).

Adaptive weighting leads under the transferred recipe. Prediction weighting improves over unweighted iMF, but adaptive weighting achieves lower FID and DINO-FID, reversing the CIFAR ordering (Table 2). Its FID advantage extends across head, body, and tail classes and persists without guidance (Table 12 and Appendix B.5). Pointwise target preservation need not yield the best finite-budget FID. At 300k, unweighted FM and prediction-weighted iMF have higher precision but lower recall than adaptive-weighted iMF, unlike the CIFAR-10-LT100 comparison (Appendix B.7).

Table 2: ImageNet-LT at 300k without perceptual losses. Val-BAL and Train-LT use balanced and LT-prior labels with validation and training-derived references, respectively. NFE counts network evaluations. DINO-FID uses DINOv2 ViT-L/14. Bold marks minima within each model family.

			Val-BAL		Train-LT	
Training rule	p	NFE	FID	DINO-FID	FID	DINO-FID
FM, Unweighted	0	50	42.45	913.84	41.24	908.60
FM, Adaptive	1	50	**18.88**	**531.91**	**15.66**	**482.96**
iMF, Unweighted	0	1	62.54	1055.43	65.98	1082.85
iMF, Adaptive	0.5	1	33.33	738.59	32.08	719.45
iMF, Adaptive	1	1	**30.71**	**662.80**	**28.60**	**633.58**
iMF, Prediction	0.5	1	51.04	950.73	51.74	958.21
iMF, Prediction	1	1	49.79	935.11	50.88	943.40

Perceptual continuation substantially narrows the gap. We add pMF's LPIPS and ConvNeXt losses to improve FID through pretrained feature similarity (Lu et al., 2026). Target preservation does not cover these feature losses. After 100k further updates with the same architecture and regression weighting rule, prediction weighting narrows the Val-BAL FID gap to 10.60 versus adaptive weighting's 9.64. FM and unweighted iMF gain little (Table 3). DINO-FID, precision, and recall still favor adaptive weighting (Appendix B.7). Protocol and unguided results are in Appendix B.6.

Table 3: ImageNet-LT at 400k after 100k perceptual updates. Samplers and label priors match Table 2, but DINO-FID uses ViT-S/14, not ViT-L/14. Lower is better.

		Val-BAL		Train-LT	
Training rule	p	FID	DINO-FID	FID	DINO-FID
FM, Unweighted	0	42.58	707.87	41.43	675.44
iMF, Unweighted	0	58.34	907.97	61.91	894.59
iMF, Adaptive	1	**9.64**	**258.95**	**8.05**	**231.01**
iMF, Prediction	1	10.60	317.92	9.71	291.63

5.5 Direct Flow Matching on CIFAR-10 and ImageNet-LT

Direct FM removes JVPs and average-velocity learning. Unweighted and adaptive $p = 1$ runs share per-dataset settings, uniform times, and 50-step Euler (Tables 1 and 2).

Weighting effects are not specific to JVP training. Adaptive weighting improves FID on ImageNet-LT and CIFAR-10-LT100 but worsens it on balanced CIFAR. On LT100, FID improves from 55.42 to 49.44 while prior error increases from 4.69 to 5.83. This divergence between quality and class-prior agreement requires neither JVPs nor average-velocity learning, though target selection and optimization remain coupled.

6 Conclusion

Adaptive weighting in MeanFlow changes not only update strength but also which conditional targets the model learns to average. We characterize the selected conditional law through sharp conditions for nonincreasing total variance and demonstrate changes in generated proportions. Prediction weighting provides a simple alternative that retains adaptive scaling while preserving the pointwise regression target. Understanding when target selection helps or harms generation remains an open question. Our results support treating adaptive weighting as a generative modeling choice, evaluated for the targets it selects as well as its optimization benefits.

AI USE STATEMENT

Through iterative discussions with ChatGPT and Codex, the authors drafted and refined the manuscript and explored theoretical and experimental questions. These exchanges supported literature searches, experimental design and implementation, result interpretation, and the development of proofs and counterexamples. The authors evaluated and revised the suggestions and reviewed the resulting text, code, and mathematical arguments, taking responsibility for the final work.

REPRODUCIBILITY STATEMENT

Appendix A gives complete proofs and assumptions. Appendix B specifies data construction, training, sampling, evaluation, and numerical-resolution checks. Additional generated samples appear in Appendix C.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL https://www.jmlr.org/papers/v26/23-1605.html.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005. URL https://www.jmlr.org/papers/v6/banerjee05b.html.

Jonathan T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1701.03077.

Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.03719.

Lukas Billera, Hedwig Nora Nordlinder, and Ben Murrell. Time dependent loss reweighting for flow matching and diffusion models is theoretically justified. *arXiv preprint arXiv:2511.16599*, 2025. URL https://arxiv.org/abs/2511.16599.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=cqDH0e6ak2.

Sergio Calvo-Ordoñez, Matthieu Meunier, Álvaro Cartea, Christoph Reisinger, Yarin Gal, and José Miguel Hernández-Lobato. Weighted conditional flow matching. *arXiv preprint arXiv:2507.22270*, 2025. URL https://arxiv.org/abs/2507.22270.

Douglas G. Chapman and Herbert Robbins. Minimum variance estimation without regularity assumptions. *The Annals of Mathematical Statistics*, 22(4):581–586, 1951. doi: 10.1214/aoms/1177729548.

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*, volume 38, pp. 75460–75482, 2025a. URL https://arxiv.org/abs/2505.13447.

Zhengyang Geng, Ashwini Pokle, Weijian Luo, Justin Lin, and J. Zico Kolter. Consistency models made easy. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=xQVxo9dSID.

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*, pp. 30467–30476, 2026. URL https://arxiv.org/abs/2512.02012.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138.

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. URL https://arxiv.org/abs/1512.03385.

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*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary Markov processes. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.20587.

Paul W. Holland and Roy E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6(9):813–827, 1977. doi: 10.1080/03610927708827533.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.10196.

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*, volume 35, pp. 26565–26577, 2022. URL https://arxiv.org/abs/2206.00364.

Beomsu Kim, Yu-Guan Hsieh, Michal Klein, Marco Cuturi, Jong Chul Ye, Bahjat Kawar, and James Thornton. Simple ReFlow: Improved techniques for fast flow models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07815.

Jin-Young Kim, Hyojun Go, Lea Bogensperger, Julius Erbach, Nikolai Kalischek, Federico Tombari, Konrad Schindler, and Dominik Narnhofer. Understanding, accelerating, and improving MeanFlow training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37992–38003, 2026. URL https://arxiv.org/abs/2511.19065.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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*, volume 32, 2019. URL https://arxiv.org/abs/1904.06991.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2405.20320.

Zhiqi Li, Yuchen Sun, Duowen Chen, Jinjin He, and Bo Zhu. Trajectory consistency for one-step generation on Euler Mean Flows. *arXiv preprint arXiv:2602.02571*, 2026. URL https://arxiv.org/abs/2602.02571.

Yexiong Lin, Jia Shi, Shanshan Ye, Wanyu Wang, Yu Yao, and Tongliang Liu. SubFlow: Sub-mode conditioned flow matching for diverse one-step generation. *arXiv preprint arXiv:2604.12273*, 2026. URL https://arxiv.org/abs/2604.12273.

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. URL https://arxiv.org/abs/2210.02747.

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. URL https://openreview.net/forum?id=XVjTT1nw5z.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019. URL https://arxiv.org/abs/1904.05160.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.11081.

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. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.22158.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. 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. URL https://arxiv.org/abs/2304.07193.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://arxiv.org/abs/2212.09748.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.00512.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.14189.

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. URL https://arxiv.org/abs/2011.13456.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pp. 32211–32252, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0bc795afae289ed465a65a3b4b1f4eb7-Abstract.html.

Ryan Turner, Jane Hung, Eric Frank, Yunus Saatchi, and Jason Yosinski. Metropolis-Hastings generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 6345–6353, 2019. URL https://proceedings.mlr.press/v97/turner19a.html.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023. URL https://arxiv.org/abs/2301.00808.

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. URL https://arxiv.org/abs/2602.05435.

Alexander Zaytsev, Dmitry Baranchuk, Alexander Korotin, and Aibek Alanov. Rethinking Pixel Mean Flows via Interval Denoiser. *arXiv preprint arXiv:2608.04818*, 2026. URL https://arxiv.org/abs/2608.04818.

Huijie Zhang, Aliaksandr Siarohin, Willi Menapace, Michael Vasilkovsky, Sergey Tulyakov, Qing Qu, and Ivan Skorokhodov. AlphaFlow: Understanding and improving MeanFlow models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=adacb4JTIv.

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 Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://arxiv.org/abs/1801.03924.

A Proofs

A.1 Weighted-target representation and preservation

Proposition 2 (Distribution selected by a stopped weight)**.** *Fix a state $Z = z$ and all stopped model quantities. Assume the relevant first moments and the normalizer $\mathbb{E}[w(f, Y, z) \mid z]$ are finite, with the latter positive. Define*

$$\frac{\mathrm{d}\mathbb{Q}_f(\cdot \mid z)}{\mathrm{d}\mathbb{P}(\cdot \mid z)}(Y) = \frac{w(f, Y, z)}{\mathbb{E}[w(f, Y, z) \mid z]}. \tag{16}$$

Every pointwise root satisfies

$$f = \mathbb{E}_{\mathbb{Q}_f}[Y \mid z]. \tag{17}$$

Proof of Proposition 2. Since gradients are stopped through w,

$$G_w(f; z) = \mathbb{E}[w(f, Y, z)(f - Y) \mid z]. \tag{18}$$

Dividing the root equation by the positive normalizer gives

$$f = \frac{\mathbb{E}[w(f, Y, z)Y \mid z]}{\mathbb{E}[w(f, Y, z) \mid z]} = \mathbb{E}_{\mathbb{Q}_f}[Y \mid z]. \tag{19}$$

□

For a teacher target $Y = m(z)$ almost surely, both conditional laws in Equation (16) are the same point mass, and the root remains $m(z)$. The condition is determinism given the student's input. Teacher-based training with remaining target randomness still falls under the general analysis.

The weighted-target representation also holds for target-first Bregman losses. Let

$$D_\Phi(Y, f) = \Phi(Y) - \Phi(f) - \langle \nabla\Phi(f), Y - f \rangle, \tag{20}$$

where Φ is twice continuously differentiable and $\nabla^2\Phi(f) \succ 0$. Since

$$\nabla_f D_\Phi(Y, f) = \nabla^2\Phi(f)(f - Y), \tag{21}$$

the stopped weighted score is $\nabla^2\Phi(f)G_w(f; z)$ and has the same roots (Banerjee et al., 2005; Gneiting, 2011).

Proposition 3 (Prediction weighting leaves the target unchanged)**.** *Suppose that, after conditioning on z and freezing stopped model quantities, $\widehat{v}_\theta(z)$ is independent of the remaining target randomness. Then the weight in Equation (14) gives*

$$\mathbb{Q}_{\mathrm{pred}}(\mathrm{d}y \mid z) = \mathbb{P}(\mathrm{d}y \mid z), \qquad \mathbb{E}_{\mathbb{Q}_{\mathrm{pred}}}[Y \mid z] = \mathbb{E}_{\mathbb{P}}[Y \mid z]. \tag{22}$$

This preserves the marginal velocity in direct CFM and the conditional-mean target in iMF's corrected velocity-regression space.

Proof. The multiplier is constant across target values at fixed z, so it cancels from the normalized conditional law in Equation (16). Equivalently, the expected update is $w_{\mathrm{pred}}(z)(f - \mathbb{E}[Y \mid z])$, whose unique root is the original conditional mean. □

Corollary 4 (When the original conditional mean is retained)**.** *Let $\mu = \mathbb{E}[Y \mid z]$. Then μ is a root if and only if*

$$\mathrm{Cov}(w(\mu, Y, z), Y \mid z) = 0. \tag{23}$$

In particular, a positive multiplier that is constant across targets at fixed z leaves the conditional law unchanged. Conversely, fix f and z. If $y \mapsto w(f, y, z)$ is continuous on an open convex domain containing f, and f is a root for every finitely supported target law on that domain with mean f, then $w(f, y, z)$ is independent of y on the domain.

Proof of Corollary 4. At $f = \mu$,

$$\mathbb{E}[w(\mu, Y, z)(\mu - Y) \mid z] = -\mathrm{Cov}(w(\mu, Y, z), Y \mid z), \tag{24}$$

because $\mathbb{E}[\mu - Y \mid z] = 0$. This proves (23). If $w(f, z)$ is target-independent, the score is $w(f, z)(f - \mu)$ and has unique root μ.

For the converse, translate f to zero and suppress z. For any nonzero y in the domain, openness places $-sy$ in the domain for all sufficiently small $s > 0$. The law with masses $s/(1+s)$ at y and $1/(1+s)$ at $-sy$ has mean zero. Its weighted stationarity implies $w(y) = w(-sy)$. Letting $s \downarrow 0$ and using continuity gives $w(y) = w(0)$ throughout the domain. □

Corollary 5 (Change in group influence)**.** *Let C be a group jointly distributed with Y under $\mathbb{P}$, and extend $\mathbb{Q}_\star$ by the adaptive weight in Equation (5). For groups c, d with positive probability,*

$$\frac{\mathbb{Q}_\star(C = c)/\mathbb{Q}_\star(C = d)}{\mathbb{P}(C = c)/\mathbb{P}(C = d)} = \frac{\mathbb{E}_{\mathbb{P}}[S_\star^{-p} \mid C = c]}{\mathbb{E}_{\mathbb{P}}[S_\star^{-p} \mid C = d]}. \tag{25}$$

Proof. The selected group probability is $\mathbb{Q}_\star(C = c) = \mathbb{P}(C = c)\mathbb{E}_{\mathbb{P}}[S_\star^{-p} \mid C = c]/\mathbb{E}_{\mathbb{P}}[S_\star^{-p}]$. Taking the ratio for two groups cancels the normalizer. □

A.2 Robust-loss geometry

Residual-dependent reweighting has a classical interpretation in robust regression (Huber, 1964; Holland & Welsch, 1977). For fixed Y, the detached inverse-residual score is $x(\|x\|^2 + \varepsilon)^{-p}$, where $x = f - Y$. It is the gradient with respect to f of the radial potential

$$\rho_p(x) = \begin{cases} \dfrac{(\|x\|^2 + \varepsilon)^{1-p} - \varepsilon^{1-p}}{2(1-p)}, & p \neq 1, \\ \dfrac{1}{2}\log(1 + \|x\|^2/\varepsilon), & p = 1. \end{cases} \tag{26}$$

The $p = 1/2$ and $p = 1$ cases are the pseudo-Huber (Charbonnier) and Cauchy forms, respectively, up to positive scale factors (Barron, 2019). Writing $r = \|x\|^2$, the score magnitude and radial Hessian eigenvalue are

$$\frac{\sqrt{r}}{(r + \varepsilon)^p}, \qquad \frac{\varepsilon + (1 - 2p)r}{(r + \varepsilon)^{p+1}}. \tag{27}$$

The score is bounded for $p \geq 1/2$ and redescends for $p > 1/2$. The potential is convex for $p \leq 1/2$ and nonconvex above it.

A.3 THE TWO FIXED-INPUT EXAMPLES

We give concrete examples of the two effects illustrated in Figure 1. For a unique optimum shifted away from the original mean, let Y be uniform on $(-1, .1)$, $(-1, -.1)$, and $(2, 0)$, with $p = 1/2$ and $\varepsilon = .01$. The Hessian of each potential is positive definite because its radial eigenvalue is $\varepsilon/(\|f - Y\|^2 + \varepsilon)^{3/2} > 0$. The expected potential is strictly convex and coercive, hence has a unique minimum. Symmetry puts that minimum on the horizontal axis. Its first coordinate solves

$$\frac{2(x+1)}{\sqrt{(x+1)^2 + .02}} + \frac{x-2}{\sqrt{(x-2)^2 + .01}} = 0. \tag{28}$$

Numerical root finding gives $x = -.9184141770$ and selected probabilities $(.48640236, .48640236, .02719527)$, rather than $(1/3, 1/3, 1/3)$. The shifted mean is therefore a consequence of the objective, not an alternative local optimum.

For the symmetric two-target example, let $\mathbb{P}(Y = -a) = \mathbb{P}(Y = a) = 1/2$ and $p = 1$. The expected detached score is

$$G(f) = \frac{1}{2}\left(\frac{f-a}{(f-a)^2 + \varepsilon} + \frac{f+a}{(f+a)^2 + \varepsilon}\right) \tag{29}$$

$$= \frac{f(f^2 + \varepsilon - a^2)}{((f-a)^2 + \varepsilon)((f+a)^2 + \varepsilon)}. \tag{30}$$

When $a^2 > \varepsilon$, its roots are 0 and $\pm\sqrt{a^2 - \varepsilon}$. At zero, $G'(0) = (\varepsilon - a^2)/(a^2 + \varepsilon)^2 < 0$. At either nonzero root the numerator derivative is $2(a^2 - \varepsilon) > 0$, and the denominator is positive. Under $\dot{f} = -G(f)$, zero is unstable and the two nonzero roots are stable; the same local classification holds for sufficiently small gradient-descent steps. The sign of a nonzero initial prediction selects its limiting root. At either stable root the selected variance is $a^2 - f_\star^2 = \varepsilon$, compared with original variance a^2. Multiple stable predictions thus coexist with the variance guarantee at $p = 1$.

A.4 INVERSE-POWER VARIANCE IDENTITIES

We bound the mean shift in Theorem 1 and derive the variance guarantee for $0 < p \leq 1$.

Proposition 6 (Additional inverse-power identities)**.** *Under the assumptions of Theorem 1, additionally assume $\mathbb{E}_{\mathbb{P}}[S_\star^p] < \infty$ when $p > 1$. Then*

$$D_\star = (T_{\mathbb{Q}} + \varepsilon)\,\chi^2(\mathbb{P}\|\mathbb{Q}_\star) + \Delta_p, \tag{31}$$

where

$$\chi^2(\mathbb{P}\|\mathbb{Q}_\star) = \mathbb{E}_{\mathbb{Q}_\star}\left[\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}_\star} - 1\right)^2\right] \tag{32}$$

and, with all moments below taken under $\mathbb{Q}_\star$,

$$\Delta_p = \frac{\mathbb{E}[S_\star^p]\mathbb{E}[S_\star^{p+1}] - \mathbb{E}[S_\star]\mathbb{E}[S_\star^{2p}]}{\mathbb{E}[S_\star^p]^2}. \tag{33}$$

Unless $R_\star$ is constant,

$$\operatorname{sgn}(\Delta_p) = \operatorname{sgn}(1 - p). \tag{34}$$

With $\Sigma_{\mathbb{Q}_\star} = \operatorname{Cov}_{\mathbb{Q}_\star}(Y)$, the classical Chapman–Robbins bound (Chapman & Robbins, 1951), applied in each direction, gives

$$\delta\delta^\top \preceq \chi^2(\mathbb{P}\|\mathbb{Q}_\star)\,\Sigma_{\mathbb{Q}_\star}. \tag{35}$$

Consequently, for $0 < p \leq 1$,

$$T_{\mathbb{P}} - T_{\mathbb{Q}} \geq \left[\varepsilon + T_{\mathbb{Q}} - \lambda_{\max}(\Sigma_{\mathbb{Q}_\star})\right]\chi^2(\mathbb{P}\|\mathbb{Q}_\star) \tag{36}$$

$$\geq \varepsilon\,\chi^2(\mathbb{P}\|\mathbb{Q}_\star). \tag{37}$$

At $p = 1$,

$$D_\star = \frac{\operatorname{Var}_{\mathbb{Q}_\star}(R_\star)}{T_{\mathbb{Q}} + \varepsilon} = (T_{\mathbb{Q}} + \varepsilon)\chi^2(\mathbb{P}\|\mathbb{Q}_\star). \tag{38}$$

Proof of Theorem 1 and Proposition 6. Write $R = R_\star$, $S = R+\varepsilon$, and $w = S^{-p}$. The root equation gives $\mathbb{E}_{\mathbb{Q}_\star} Y = f_\star$, hence

$$T_{\mathbb{Q}} = \mathbb{E}_{\mathbb{Q}_\star} R, \qquad T_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}} R - \|f_\star - \mu\|_2^2 . \tag{39}$$

Rearrangement proves (8). Since w is decreasing in R,

$$D_\star = -\frac{\mathrm{Cov}_{\mathbb{P}}(w, R)}{\mathbb{E}_{\mathbb{P}} w} \geq 0, \tag{40}$$

with equality exactly for constant R.

Let $H = S^p$ and $m = \mathbb{E}_{\mathbb{Q}_\star} H$. Direct change of measure gives

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}_\star} = \frac{H}{m}. \tag{41}$$

Set $X = Y - f_\star$, so $\mathbb{E}_{\mathbb{Q}_\star} X = 0$. Then

$$m^2(T_{\mathbb{P}} - T_{\mathbb{Q}}) = m \,\mathrm{Cov}_{\mathbb{Q}_\star}(H, S) - \|\mathbb{E}_{\mathbb{Q}_\star}[HX]\|_2^2 . \tag{42}$$

Matrix Cauchy–Schwarz gives

$$\mathbb{E}[HX]\mathbb{E}[HX]^\top \preceq \mathrm{Var}(H)\,\mathrm{Cov}(X). \tag{43}$$

Because $\mathbb{E}[HX] = -m\delta$ and $\mathrm{Var}(H) = m^2\chi^2(\mathbb{P}\|\mathbb{Q}_\star)$, this proves (35).

Let $M(q) = \mathbb{E}_{\mathbb{Q}_\star}[S^q]$ and $\phi(q) = \log M(q)$. Convexity of ϕ implies that $\phi(c-a) + \phi(c+a)$ is nondecreasing in a. Comparing exponent pairs $(p, p+1)$ and $(1, 2p)$, both centered at $c = p + 1/2$, yields

$$\mathrm{sgn}\big(M(p)M(p+1) - M(1)M(2p)\big) = \mathrm{sgn}(1-p) \tag{44}$$

for nonconstant S. Adding and subtracting $\mathbb{E}[S]\,\mathrm{Var}(H)$ in (42) gives (31)–(34).

For $0 < p \leq 1$, $\Delta_p \geq 0$. Taking the largest eigenvalue in (35) gives

$$\|\delta\|_2^2 \leq \lambda_{\max}(\Sigma_{\mathbb{Q}_\star})\chi^2(\mathbb{P}\|\mathbb{Q}_\star). \tag{45}$$

Combining this inequality with (8) and (31) proves (36). The bound $\lambda_{\max}(\Sigma_{\mathbb{Q}_\star}) \leq T_{\mathbb{Q}}$ gives (37).

At $p = 1$, $H = S$ and $m = T_{\mathbb{Q}} + \varepsilon$. Equation (41) gives

$$\mathrm{Var}_{\mathbb{Q}_\star}(R) = m^2\chi^2(\mathbb{P}\|\mathbb{Q}_\star), \tag{46}$$

and $\Delta_1 = 0$, proving (38). The strict-local-minimum construction in Proposition 8 proves sharpness above $p = 1$. □

A.5 General radial characterization

Theorem 7 (Radial weights with a uniform variance guarantee)**.** *Let $\eta : [0,\infty) \to (0,\infty)$ be continuous. For a target law $\mathbb{P}$ and a stationary prediction a satisfying*

$$\mathbb{E}_{\mathbb{P}}\Big[\eta(\|Y-a\|_2^2)(a-Y)\Big] = 0, \tag{47}$$

define

$$\frac{\mathrm{d}\mathbb{Q}_a}{\mathrm{d}\mathbb{P}} = \frac{\eta(\|Y-a\|_2^2)}{\mathbb{E}_{\mathbb{P}}[\eta(\|Y-a\|_2^2)]}. \tag{48}$$

The following are equivalent:

(i) For every finitely supported $\mathbb{P}$ on every $\mathbb{R}^d$ and every stationary a,

$$\mathrm{tr}\,\mathrm{Cov}_{\mathbb{Q}_a}(Y) \leq \mathrm{tr}\,\mathrm{Cov}_{\mathbb{P}}(Y). \tag{49}$$

(ii) $\eta(r)$ is nonincreasing and $r\eta(r)$ is nondecreasing on $[0,\infty)$.

When (ii) holds, the inequality also holds for every target law with a finite second moment. If both monotonicities are strict on the essential range of the squared residual, the inequality is strict whenever that squared residual is nonconstant.

Proof of Theorem 7. Let $X = Y - a$, $R = \|X\|_2^2$, $h = 1/\eta$, and $c = \mathbb{E}_{\mathbb{Q}_a}[h(R)]$. Since $a = \mathbb{E}_{\mathbb{Q}_a} Y$, we have $\mathbb{E}_{\mathbb{Q}_a} X = 0$ and

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}_a} = \frac{h(R)}{c}. \tag{50}$$

Write $m = \mathbb{E}_{\mathbb{Q}_a} R$. Direct calculation gives the exact identity

$$\begin{aligned}\operatorname{tr}\operatorname{Cov}_{\mathbb{P}}(Y) - \operatorname{tr}\operatorname{Cov}_{\mathbb{Q}_a}(Y) = \frac{m \operatorname{Var}_{\mathbb{Q}_a}(h(R)) - \|\operatorname{Cov}_{\mathbb{Q}_a}(h(R), X)\|_2^2}{c^2} \\ + \operatorname{Cov}_{\mathbb{P}}\left(h(R), \frac{R}{h(R)}\right).\end{aligned} \tag{51}$$

The first term is nonnegative by vector Cauchy–Schwarz. If η is nonincreasing and $r\eta(r)$ is nondecreasing, then $h(r)$ and $r/h(r)$ are both nondecreasing. The second term is therefore nonnegative by the covariance identity for comonotone functions. Strict monotonicity gives strictness for nonconstant R. These conditions also give $\eta(r) \le \eta(0)$ and $h(r) \le h(1)(1 + r)$, so every term is finite under a finite second moment of $\mathbb{P}$.

For necessity, choose in one dimension a mean-zero law $\mathbb{Q}$ with

$$\mathbb{Q}(X = -u) = \frac{v}{u+v}, \qquad \mathbb{Q}(X = v) = \frac{u}{u+v}, \qquad 0 < u < v, \tag{52}$$

and define $\mathbb{P}$ by $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q} \propto h(X^2)$. Then $a = 0$ is stationary under $\mathbb{P}$. Writing $h_u = h(u^2)$ and $h_v = h(v^2)$, direct calculation gives

$$\operatorname{Var}_{\mathbb{P}}(X) - \operatorname{Var}_{\mathbb{Q}}(X) = \frac{uv(h_u - h_v)(u^2 h_v - v^2 h_u)}{(v h_u + u h_v)^2}. \tag{53}$$

The expression is nonnegative exactly when $1 \le h_v/h_u \le v^2/u^2$. Requiring this for every $0 < u < v$ makes h and $r/h(r)$ nondecreasing, equivalently η nonincreasing and $r\eta(r)$ nondecreasing. Continuity extends the conclusion to $r = 0$. □

A.6 Failure above $p = 1$

Proposition 8 (Failure of a uniform guarantee for $p > 1$)**.** *For every $p > 1$ and $\varepsilon > 0$, there exists a two-point target law and a strict local minimum $f_\star$ of the corresponding expected radial potential such that the target distribution selected by adaptive weighting has larger total variance than the original law.*

Proof. Choose $0 < a < \sqrt{\varepsilon/(2p-1)}$ and $L > a$. Let

$$\mathbb{Q}_L\{-a\} = \frac{L}{L+a}, \qquad \mathbb{Q}_L\{L\} = \frac{a}{L+a}, \tag{54}$$

and define

$$\frac{\mathrm{d}\mathbb{P}_L}{\mathrm{d}\mathbb{Q}_L} = \frac{(Y^2+\varepsilon)^p}{\mathbb{E}_{\mathbb{Q}_L}[(Y^2+\varepsilon)^p]}. \tag{55}$$

Reweighting $\mathbb{P}_L$ by $(Y^2 + \varepsilon)^{-p}$ recovers $\mathbb{Q}_L$, whose mean is zero, so $f_\star = 0$ is stationary. As $L \to \infty$,

$$\frac{\operatorname{Var}_{\mathbb{P}_L}(Y)}{\operatorname{Var}_{\mathbb{Q}_L}(Y)} \to 0. \tag{56}$$

The second derivative of the expected potential at zero has the sign of

$$\mathbb{E}_{\mathbb{Q}_L}\left[1 - \frac{2pY^2}{Y^2+\varepsilon}\right], \tag{57}$$

which converges to $1 - 2pa^2/(a^2 + \varepsilon) > 0$. Thus zero is a strict local minimum while the weighted target has larger variance for all sufficiently large L. □

A two-point example. For $p = 2$ and $\varepsilon = 1/20$, let $Y \in \{0, 1\}$ with probabilities $(81, 1849)/1930$. At $f_\star = 1/10$, adaptive weighting gives selected probabilities $(9/10, 1/10)$, so $\mathbb{E}_{\mathbb{Q}_\star} Y = f_\star$. Direct calculation yields

$$\mathrm{Var}_{\mathbb{P}}(Y) = \frac{149769}{3724900} \approx .0402, \qquad \mathrm{Var}_{\mathbb{Q}_\star}(Y) = .09. \tag{58}$$

The expected potential has second derivative $2500/8299 > 0$ at $f_\star$. Thus a strict local minimum more than doubles the variance in a distribution with only two outcomes.

B Experimental Details

B.1 Gaussian-mixture diagnostics

Fixed-state mixture. We use Gaussian quadrature for a five-component mixture with imbalance ratio 100. At the evaluated state, component identity is hidden and its posterior equals the long-tailed prior (Table 4).

Table 4: Conditional target selection on a long-tailed mixture. The variance change compares each law around its own mean. The original majority and rarest probabilities are 68.59% and 0.686%.

p	Variance decrease $T_{\mathbb{P}} - T_{\mathbb{Q}}$	Squared mean shift $\|\delta\|^2$	Majority (%)	Rarest (%)
0.5	13.8740	3.4772	96.87	.073
1	15.6833	4.7474	99.88	.003

Conditioning controls. At $x = (0, 0)$ and $t = .5$, we compare three conditional laws. The hidden-identity law mixes component-conditioned Gaussian quadrature laws using their posterior probabilities. The observed-identity control evaluates each component separately. The symmetric control mixes the hidden law equally with its reflection $Y \mapsto 2\mu - Y$, preserving the mean, covariance, and distribution of $\|Y - \mu\|^2$. Hidden-law roots use mean-initialized fixed-point iteration with relaxation .5, a 10,000-iteration limit, and covariance-normalized step tolerance 10^{-13}. Displacements use the original-law covariance as their metric.

At radial quadrature order 128 with 512 angular points, hidden-identity displacements are .538264 and .630189 for $p = 0.5$ and $p = 1$. The observed and symmetric controls have normalized score residuals below 10^{-12} at their exact means. These are stationarity checks, not root-search or stability tests. Radial orders 96 and 160, with angular counts four times each radial order, change either hidden displacement by less than 3×10^{-7}.

Population illustration. The calculation in Figure 2 uses five isotropic Gaussian components on a radius-three circle, component standard deviation .25, and imbalance ratio 100. The path interpolates data to standard Gaussian noise. We evaluate the analytic conditional-mean field and approximate adaptive-$p = 0.5$ roots with 1024 conditional target samples on an 81×81 spatial grid over $[-5, 5]^2$ and 41 equally spaced times. Each root starts at its conditional sample mean and reuses those samples for at most 640 fixed-point iterations. Individual roots stop when their Euclidean update falls below 10^{-7}. Conditional targets use field seed 7 and batches of 128 grid points.

Fields are bilinearly interpolated in space and linearly interpolated in time. Both are integrated from the same 20,000 Gaussian starting points with 320 explicit Euler steps from $t = 1$ to 0. Starting particles use seed 8 and the independent data panel uses seed 10. Components are assigned by the nearest mixture mean in Euclidean distance. The scatter and component table use the same samples. The implementation clips field queries and particles to the spatial domain, but no clipping occurs in the reported trajectories.

Resolution checks. To compare field estimators, we also compute the sampled conditional mean ($p = 0$) using the same targets as adaptive $p = 0.5$. Table 5 varies the conditional sample count, root iterations, spatial/time resolution, and Euler steps. All settings use identical starting particles. The displayed population figure uses the final row.

Table 5: Resolution sensitivity of population generation at $p = 0.5$. Grid/times gives the spatial grid size and number of time points. Iter./Euler gives the root-iteration budget and integration steps. Δ_0 is the largest absolute component-mass difference between sampled $p = 0$ and analytic CFM, in percentage points. The last two columns are adaptive-$p = 0.5$ component masses.

Setting	Targets	Grid/times	Iter./Euler	Δ_0 (pp)	Largest (%)	Rarest (%)
Base	256	$41^2/21$	$40/80$	0.135	84.220	0.240
More targets	1024	$41^2/21$	$40/80$	0.070	84.580	0.220
More iterations	1024	$41^2/21$	$640/80$	0.070	84.640	0.220
Finer grid	1024	$81^2/41$	$640/160$	0.040	85.825	0.135
More Euler steps	1024	$81^2/41$	$640/320$	0.025	85.800	0.130

The maximum empirical fixed-point residual falls from about .013 at 40 iterations to below 10^{-7} at 640 on both grids. Increasing the iteration budget changes component masses by at most .065 percentage points on the coarse grid. Doubling Euler steps from 160 to 320 changes them by at most .030 points on the fine grid. These changes and the sampled-$p = 0$ error are small relative to the redistribution under adaptive weighting. Rare-mode suppression persists at every resolution, although its exact magnitude varies.

Finite CFM and iMF. CFM uses a two-hidden-layer MLP of width 48, 2,500 updates, batch size 512, and Adam learning rate 2×10^{-3}. iMF uses the same width, 2,000 updates, batch size 256, and learning rate 10^{-3}. Both use $\varepsilon = 10^{-2}$ and 80-step Euler sampling. Initialization, minibatches, and time tuples are matched across rules within each repeat. CFM and iMF average three and two training repeats, respectively, with additional matched sampling repeats for distribution metrics.

Table 6: Learned Gaussian-mixture flows, mean $\pm$ standard deviation. Field MSE compares against the ordinary conditional-mean velocity. Mode TV compares generated and target component probabilities. Tail ratio is generated divided by target rare-mode mass.

Objective	Training rule	p	Field MSE	Mode TV	Tail ratio
CFM	Unweighted	0	0.097 ± 0.004	0.034 ± 0.013	1.28 ± 0.34
CFM	Adaptive	0.5	0.434 ± 0.015	0.160 ± 0.013	0.15 ± 0.04
CFM	Adaptive	1	1.069 ± 0.031	0.236 ± 0.008	0.09 ± 0.03
CFM	Prediction	0.5	0.130 ± 0.012	0.035 ± 0.013	1.19 ± 0.28
CFM	Prediction	1	0.530 ± 0.105	0.027 ± 0.009	1.66 ± 0.78
iMF	Unweighted	0	0.188 ± 0.019	0.031 ± 0.009	0.89 ± 0.13
iMF	Adaptive	0.5	0.579 ± 0.023	0.161 ± 0.018	0.35 ± 0.13
iMF	Adaptive	1	1.691 ± 0.113	0.270 ± 0.007	0.07 ± 0.03
iMF	Prediction	0.5	0.338 ± 0.050	0.027 ± 0.003	1.82 ± 0.69
iMF	Prediction	1	0.688 ± 0.060	0.054 ± 0.000	2.00 ± 0.25

B.2 Spiral distribution

Data construction. We generate eight groups indexed by $g = 0, \dots, 7$. Let $\theta_{\min} = -.45\pi$, $\theta_{\max} = 2.95\pi$, and draw $s \sim \mathcal{U}[-1, 1]$ independently for each point. The angle and radius are

$$c_g = \theta_{\min} + (\theta_{\max} - \theta_{\min})(g + .5)/8, \qquad a_g = .50 + .08(g \bmod 3),$$
$$\theta = c_g + a_g s + .18 \sin(3s + .7g), \qquad \rho = .42 + .20(\theta - \theta_{\min}) + .12 \sin(2.5\theta).$$

With $e_r = (\cos\theta, \sin\theta)$ and $e_\theta = (-\sin\theta, \cos\theta)$, independent $\xi, \zeta \sim \mathcal{N}(0, 1)$ give noisy points

$$q = \rho e_r + (.055 + .012g)\xi e_\theta + \big(.030 + .006(g \bmod 2)\big)\zeta e_r.$$

We then apply the nonlinear deformation

$$\widetilde{q}_1 = q_1 + .28\sin(1.7q_2) + .055q_1q_2,$$
$$\widetilde{q}_2 = q_2 + .22\sin\big(1.35q_1 + .4\sin\theta\big) - .035q_1^2.$$

Groups denote these generating components, not disjoint spatial partitions.

The fixed training set contains 24,000 points with counts proportional to $80^{-g/7}$. Flooring the normalized counts and assigning the remainder in group order gives

$$(n_0, \ldots, n_7) = (11242, 6012, 3215, 1719, 919, 491, 262, 140).$$

Training shuffles this dataset without group rebalancing or group labels. We standardize each coordinate using the training mean and standard deviation, then map the points to $\mathbb{R}^{16}$ using an orthonormal 16×2 basis obtained by QR decomposition of a Gaussian matrix. No noise is added outside this two-dimensional subspace. Evaluation points use the same transformation.

Spiral training and evaluation. Each method uses a six-block residual MLP of width 256, batch size 512, AdamW with learning rate 10^{-4}, 1,000-step warmup, EMA decay .999, and $\varepsilon = 10^{-2}$. Training lasts 100,000 updates. The qualitative panels use seed 42. FM and diagonal iMF times are uniform. Half of each iMF batch is off diagonal ($r < t$), with endpoints obtained by ordering two independent uniform draws. Metrics use 12,000 generated samples and an independent 12,000-point real set constructed with the same count rule. For evaluation only, the same 7-nearest-neighbor classifier assigns groups to both real and generated samples, using a separate balanced reference with 2,048 labeled points per group.

SW averages one-dimensional Wasserstein-1 distances over 256 random directions in $\mathbb{R}^{16}$ (Karras et al., 2018). Group JS compares the eight assigned group probabilities of real and generated samples (Turner et al., 2019), measuring composition rather than within-group geometry.

Table 7: Spiral results, mean $\pm$ standard deviation over three seeds. SW is sliced Wasserstein distance. JS compares real and generated group proportions under the same 7-NN assignment. Euler integrates the diagonal field for 100 steps. Lower is better.

		One call		Diagonal Euler	
Training rule	p	SW	JS (10^{-3})	SW	JS (10^{-3})
FM, Unweighted	0	–	–	$\mathbf{.0050 \pm .0015}$	$\mathbf{.14 \pm .06}$
iMF, Unweighted	0	$.0174 \pm .0030$	$.73 \pm .46$	$.0103 \pm .0013$	$.35 \pm .18$
iMF, Adaptive	0.5	$.0512 \pm .0022$	$16.62 \pm .65$	$.0523 \pm .0014$	$15.85 \pm .62$
iMF, Adaptive	1	$.0995 \pm .0087$	56.00 ± 5.32	$.0955 \pm .0096$	48.16 ± 5.22
iMF, Prediction	0.5	$\mathbf{.0096 \pm .0010}$	$\mathbf{.18 \pm .10}$	$.0062 \pm .0008$	$.19 \pm .11$
iMF, Prediction	1	$.0115 \pm .0017$	$.36 \pm .02$	$.0065 \pm .0014$	$.19 \pm .10$
iMF, Adaptive (off-diag.)	1	$.0523 \pm .0041$	5.60 ± 3.44	$.0095 \pm .0013$	$.38 \pm .16$

B.3 Image-scale weighting rules

CIFAR-10 and ImageNet-LT training use seed 2026 throughout.

Public training recipes. MeanFlow uses inverse-residual exponent $p = 0.75$ on CIFAR-10 and $p = 1$ on ImageNet (Geng et al., 2025a). The official iMF and pMF losses default to $p = 1$ and $\varepsilon = 10^{-2}$, normalizing the main and auxiliary velocity losses separately. The training-dynamics study of Kim et al. (2026) also uses $p = 1$.

For iMF, we use the corrected regression prediction in Equation (2), with the JVP holding r and all other conditions fixed. Direct FM uses the diagonal prediction $u_\theta(x, t, t)$. In class-conditional runs, z also includes the class and guidance variables. The guided diagonal tangent and target construction are specified in Appendix B.5.

Within each model family, the squared regression error is fixed and only the detached denominator changes. Both scales sum over image coordinates:

$$s_\theta^{\text{adapt}} = \|\widehat{v}_\theta(Z) - T\|_2^2, \qquad s_\theta^{\text{pred}} = \|\widehat{v}_\theta(Z)\|_2^2, \tag{59}$$

$$w_\theta = \big(\text{sg}(s_\theta) + \varepsilon\big)^{-p}. \tag{60}$$

Because the JVP tangent is model-predicted, $\widehat{v}_\theta$ is independent of the remaining target randomness at fixed state and stopped model. Thus Proposition 3 applies.

B.4 CIFAR-10 AND CIFAR-10-LT100

Models are trained separately on balanced CIFAR-10 (Krizhevsky, 2009) and its imbalance-100 split. We adapt MeanFlow's CIFAR architecture and optimizer (Geng et al., 2025a), with the iMF objective for average-velocity models. All runs use the same 55.9M-parameter SongUNet (Song et al., 2021; Karras et al., 2022), batch size 512, 200k updates, Adam with learning rate 10^{-4} and 10k-step warmup, BF16, horizontal flips, and EMA evaluation. Adam uses betas $(.9, .999)$ and no weight decay. Dropout is disabled and EMA decay is .99995. Weighting uses $\varepsilon = 10^{-2}$ with sum-reduced squared norms. Data splits and random streams are fixed across weighting rules. Training on one RTX A6000 takes about 75 GPU-hours for FM and 169–170 for iMF, with 40.4GB peak allocated memory.

Each iMF batch contains equal numbers of diagonal and off-diagonal examples. Diagonal times are uniform, while off-diagonal (r, t) are the ordered values of two independent $\text{sigmoid}(\mathcal{N}(-.4, 1))$ draws. FM uses only uniform diagonal times. The off-diagonal proportion and sampler follow iMF. Relative to the original MeanFlow CIFAR recipe, we also shorten training from 800k to 200k updates, disable dropout, and retain only horizontal-flip augmentation. All settings are shared across weighting rules within each model family.

We generate 50,000 samples per checkpoint. Inception FID (Heusel et al., 2017) uses the 50,000-image balanced training set and the 12,408-image LT split. DINO-FID uses the same samples and references with frozen DINOv2 ViT-S/14 features (Oquab et al., 2024; Stein et al., 2023), bicubic resizing from 32×32 to 224×224, and ImageNet normalization. A fixed ResNet-34 (He et al., 2016) with 95.07% test accuracy estimates class priors. Given observed class frequencies q and row-normalized test confusion matrix C, we minimize $\|C^\top \pi - q\|_2^2 + 10^{-3}\,\|\pi - q\|_2^2$ by projected gradient descent over nonnegative π summing to one. Figures 7 and 8 show generated samples for both training distributions.

Table 8: CIFAR-10-LT100 training at 200k updates. Each model generates 50,000 samples, evaluated against the LT and balanced (BAL) training references. DINO-FID uses DINOv2 ViT-S/14. Prior error is $100\times$ confusion-corrected class-prior MAE against the training distribution. Lower is better; bold marks the best result within each model family.

			vs. LT		vs. BAL		
Training rule	p	NFE	FID	DINO-FID	FID	DINO-FID	Prior error
FM, Unweighted	0	50	55.42	630.87	57.99	815.59	**4.69**
FM, Adaptive	0.5	50	68.55	721.80	68.02	890.54	6.23
FM, Adaptive	1	50	**49.44**	**617.83**	**49.22**	**796.86**	5.83
iMF, Unweighted	0	1	29.85	499.09	42.88	759.53	2.80
iMF, Adaptive	0.5	1	29.16	494.16	32.86	659.18	2.23
iMF, Adaptive	1	1	53.05	748.23	49.69	864.12	4.44
iMF, Adaptive	1.5	1	57.15	825.78	55.57	935.49	4.43
iMF, Prediction	0.5	1	**18.38**	**347.36**	**28.51**	**563.17**	**.78**
iMF, Prediction	1	1	23.22	413.92	29.25	594.52	1.63
iMF, Prediction	1.5	1	26.92	452.53	30.75	615.85	2.23

Table 10 reports sampler FID and estimated group masses. Confusion correction uses real test data and may not fully correct classification errors on generated images.

Table 9: Balanced CIFAR-10 training at 200k updates, evaluated against BAL and LT training references. Sample counts, metrics, and boldface follow Table 8. Prior error is measured against the balanced training distribution.

			vs. BAL		vs. LT		
Training rule	p	NFE	FID	DINO-FID	FID	DINO-FID	Prior error
FM, Unweighted	0	50	**8.55**	**184.06**	**27.96**	**427.72**	**1.21**
FM, Adaptive	0.5	50	9.96	209.81	29.73	448.83	1.69
FM, Adaptive	1	50	11.88	243.95	31.70	470.06	2.30
iMF, Unweighted	0	1	16.48	352.76	30.33	554.88	2.43
iMF, Adaptive	0.5	1	9.73	210.16	25.32	**428.13**	.78
iMF, Adaptive	1	1	12.31	245.76	30.06	472.71	1.71
iMF, Adaptive	1.5	1	18.11	345.79	35.88	553.57	2.69
iMF, Prediction	0.5	1	8.99	208.61	24.67	439.00	.68
iMF, Prediction	1	1	8.35	200.25	**24.26**	434.71	**.47**
iMF, Prediction	1.5	1	**7.92**	**199.64**	24.45	435.90	.53

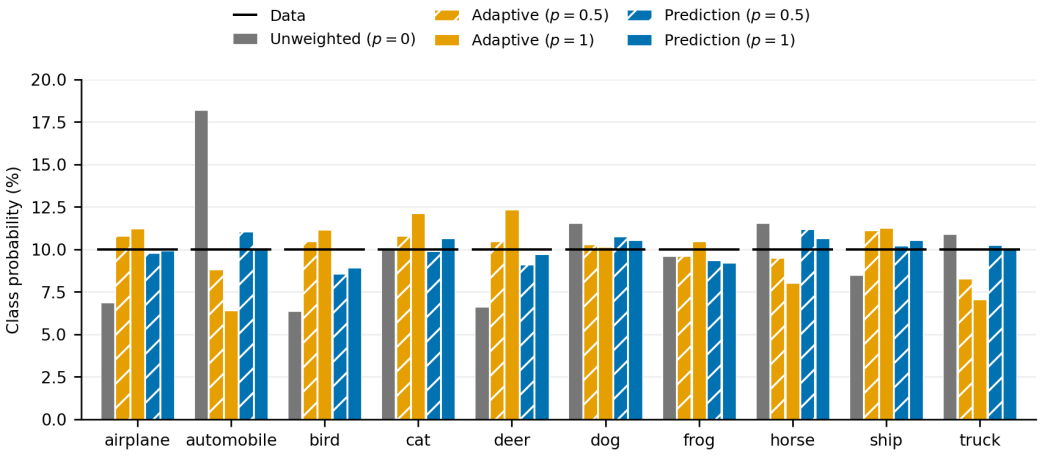


Figure 5: Estimated class probabilities for one-call iMF after balanced CIFAR-10 training, with the same classifier and correction as Figure 4. The axis is linear. Prediction weighting improves overall prior agreement. Black marks indicate the uniform training prior.

Table 10: LT-trained iMF at 200k updates. FID uses the LT reference. I-1 and I-4 use one and four interval calls, and D-E50 uses 50 diagonal Euler steps. Group masses are confusion-corrected estimates from I-1 samples. Head and tail are the three most and three least frequent training classes.

		FID by sampler			I-1 group mass (%)	
Training rule	p	I-1	I-4	D-E50	Head	Tail
Training distribution	–	–	–	–	78.93	2.19
Unweighted	0	29.85	24.88	22.36	80.12	2.92
Adaptive	0.5	29.16	24.75	25.72	67.85	4.37
Adaptive	1	53.05	48.77	42.03	57.18	6.90
Prediction	0.5	**18.38**	**11.25**	**11.87**	75.64	2.40
Prediction	1	23.22	16.23	18.10	70.97	3.41

Extension to $p = 1.5$. The four additional iMF runs use the same training configuration as their respective $p = 1$ controls, changing only the exponent. Evaluation uses the same 200k EMA endpoint, sampling seed, 50,000-image count, references, and classifier. Table 11 compares one-call, four-call interval, and 50-step diagonal sampling. On LT, the one-call adaptive model has estimated head and tail masses of 56.81% and 8.23%, respectively. Prediction weighting gives 68.12% and 3.58%.

Table 11: CIFAR iMF at $p = 1.5$ and 200k updates. FID uses each model's training-distribution reference, with the same sampler notation as Table 10.

		LT training			BAL training		
Training rule	p	I-1	I-4	D-E50	I-1	I-4	D-E50
Adaptive	1.5	57.15	46.72	26.97	18.11	14.74	12.59
Prediction	1.5	26.92	23.48	26.33	7.92	4.61	5.20

Training trajectories. Figure 6 evaluates unweighted iMF and both weighting rules at $p = 0.5, 1$, using checkpoints at 10k updates and every 20k updates from 20k to 200k. Early FID reduction is broadly similar, and weighting does not consistently accelerate it. Later, unweighted LT FID rises from 23.04 at 140k to 84.30 at 160k, while prediction $p = 0.5$ remains at 19.24 at both checkpoints. From 120k to 200k, LT prior error increases from 1.19 to 2.24 for adaptive $p = 0.5$ and from .53 to .77 for prediction $p = 0.5$. The trajectories show differences in late-stage quality and composition rather than a uniform speedup.

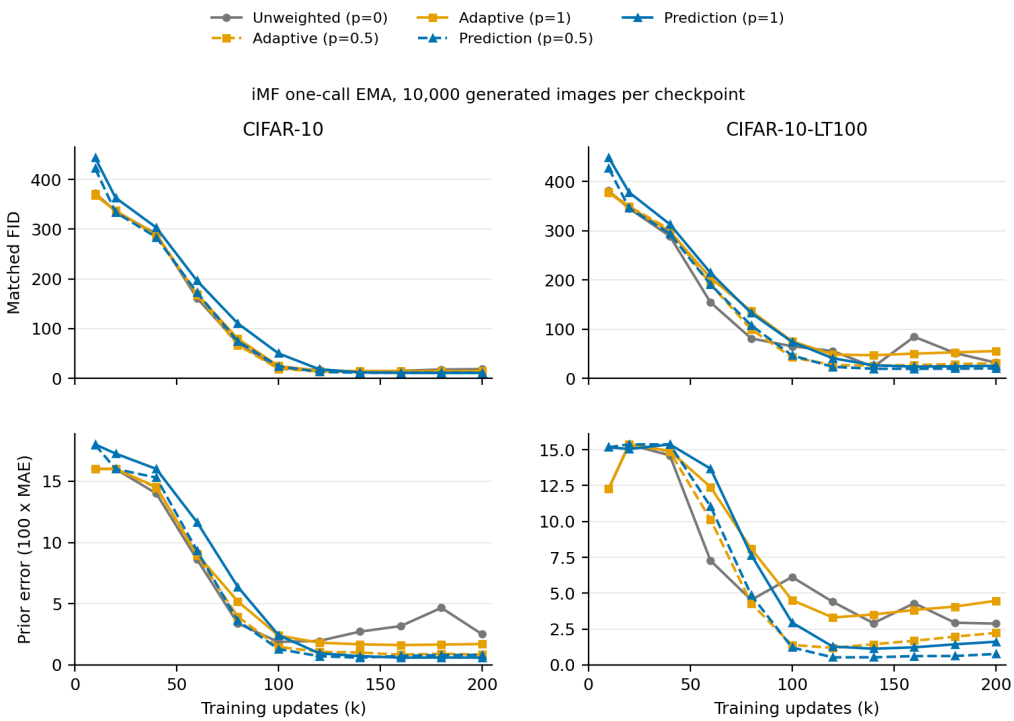


Figure 6: CIFAR iMF training trajectories. Each checkpoint uses 10,000 one-call EMA samples (50,000 in the main tables), with a shared sampling seed. FID uses each model's training-distribution reference. Prior error is $100\times$ confusion-corrected class-prior MAE from the same generated samples.

B.5 IMAGENET-LT

We use the ImageNet-LT split (Liu et al., 2019) of ImageNet (Russakovsky et al., 2015). We train and generate 256×256 RGB images. Preprocessing applies box downsampling when needed, bicubic resizing of the shorter side to 256, and a 256×256 center crop. Our 118M-parameter DiT-B/16 (Peebles & Xie, 2023) is derived from pMF (Lu et al., 2026; Geng et al., 2026), using pixel-space x-prediction, a 128-channel patch bottleneck, Muon with AdamW for the remaining parameters, class dropout, and flexible CFG. We use one EMA and no auxiliary velocity head. The 300k comparison has no perceptual losses. Batch size, warmup, and time sampling also differ from the original pMF protocol, as specified below.

All seven runs use batch size 512, BF16, horizontal flips, $\varepsilon = 10^{-2}$, no gradient clipping, and EMA decay .9995. Both optimizers use learning rate 10^{-3}, no weight decay, and 10k-step warmup. AdamW uses betas $(.9, .95)$ and Muon uses momentum .95. On one B200, 300k updates take about 112 GPU-hours for FM and 143–144 for iMF, with 146–147GB peak allocated memory.

FM uses uniform diagonal times and 50-step Euler sampling. Exactly half of each iMF batch is diagonal with uniform time. The other half orders two independent $\mathrm{sigmoid}(\mathcal{N}(-.4, 1))$ draws to obtain $r < t$. The JVP tangent is model-predicted. Conversion from x-predictions to velocities clamps the time denominator below at .05.

Flexible CFG is distilled into the model during training, with scale $\omega \in [1, 8]$ drawn using $\beta = 1$ and class-drop probability .1. The model is conditioned on ω and a guidance interval I, distinct from the flow interval $[r, t]$. Every diagonal example, including all FM examples, uses $I = [0, 1]$. Off-diagonal iMF examples draw its endpoints independently from $[0, .5]$ and $[.5, 1]$. At the sampled time, the effective scale ω_t equals ω for $t \in I$ and 1 otherwise. The target adds $(1 - 1/\omega_t)(v_c - v_u)$ to the sampled path velocity, where v_c is the stopped diagonal prediction at scale ω_t and v_u is the stopped null-class prediction at scale 1; both are evaluated with guidance interval $[0, 1]$. Dropped-

class examples retain the base target. The JVP direction uses v_c, not the sampled velocity. The average-velocity model receives the original ω and I, so the guidance interval enters its learned regression through both conditioning and the target construction.

At evaluation, direct FM passes $I = [0, 1]$ on every Euler step and gates the scale by the step time: 8.5 for $t \in [.1, .7]$, otherwise 1. One-call iMF instead passes scale 8.5 and $I = [.1, .7]$ to its single evaluation at $(r, t) = (0, 1)$. It does not disable guidance because $t = 1$ is outside the interval; the model is conditioned on the interval over which guidance was distilled. Both samplers use one distilled-model evaluation per step, not separate conditional and unconditional calls.

Global metrics use 50,000 generated images and frequency-group metrics use the corresponding subsets. Val-BAL FID compares balanced generation with the validation set. Train-LT FID compares LT-prior generation with 50,000 images sampled without replacement from the 115,846-image ImageNet-LT training split, using proportional per-class quotas rounded by the largest-remainder rule. The 300k DINO-FID uses frozen DINOv2 ViT-L/14 features (Oquab et al., 2024; Stein et al., 2023), bicubic resizing to 224×224, and ImageNet normalization. It uses the same sample and reference manifests as FID. Label agreement uses the ViT-L/14 ImageNet linear classifier, whose validation accuracy is 86.45%.

The unguided 300k comparison evaluates the same adaptive and prediction $p = 1$ EMA checkpoints at CFG $= 1$, using 50,000 images per label policy and the same references. Export batch sizes are 40 without guidance and 4096 with guidance. Adaptive and prediction give Val-BAL FIDs of $56.12/83.92$ and Train-LT FIDs of $61.00/90.67$, respectively.

Table 12: ImageNet-LT at 300k. Head, body, and tail contain classes with more than 100, 20–100, and fewer than 20 training images. Group FIDs use 50 samples per class and matching validation references. Label agreement is DINOv2 linear-probe accuracy against the requested labels over all balanced-generation samples.

Training rule	p	Head FID	Body FID	Tail FID	Label agree.
Direct FM, 50-step Euler					
FM, Unweighted	0	42.02	49.75	76.29	.238
FM, Adaptive	1	**16.95**	**24.24**	**49.78**	**.512**
iMF, one call					
iMF, Unweighted	0	65.23	71.06	92.10	.110
iMF, Adaptive	0.5	31.93	40.98	65.39	.291
iMF, Adaptive	1	**29.03**	**37.52**	**63.41**	**.380**
iMF, Prediction	0.5	51.79	59.09	84.29	.153
iMF, Prediction	1	50.24	57.84	82.82	.170

B.6 ImageNet-LT perceptual continuation

We continue the 300k checkpoints and optimizer states of FM, unweighted iMF, adaptive $p = 1$, and prediction $p = 1$ for 100k updates at effective batch size 512, without changing the architecture. Each continuation run uses one RTX A6000, accumulating gradients over eight microbatches of 64 to retain this batch size within GPU memory. The pMF perceptual terms (Lu et al., 2026) use LPIPS (Zhang et al., 2018) and ConvNeXt-V2 Base (Woo et al., 2023) with coefficients .4 and .1, active for $t \leq .8$ and each normalized as $L/\operatorname{sg}(L + .01)$. They compare $X_t - tu_\theta(X_t, r, t)$ with the clean target X_0. Feature networks are frozen, with gradients passing through their prediction inputs. The primary regression weighting rule is unchanged. The target-preservation result applies to that regression term, not to the augmented perceptual objective.

Table 3 uses the 400k endpoint and the same 50,000-sample evaluation policies as at 300k. DINO-FID uses frozen DINOv2 ViT-S/14 features with 224×224 inputs. This is distinct from the ViT-L/14 ImageNet linear classifier used for label agreement. Figure 10 shows samples after perceptual continuation.

Longer continuation. Extending training from 400k to 425k updates changes Val-BAL FID from 9.64 to 9.28 for adaptive weighting and from 10.60 to 10.54 for prediction weighting. Train-LT FID changes from 8.05 to 7.74 and from 9.71 to 9.72, respectively. Both endpoints use the same generation policies and 50,000-sample protocol. We retain 400k as the common endpoint.

We also evaluate the same 400k EMA checkpoints without guidance (Table 13), retaining the label schedules and sampling seeds. At CFG $= 1$, adaptive and prediction $p = 1$ have nearly identical Val-BAL FID, 11.12 and 11.13. This comparison uses the perceptually fine-tuned models and does not replace the 300k primary-regression results.

Table 13: Guidance comparison for the 400k perceptual checkpoints. Each setting uses 50,000 samples, with balanced and LT-prior labels evaluated against their matching references. FM uses 50-step Euler and iMF uses one call. All entries are FID.

		Val-BAL		Train-LT	
Training rule	p	CFG=1	CFG=8.5	CFG=1	CFG=8.5
FM, Unweighted	0	87.58	42.58	94.64	41.43
iMF, Unweighted	0	90.40	58.34	97.85	61.91
iMF, Adaptive	1	11.12	9.64	9.82	8.05
iMF, Prediction	1	11.13	10.60	10.39	9.71

B.7 Precision and Recall

We report nearest-neighbor precision and recall (Kynkäänniemi et al., 2019) with $k = 5$, using frozen DINOv2 ViT-S/14 features for CIFAR and ViT-L/14 for ImageNet. Inputs are resized to 224×224 and ImageNet-normalized. All models generate 50,000 samples, with 50-step Euler for FM and one call for iMF. CIFAR compares LT-trained models with the 12,408-image LT training split and BAL-trained models with the 10,000-image test set. The latter differs from the training reference used for CIFAR FID. ImageNet uses the same 50,000-image references and guidance scale 8.5 as the main tables.

Tables 14 to 16 report the results. Notably, unweighted iMF has the highest ImageNet precision but the lowest recall at 300k. High precision alone therefore does not indicate broad distributional coverage.

Table 14: CIFAR precision and recall at 200k updates, using DINOv2 ViT-S/14. References are LT training data and the BAL test set, respectively.

		LT training		BAL training	
Training rule	p	Precision	Recall	Precision	Recall
FM, Unweighted	0	.848	.529	.878	.563
FM, Adaptive	0.5	.909	.646	.884	.533
FM, Adaptive	1	.938	.639	.880	.501
iMF, Unweighted	0	.772	.110	.785	.222
iMF, Adaptive	0.5	.880	.265	.861	.422
iMF, Adaptive	1	.928	.241	.867	.387
iMF, Prediction	0.5	.854	.284	.855	.414
iMF, Prediction	1	.861	.269	.852	.419

Table 15: ImageNet-LT precision and recall at 300k updates without perceptual losses, using DINOv2 ViT-L/14 and the sample sets in Table 2.

		Val-BAL		Train-LT	
Training rule	p	Precision	Recall	Precision	Recall
FM, Unweighted	0	.623	.133	.643	.132
FM, Adaptive	1	.566	.375	.589	.357
iMF, Unweighted	0	.692	.067	.709	.077
iMF, Adaptive	0.5	.566	.208	.574	.231
iMF, Adaptive	1	.563	.272	.560	.248
iMF, Prediction	0.5	.623	.095	.643	.123
iMF, Prediction	1	.633	.116	.647	.130

Table 16: ImageNet-LT precision and recall after perceptual continuation to 400k, using DINOv2 ViT-L/14 and the sample sets in Table 3.

		Val-BAL		Train-LT	
Training rule	p	Precision	Recall	Precision	Recall
FM, Unweighted	0	.628	.125	.641	.133
iMF, Unweighted	0	.678	.076	.701	.110
iMF, Adaptive	1	.522	.327	.547	.306
iMF, Prediction	1	.491	.302	.508	.294

C Additional Generated Samples

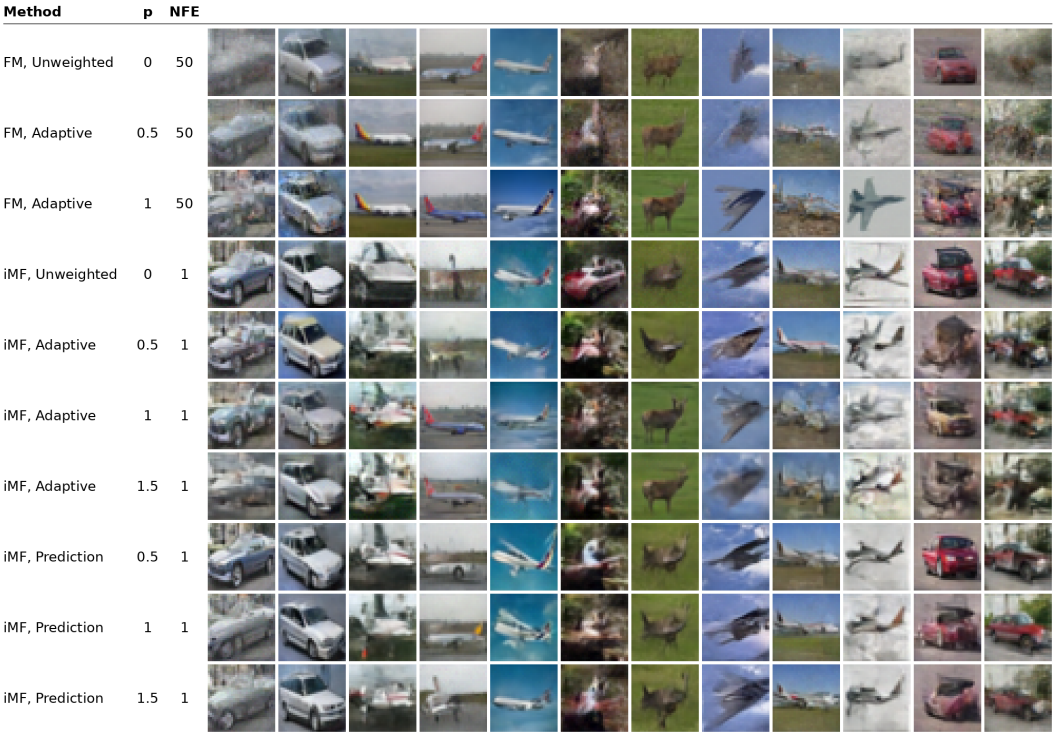


Figure 7: Unconditional CIFAR-10-LT100 samples at 200k updates. Rows show all methods in Table 1, using 50-step Euler for FM and one call for iMF. Columns use shared Gaussian initializations and the first 12 evaluation indices, without class-based selection. The 32×32 images are enlarged by nearest-neighbor interpolation.

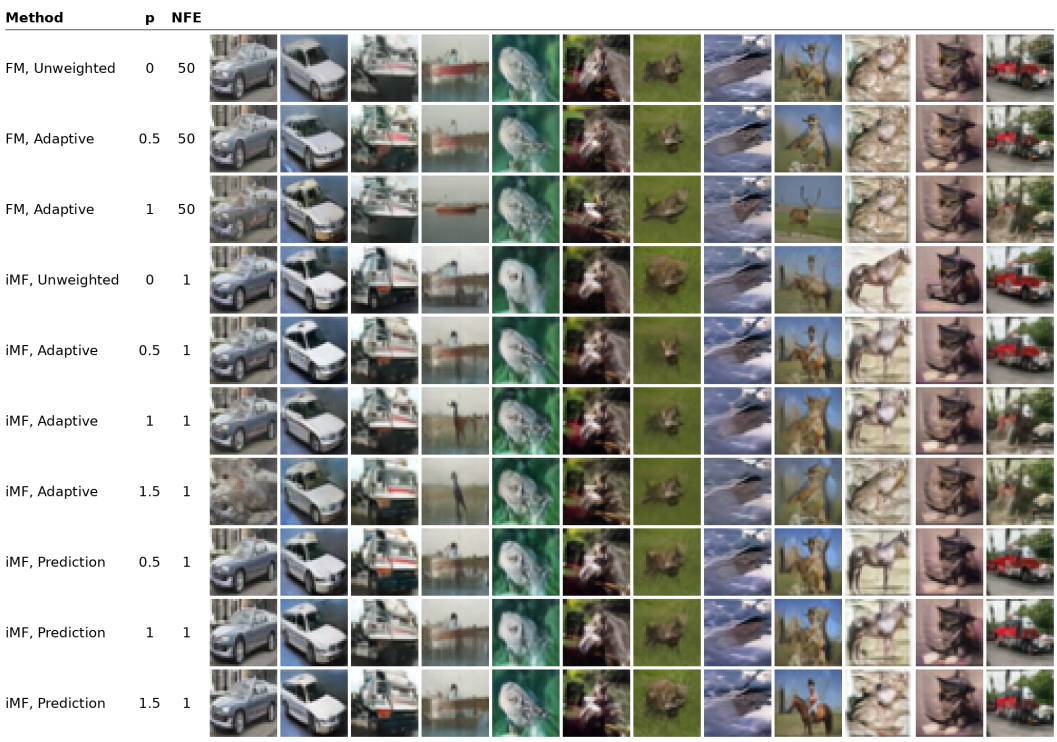


Figure 8: Unconditional samples after balanced CIFAR-10 training at 200k updates. The methods, initializations, sample indices, and display scale match Figure 7. FM uses 50-step Euler and iMF uses one call.

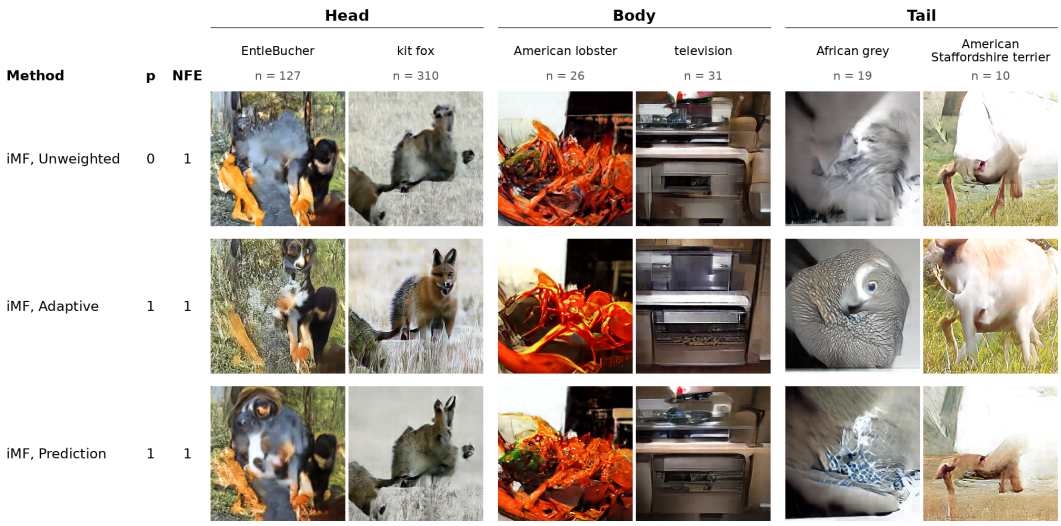


Figure 9: One-call iMF samples on ImageNet-LT at 300k updates. Two classes per frequency group are selected using a fixed random seed. Rows share labels and sampling seed.

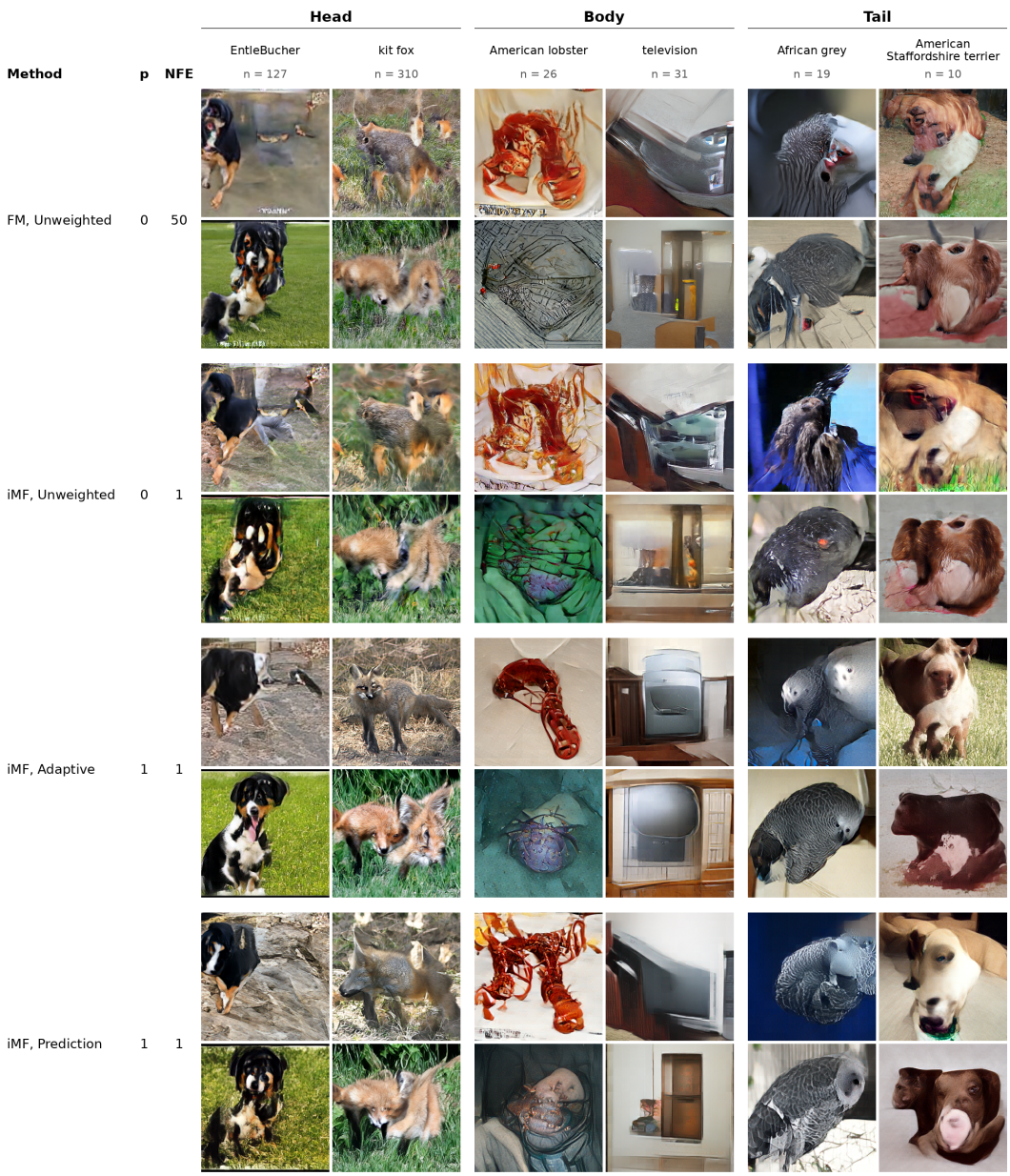


Figure 10: ImageNet-LT samples at 400k updates, after 100k updates with perceptual losses. Each method shows the first two evaluation samples per selected class, with shared labels and initial noise across methods. FM uses 50-step Euler and iMF uses one call, with CFG 8.5 and guidance interval $[.1, .7]$. Classes match Figure 9, but the 300k and 400k grids use different initial noise.