# WHEN MORE CALLS HELP: DATASET-FREE LOW-COMPUTE SELF-DISTILLATION FOR MEANFLOW

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Distillation has become an effective way to compress multi-step generative models into fewer steps while retaining much of the teacher's sample quality. We observe a complementary opportunity in MeanFlow: selected multi-step samplers of an already one-step-capable checkpoint produce better samples than its one-step mode. Together, these observations suggest a simple route to a stronger one-step model: distill the better multi-step mode back into one step, keeping the quality lost in distillation smaller than the quality gained by sampling. We call this recipe Many-to-One self-distillation (M2O). This transfer requires only a small adaptation budget. Using only random noise, class labels, and a frozen self-teacher from the same checkpoint, M2O matches generated distributions in a frozen visual feature space *without accessing training data*. On ImageNet-256, M2O improves iMF-B/2 from 3.35 to 3.00 FID-50K in 0.60 H200-hour. On iMF-L/2, it reaches 1.72 in 0.93 H200-hour, improving on the guidance-calibrated one-step baseline of 1.88 by 0.16 FID. A fixed-guidance control improves from 2.31 to 1.72 at CFG 5.00. Both students preserve single-step, single-branch inference. Our results show that a checkpoint's multi-step advantage can become a practical source of low-compute, dataset-free improvement for one-step generation.

## 1 INTRODUCTION

Distillation has made large reductions in generative sampling cost possible. Progressive distillation preserves much of a diffusion teacher's perceptual quality with far fewer sampling steps (Salimans & Ho, 2022); distribution-based methods extend this idea to strong one-step generators (Yin et al., 2024b; Zhou et al., 2024; Yin et al., 2024a). These results suggest that compressing a teacher need not erase most of its quality advantage. The practical question is where to find that advantage, and how much additional training is needed to retain it.

MeanFlow offers a natural source within a single checkpoint. It models interval-average velocity, allowing either a single jump from noise to a sample or several composed interval calls (Geng et al., 2025a;b). The released one-step generator is therefore also a candidate multi-step teacher. We find selected multi-step iMF settings that improve on the checkpoint's one-step mode: on B/2, FID-50K changes from 3.35 to 2.52; on L/2, a fixed-guidance comparison changes from 2.31 to 1.59. This creates a measurable quality margin that distillation can transfer to the original one-step deployment path.

The resulting idea is simple (Figure 1). Let sampling improve a fixed evaluation score by $A$, and let distillation give back $D$ of that improvement. The student still beats the original one-step model whenever $A > D$. Here, $D$ is the quality gap between student and teacher under the same evaluation metric. This view makes the empirical question precise: *can we retain a positive sampling gain with a very small adaptation budget?*

We answer this question with *Many-to-One self-distillation* (M2O). A frozen multi-step mode of the released iMF checkpoint supervises a one-step student initialized from the same weights. Adaptation uses random noise, class labels, and teacher-generated endpoints, comparing student and teacher distributions through a frozen visual encoder. Adaptation accesses no real training images or reference statistics. Here, dataset-free refers to this adaptation stage: the generator, VAE, and visual encoder are pretrained, and sampler selection uses real-reference FID.



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.

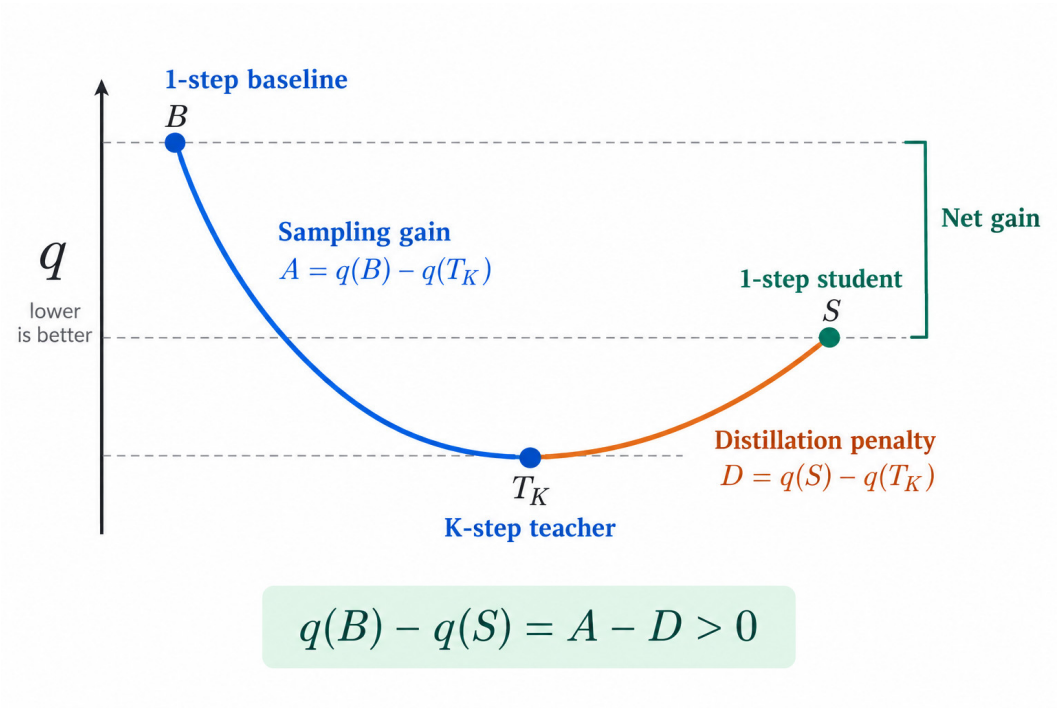



Figure 1: The M2O transfer criterion. A selected $K$-step self-teacher gains $A = q(B) - q(T_K)$ over the released one-step baseline. Distillation incurs an evaluation penalty $D = q(S) - q(T_K)$ relative to that teacher. The one-step student improves on the baseline when $A > D$, retaining net gain $A - D$. The schematic uses a shared lower-is-better evaluation scale $q$ (e.g., FID).

Our main finding is that this transfer takes less than one H200-hour after teacher selection: B/2 reaches FID-50K 3.00 in 0.60 hour, and fixed-guidance L/2 reaches 1.72 in 0.93 hour, below its guidance-calibrated baseline of 1.88. Deployment keeps the original single branch and one network evaluation. Three sets of evidence support this finding: short-budget curves show that improvement appears early; matched one-step self-targets give little or no gain while multi-step targets improve quality; and objective controls identify which supervision retains the advantage. The contribution is this measured, low-compute checkpoint-improvement regime and its controlled validation.

## 2 Related Work

**Distillation and data-free transfer.** Progressive distillation repeatedly compresses a trained sampler into fewer steps (Salimans & Ho, 2022). DMD combines distribution matching with regression to teacher-generated outputs (Yin et al., 2024b); DMD2 removes that regression requirement and adds real-image adversarial supervision (Yin et al., 2024a). Diff-Instruct and SiD show that a pretrained diffusion model can supervise a generator without real training samples (Luo et al., 2023; Zhou et al., 2024). Recent flow distillation methods learn network-free velocity policies ($\pi$-Flow) or stabilize consistency with flow supervision (FACM) (Chen et al., 2025; Peng et al., 2025). This literature establishes the feasibility of retaining teacher quality, including successful one-step and data-free settings. Our question concerns the additional budget needed to improve an already useful one-step checkpoint using its own better multi-step sampler.

**MeanFlow and few-step generation.** MeanFlow learns interval-average velocities for one-step generation from scratch (Geng et al., 2025a). Improved MeanFlow revises its objective, guidance conditioning, and architecture to strengthen that mode (Geng et al., 2025b). MeanFlow also supports distillation from pretrained flow or diffusion models: MF-RAE combines representation autoencoders with teacher distillation and consistency mid-training (Hu et al., 2025), while recent methods study stable average- velocity transfer and large-model distillation (Zhao et al., 2026; He

et al., 2026). We keep the released iMF architecture and one-step inference path, and ask whether its own sampling margin can fund a short self-distillation stage.

**Feature matching and self-improvement.** SwD matches teacher-synthesized samples using MMD in frozen diffusion features, including an MMD-only variant (Starodubcev et al., 2026). Its patch-level comparisons motivate a close methodological comparison with our global-feature minibatch matching. Perceptual feature supervision is also established by LPIPS and BiFlow (Zhang et al., 2018; Lu et al., 2025). We use a cubic KID-style kernel (Bińkowski et al., 2018) with teacher-RMS normalization and paired-diagonal exclusion. Neon improves generators with a short self-training stage followed by negative weight extrapolation (Alemohammad et al., 2025). M2O instead learns toward a verified better sampling mode of the same checkpoint. Our paired-feature, random-encoder, and teacher-step controls test the ingredients needed in this specific low-compute setting.

# 3 Many-to-One self-distillation

## 3.1 Retaining the sampling gain

Let $B$ denote the released one-step model, $T_K$ its frozen $K$-step sampling mode, and $S$ the adapted one-step student. For a fixed lower-is-better evaluation $q$, define sampling gain $A = q(B) - q(T_K)$ and distillation penalty $D = q(S) - q(T_K)$. Then

$$q(B) - q(S) = A - D > 0 \quad \Longleftrightarrow \quad A > D. \tag{1}$$

This identity separates the teacher's available advantage from the part lost during transfer. It does not require the student to exactly match its teacher, and $D$ can be negative if the student surpasses it. For B/2, the measured gain is 0.84 and the residual penalty is 0.48, leaving a 0.36 improvement (computed before rounding).

M2O first selects a multi-step teacher with positive measured advantage, then adapts a student from the same released weights. Figure 2 shows the training paths. For Gaussian noise $z$ and a uniformly drawn class label $y$, the teacher produces endpoint $T_K(z, y)$ online and the student produces $G_\theta(z, y)$ in one call. The teacher stays fixed; only selected student parameters change.

## 3.2 Matching generated distributions

The frozen VAE (Rombach et al., 2022) decodes the endpoints using the released iMF latent codec. A frozen ImageNet-22K-pretrained ConvNeXtV2-Tiny (Woo et al., 2023) maps the decoded images to global features $s_i = \phi(G_\theta(z_i, y_i))$ and $t_i = \phi(T_K(z_i, y_i))$; $\phi$ includes decoding and image preprocessing. We normalize both feature batches by the teacher RMS, giving $\tilde{s}_i$ and $\tilde{t}_i$. Using the cubic kernel $k(a, b) = (a^\top b/d_\phi + 1)^3$ for feature dimension $d_\phi$, we match their minibatch distributions:

$$\widehat{\mathcal{L}}_{\text{dist}} = \frac{1}{n(n-1)} \sum_{i \neq j} \big[k(\tilde{s}_i, \tilde{s}_j) + k(\tilde{t}_i, \tilde{t}_j) - 2k(\tilde{s}_i, \tilde{t}_j)\big]. \tag{2}$$

Teacher features and the normalizer are stop-gradient. Student gradients pass through the frozen encoder and decoder. Excluding $i = j$ removes the self diagonals and the paired teacher–student cross diagonal. The primary recipe uses this KID-style distribution objective alone. Appendix C gives the normalization and index-mask details.

## 3.3 Adaptation and deployment

We initialize the student from the released checkpoint and adapt selected generator parameters using online noise–label pairs and teacher outputs. The backbone outside this parameter subset remains frozen. At deployment, the teacher and visual encoder are discarded; the student retains the original one-NFE, single-branch generator and VAE decoder. Appendix A provides the training and evaluation protocol.

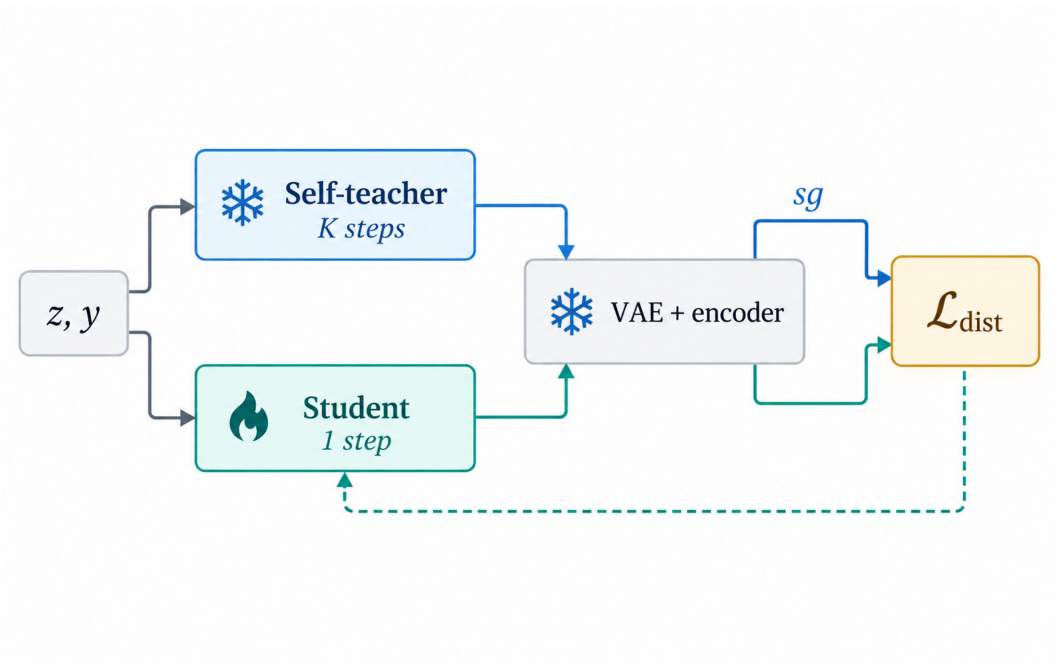



Figure 2: M2O adaptation. Shared noise $z$ and labels $y$ feed a frozen $K$-step self-teacher and a one-step student initialized from the same released checkpoint. Both branches separately pass through the same VAE decoder and visual encoder to obtain features for $\widehat{\mathcal{L}}_{\text{dist}}$. Snowflakes denote frozen parameters; sg stops gradients on the teacher branch. Student gradients pass through the frozen decoder and encoder; the dashed path updates selected student parameters. Deployment retains one generator NFE, one conditional branch, and the usual VAE decoder.

# 4 Experiments

## 4.1 Setup and evaluation

We test released iMF-B/2 and iMF-L/2 on class-conditional ImageNet-256. We report FID (Heusel et al., 2017) on 50,000 generated images using a shared evaluator. The primary B/2 and L/2 runs use 6K and 16K updates, respectively. Detailed sampler, optimizer, and evaluation settings appear in Appendix A.

B/2 teacher and student use separately selected sampling settings. The primary L/2 comparison fixes CFG=5.00 and all other sampling settings across baseline, teacher, and student, isolating changes beyond guidance calibration. A separate L/2 baseline with CFG=10.50 scores 1.88. Table 5 gives the full settings.

## 4.2 How little compute retains a useful gain?

Table 1 gives the primary results. On B/2, the teacher improves FID-50K from 3.35 to 2.52, and the one-step student reaches 3.00 in 0.60 H200-hour. It retains about 42% of the teacher's advantage. On fixed-guidance L/2, the teacher improves 2.31 to 1.59, and the student reaches 1.72 in 0.93 hour, retaining about 83% of the fixed-guidance gain. Relative to the separately calibrated 1.88 baseline, the improvement is 0.16 FID; the larger 0.59 change isolates adaptation at fixed CFG=5.00. Both students retain a positive net gain, satisfying $A > D$.

Improvement already appears at small budgets (Figure 3). B/2 reaches 3.11 after 1K updates/0.16 H200-hour and 3.03 after 3K/0.33 hour. L/2 reaches 1.83 after 4K/0.33 hour and 1.75 after 8K/0.53 hour. The primary budgets are practical operating points: extending B/2 to 12K updates gives 2.93 in 1.07 hours.

Table 1: Dataset-free adaptation on ImageNet-256 × 256, measured by FID-50K (lower is better). Green numbers show changes from the primary one-step baselines. H200 h measures adaptation after teacher selection, including setup and compilation. L/2 uses fixed guidance at 5.00.

| Model | Baseline 1-step | Self-teacher | 1-step student | H200 h |
|---|---|---|---|---|
| B/2 | 3.35 | 2.52 | **3.00** (−0.35) | 0.60 |
| L/2 (fixed CFG) | 2.31 (1.88)† | 1.59 | **1.72** (−0.59) | 0.93 |

† 2.31 uses fixed CFG=5.00, matching the teacher and student; 1.88 in parentheses uses separately calibrated CFG=10.50. The student improves on this calibrated baseline by 0.16 FID.

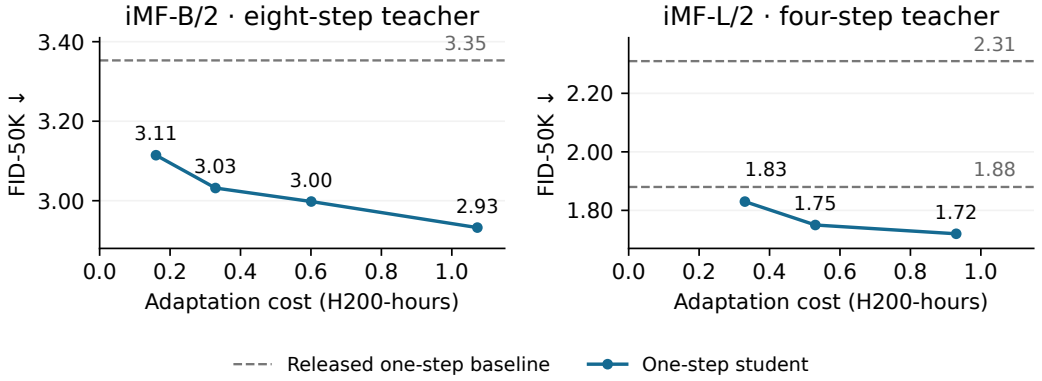



Figure 3: Adaptation cost and FID-50K on both model sizes. B/2 uses an eight-step teacher; L/2 uses a four-step teacher. Dashed lines mark the released one-step baselines: 3.35 for B/2 and 2.31 for L/2 at fixed CFG=5.00, plus the L/2 baseline of 1.88 at CFG=10.50. L/2 intermediate-checkpoint times are approximate; final-run costs include setup and compilation.

These costs measure adaptation after teacher selection. The recorded L/2 sampler searches cost an additional 6.24 H200-hours (5.17 plus 1.07); B/2's earlier discovery experiments are not aggregated into a total search cost. Appendix A details this accounting. A same-seed B/2 rerun on a second H200 host also reaches 3.00 in 0.59 hour, reproducing the primary result across hosts.

### 4.3 Does the gain come from the multi-step teacher?

Table 2 varies the frozen target while matching updates and synthetic-pair counts within each model size. At the B/2 6K budget, a one-step self-target yields 3.33, close to the released 3.35. Two-, four-, and eight-step targets yield 3.07, 2.98, and 3.00.

The fixed-setting L/2 control is more direct. One-step self-matching gives 2.35 against baseline 2.31, while $K = 2, 4, 8$ teachers score 1.66, 1.59, and 1.58 and their students reach 1.79, 1.72, and 1.69, respectively. For $K = 2$, the student retains about 81% of the teacher's advantage in 0.82 hour. Across these controls, a better multi-step target supplies the useful margin; self-matching alone does not give a comparable improvement. A separate B/2 same-guidance control supports the same attribution (Appendix B).

### 4.4 Which supervision preserves the advantage?

Table 3 compares objectives and representations at matched updates and generated-pair counts. Moving the kernel from pretrained visual features to normalized latent endpoints worsens FID from 3.00 to 3.90 on B/2 and from 1.72 to 2.47 on L/2. Frozen random ConvNeXt features yield 3.38 and 2.25. Pretrained visual features retain the largest gain among these representations.

On B/2 at 6K updates, paired ConvNeXt regression and LPIPS-VGG score 3.84 and 5.43 after 0.58 H200-hour each. Both lose the selected teacher's advantage over the 3.35 baseline, while distribution matching reaches 3.00. For L/2, the matched-time comparison gives 1.72 for distribution matching,

Table 2: Teacher-step controls at matched update and sample budgets within each model size. L/2 fixes all sampling settings except teacher step count; B/2 uses selected teacher settings.

| Model | Teacher $K$ | Teacher FID-50K | Student FID-50K | H200 h |
|---|---|---|---|---|
| B/2 | 1 | 3.35 | 3.33 | 0.50 |
| | 2 | 2.89 | 3.07 | 0.50 |
| | 4 | 2.63 | 2.98 | 0.52 |
| | 8 | 2.52 | 3.00 | 0.60 |
| L/2 | 1 | 2.31 | 2.35 | 0.81 |
| | 2 | 1.66 | 1.79 | 0.82 |
| | 4 | 1.59 | 1.72 | 0.93 |
| | 8 | 1.58 | 1.69 | 1.18 |

Table 3: Component controls at matched updates and synthetic-pair counts. B/2 uses 6K updates; L/2 uses 16K updates with the primary four-step teacher. Training time is reported for each objective.

| | B/2 | | L/2 | |
|---|---|---|---|---|
| Arm | H200 h | FID-50K | H200 h | FID-50K |
| Latent distribution | 0.22 | 3.90 | 0.68 | 2.47 |
| Random ConvNeXt distribution | 0.60 | 3.38 | 0.93 | 2.25 |
| Pretrained ConvNeXt distribution | 0.60 | **3.00** | 0.93 | **1.72** |
| Paired cross diagonal retained | 0.61 | 3.33 | 0.94 | 1.99 |
| Endpoint + distribution | 0.58 | 3.04 | 0.97 | 1.83 |
| Paired ConvNeXt regression | 0.58 | 3.84 | 1.07 | 2.06 |
| Paired LPIPS-VGG | 0.58 | 5.43 | 1.03 | 3.40 |

2.04 for paired ConvNeXt, and 3.37 for LPIPS-VGG at approximately 0.93 H200-hour. Appendix A reports the short-budget comparisons and full timing protocol.

With paired endpoint regression added, B/2 scores 3.04 versus 3.00 and L/2 scores 1.83 versus 1.72. Retaining the paired cross diagonal gives 3.33 and 1.99. A sparse hidden-alignment control reduces its own error without improving FID. Endpoint-only and randomized-student controls further support the role of the pretrained starting point and visual supervision (Appendix B). The primary M2O recipe therefore uses only the generated-distribution objective.

Independent class-fidelity diagnostics also improve for both students; Section 5.2 examines these results.

### 4.5 Context among ImageNet generators

Table 4 compares selected published generators with our local baselines and students. Literature rows retain each source's training and evaluation protocols. Both M2O rows use the primary results from Table 1. The 409M-parameter L/2 student reaches 1.72 FID, numerically matching the published 610M-parameter iMF-XL/2 result.

## 5 Analysis

### 5.1 Teacher quality and transferability

Table 2 separates the quality available in the teacher from the quality retained by the student. On B/2, increasing the teacher from two to four steps improves both teacher and student FID, from 2.89 to 2.63 and from 3.07 to 2.98, respectively. Moving to eight steps further improves the teacher to 2.52, while the student reaches 3.00. The primary eight-step setting uses the stronger measured teacher. The four-step arm provides a lower-cost operating point: 2.98 student FID in 0.52 hour versus 3.00 in 0.60 hour. Its smaller sampling gain is offset by a smaller measured distillation penalty. This

Table 4: Selected class-conditional ImageNet-$256 \times 256$ results. Parameters are generator counts; NFE excludes the image decoder and $250 \times 2$ counts both CFG branches. Published rows use their source protocols. Stars mark reproduced baselines and bold marks our adapted L/2 result.

| Model | Params | NFE | FID-50K ↓ |
|---|---|---|---|
| DiT-XL/2 (Peebles & Xie, 2023) | 675M | $250 \times 2$ | 2.27 |
| SiT-XL/2 (SDE) (Ma et al., 2024) | 675M | $250 \times 2$ | 2.06 |
| MeanFlow-B/2 (Geng et al., 2025a) | 131M | 1 | 6.17 |
| MeanFlow-M/2 (Geng et al., 2025a) | 308M | 1 | 5.01 |
| MeanFlow-L/2 (Geng et al., 2025a) | 459M | 1 | 3.84 |
| MeanFlow-XL/2 (Geng et al., 2025a) | 676M | 1 | 3.43 |
| iMF-B/2 (Geng et al., 2025b) | 89M | 1 | 3.35* |
| iMF-M/2 (Geng et al., 2025b) | 174M | 1 | 2.27 |
| iMF-L/2 (Geng et al., 2025b) | 409M | 1 | 1.88* |
| iMF-XL/2 (Geng et al., 2025b) | 610M | 1 | 1.72 |
| ARD (Kim et al., 2025) | 675M | 4 | 1.84 |
| $\pi$-Flow-XL/2 (Chen et al., 2025) | 675M | 1 | 2.85 |
| FACM (Peng et al., 2025) | 675M | 1 | 1.70 |
| MF-RAE (unguided) (Hu et al., 2025) | 841M | 1 | 2.03 |
| iMF-B/2 + M2O (0.60 H200 h) | 89M | 1 | 3.00 |
| iMF-L/2 + M2O (0.93 H200 h) | 409M | 1 | **1.72** |

* Our reproductions of the released checkpoints; the original iMF paper reports 3.39 for B/2 and 1.86 for L/2. The reproduced L/2 value 1.88 uses calibrated CFG=10.50; the fixed-CFG=5.00 baseline for M2O is 2.31 (Table 1). With an eight-step teacher, the L/2 student reaches 1.69 in 1.18 H200-hours (Table 2).

illustrates the two components of Equation 1: a useful teacher supplies a quality margin, and the transfer objective determines how much survives the adaptation budget.

L/2 exhibits a different pattern under fixed guidance: student FID improves from 1.79 to 1.72 to 1.69 as the teacher grows from two to four to eight steps. The four-step setting captures most of this improvement within 0.93 H200-hour; eight steps add 0.25 hour for a further 0.03 FID reduction. Across both sizes, the student keeps the same one-call inference path. Teacher steps thus provide an adaptation-time quality–compute trade-off whose benefit is retained at a fixed deployment cost. These per-run comparisons identify practical operating points; independent training seeds are needed to resolve the small differences between nearby settings.

The reference baseline also determines how much advantage is available. For primary L/2 at fixed CFG=5.00, the teacher improves 2.31 to 1.59 and the student reaches 1.72. Against the separately guidance-calibrated 1.88 baseline, the same teacher has a smaller 0.29-FID advantage, of which the student retains 0.16, or about 55% using the rounded table values. The fixed-guidance comparison isolates adaptation within one sampling configuration; the calibrated reference measures improvement over a stronger available one-step deployment setting. Both comparisons support a positive gain, with different magnitudes.

### 5.2 Class fidelity and batch composition

The primary kernel matches minibatch feature distributions with independently sampled labels, so most off-diagonal comparisons cross class boundaries. We therefore examine conditional fidelity using a separate frozen ResNet-50V2 classifier on 50K generated images per arm (Appendix A). Requested-class top-1 accuracy rises from 87.21% to 87.86% on B/2 and from 84.91% to 87.08% on fixed-guidance L/2. Top-5 accuracy rises from 97.16% to 97.43% and from 96.51% to 97.31%, respectively. Both students retain coverage of all 1,000 predicted classes. These results show that the FID improvement accompanies better class agreement under a representation distinct from the training encoder.

Within-class feature distance contracts from 0.47 to 0.46 on B/2 and from 0.51 to 0.49 on L/2. Class agreement and feature spread therefore capture different aspects of adaptation. Appendix B examines same-class minibatches; real-reference precision and recall remain unmeasured.

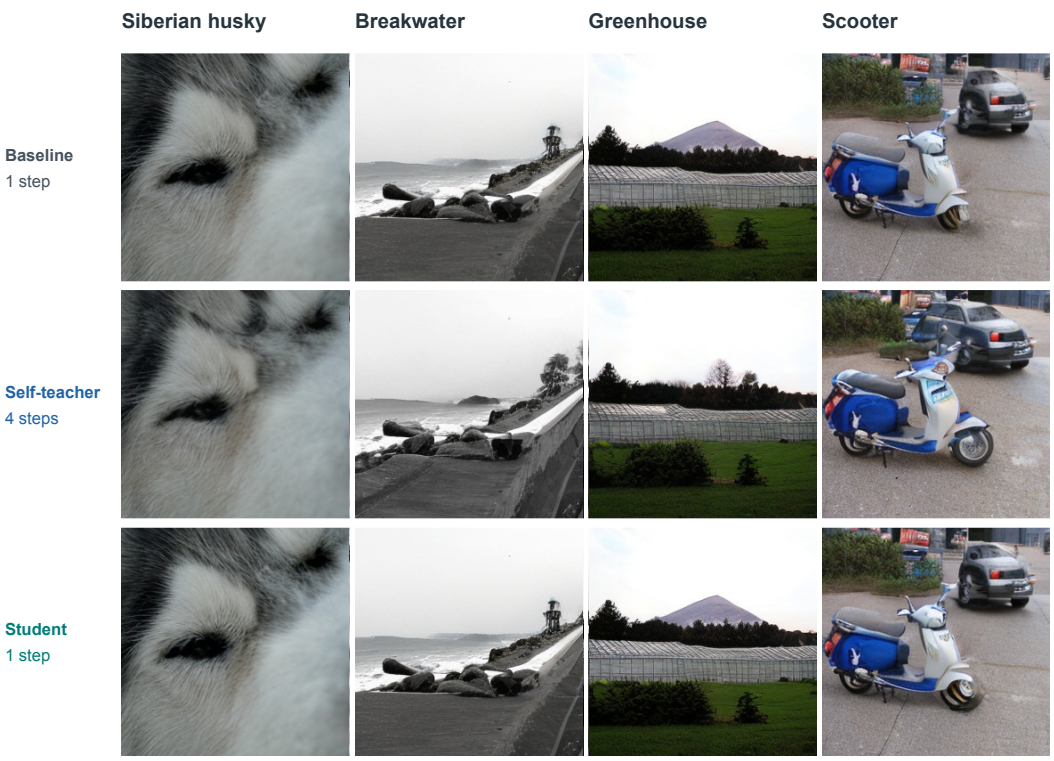



Figure 4: Complete L/2 images from shared noise and class labels. Rows show the one-step baseline, four-step self-teacher, and one-step M2O student. All use fixed CFG=5.00; the student is the primary 0.93-H200-hour model.

### 5.3 Early progress and sustained transfer

The short-budget L/2 controls reveal a dependence on training duration. At 4K updates, paired ConvNeXt regression reaches 1.82 FID in 0.33 hour and distribution matching reaches 1.83 at the same cost. Paired perceptual supervision therefore gives useful early improvement. At approximately 0.93 hour, distribution matching reaches 1.72 while the paired checkpoint, selected by elapsed time, scores 2.04. Its full 16K-update run scores 2.06. Thus the longer-budget difference reflects distinct quality trajectories: distribution matching improves further, while paired regression gives back part of its early gain. This supports evaluating objectives across training horizons and motivates the matched-time comparison alongside the matched-update controls in Table 3.

### 5.4 From full images to local details

We compare the baseline, self-teacher, and student using identical noise and class labels. Figure 4 shows complete L/2 outputs across an animal close-up, a coastal scene, a building, and a vehicle. Figure 5 then examines matched local regions. Appendices E and F provide complete comparisons and uncurated samples.

The student largely retains the baseline's composition while changing local appearance, including the greenhouse roof pattern and scooter wheel contours. The shared-noise comparisons distinguish this behavior from copying every teacher-specific arrangement.

In the crab example, leg contours and gravel texture change with similar body pose. The zebra student does not copy the teacher's muzzle alteration. Transfer is incomplete for the water tower: the student retains the baseline's irregular upper structure while the teacher produces a more regular cylindrical tank. The full uncurated grids in Appendix F complement these local comparisons.

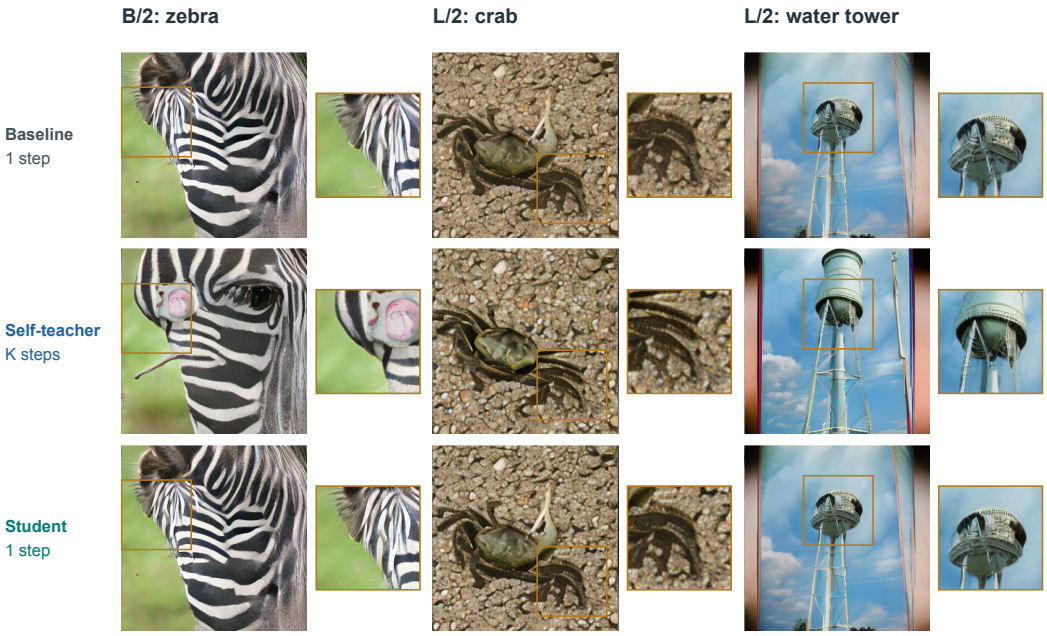



Figure 5: Shared-noise qualitative comparisons. Each column contains a released one-step baseline, its frozen self-teacher, and the adapted one-step student. Boxes locate the detail crops shown to the right. The zebra illustrates a teacher-specific local change that the student does not copy; the crab highlights leg and surface detail; the water tower shows an untransferred change in object geometry.

### 5.5 The target and the representation

The teacher-step and feature controls probe complementary parts of the transfer. With pretrained visual features but a one-step self-target, L/2 scores 2.35 against the fixed-guidance baseline of 2.31. With the better four-step teacher but a frozen random encoder, it reaches 2.25, still above the calibrated one-step reference of 1.88. Matching latent distributions gives 2.47, whereas combining the four-step target with pretrained visual features reaches 1.72. The B/2 controls show the same ordering across these representations. Thus a stronger target alone does not determine the adapted model's quality: the representation used to compare generated distributions affects how much of its advantage survives. This connects the component experiments to the sampling-gain view and explains why M2O uses both a verified better sampler and a frozen pretrained visual representation.

## 6 Conclusion

A useful one-step checkpoint can contain its own better teacher in a multi-step sampling mode. When the quality lost in distillation is smaller than the sampling gain, that advantage survives in one-step generation. Our central empirical finding is that this can happen with very little additional compute: M2O improves iMF-B/2 from 3.35 to 3.00 FID-50K in 0.60 H200-hour and reaches 1.72 on iMF-L/2 in 0.93 hour, improving on its guidance-calibrated 1.88 baseline. The fixed-CFG=5.00 control changes 2.31 to 1.72. These are adaptation costs after teacher selection, with no training data accessed during adaptation or extra inference calls. Teacher-step and objective controls identify when this short transfer works. These results motivate treating a checkpoint's multi-step quality margin as a practical resource for improving its one-step mode.

## Reproducibility statement

The supplementary archive contains source code, configs, training reports, evaluation protocols, and the compute ledger. Appendix A specifies the settings needed to reproduce the reported results.

## AI USE STATEMENT

Generative AI tools assisted with experiment design, code implementation, analysis of local records and literature, and drafting. They were also used to generate the schematic figures. The authors reviewed all AI-assisted work and take responsibility for the final experiments, source code, citations, claims, and text.

## REFERENCES

Sina Alemohammad, Zhangyang Wang, and Richard G. Baraniuk. Neon: Negative extrapolation from self-training improves image generation. *arXiv preprint arXiv:2510.03597*, 2025.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1lUOzWCW`.

Hansheng Chen, Kai Zhang, Hao Tan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. $\pi$-flow: Policy-based few-step generation via imitation distillation. *arXiv preprint arXiv:2510.14974*, 2025.

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

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b.

Xiao He, Yang Li, Peizhen Zhang, Songtao Liu, Zhao Zhong, and Nannan Wang. Stabilizing, scaling & enhancing MeanFlow for large-scale diffusion distillation. *arXiv preprint arXiv:2605.17834*, 2026.

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

Zheyuan Hu, Chieh-Hsin Lai, Ge Wu, Yuki Mitsufuji, and Stefano Ermon. MeanFlow transformers with representation autoencoders. *arXiv preprint arXiv:2511.13019*, 2025.

Yeongmin Kim, Sotiris Anagnostidis, Yuming Du, Edgar Schönfeld, Jonas Kohler, Markos Georgopoulos, Albert Pumarola, Ali Thabet, and Artsiom Sanakoyeu. Autoregressive distillation of diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yiyang Lu, Qiao Sun, Xianbang Wang, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Bidirectional normalizing flow: From data to noise and back. *arXiv preprint arXiv:2512.10953*, 2025.

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

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40, 2024.

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.

Yansong Peng, Kai Zhu, Yu Liu, Pingyu Wu, Hebei Li, Xiaoyan Sun, and Feng Wu. FACM: Flow-anchored consistency models. *arXiv preprint arXiv:2507.03738*, 2025.

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

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Nikita Starodubcev, Ilya Drobyshevskiy, Denis Kuznedelev, Artem Babenko, and Dmitry Baranchuk. Scale-wise distillation of diffusion models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Z06LNjqU1g.

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.

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

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

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*, 2018.

An Zhao, Shengyuan Zhang, Zhongjian Sun, Yixiang Zhou, Zejian Li, Ling Yang, Tianrun Chen, and Lingyun Sun. Mean flow distillation: Robust and stable distillation for flow matching models. *arXiv preprint arXiv:2606.11155*, 2026.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/zhou24x.html.

# A IMPLEMENTATION AND EVALUATION PROTOCOL

**Optimizer and parameter mask.** Both sizes initialize from their corresponding released iMF checkpoint. The AdamW learning rate is $10^{-6}$ with zero weight decay and $\beta_2 = 0.95$. The archived setting `trainable_imeanflow_u_head_from: 0` updates all `u_heads` blocks and `u_final_layer` within `net`, and any leaves outside `net`; other generator leaves remain frozen. The B/2 mask updates 56,700,992 of 88,974,400 parameters; the L/2 mask updates 100,698,128 of 409,054,224 parameters. These are exact counts. All adaptation runs use this mask. The main B/2 run uses batch 16 and 6K updates (96K noise–label pairs), whereas L/2 uses batch 4 and 16K updates (64K pairs). All reported students use raw parameters.

**Synthetic data and frozen representations.** Training uses deterministic noise–label pairs from sample IDs starting at zero, with training seed 42 and sampling seed 20260831. Labels are drawn uniformly and independently in the primary runs. No target images, target latents, real-feature caches, or real-reference moments enter adaptation. Teacher and student endpoints are decoded using the released iMF codec, then resized to $224 \times 224$ with bicubic interpolation for the frozen ImageNet-22K-pretrained ConvNeXtV2-Tiny global-mean representation. Teacher features and their RMS normalizer are stop-gradient. Gradients through the frozen image encoder and VAE still reach the student. The finite-batch objective and feature normalization are specified in Appendix C.

**Paired perceptual comparator protocol.** For the L/2 fixed-$K = 4$ setting, two controls retain the released initialization, independent noise–label IDs, optimizer mask, one-step inference settings, and zero endpoint, anchor, and hidden-loss weights. The paired ConvNeXt control uses the same decoded images, $224 \times 224$ resize, frozen global-mean representation, and stop-gradient teacher RMS as the distribution arm, replacing only its minibatch kernel with normalized matched-pair feature MSE. The LPIPS control applies frozen official LPIPS-VGG v0.1 weights to paired decoded RGB images clipped to $[-1, 1]$; the VAE and LPIPS computation remain differentiable with respect to the student. On synthetic image pairs, its JAX value and image-input gradient agree with the official PyTorch implementation within $10^{-8}$ RMS/absolute error. Teacher-image gradients are stopped. Both controls use unit loss weight. We record full single-H200 training wall time and separately audit raw one-NFE FID-50K. For a longer-run equal-time comparison, a checkpoint is chosen by proximity to the distribution arm's measured training time before reading that checkpoint's FID. At 4K updates, distribution matching and paired ConvNeXt score 1.83 and 1.82 after 0.33 hour each, while LPIPS scores 3.58 after 0.32 hour. The preselected longer comparison uses distribution step16K at 0.9336 hour, paired ConvNeXt step13K at 0.9167 hour, and LPIPS step14K at 0.9308 hour; their FID-50K values are 1.72, 2.04, and 3.37. Direct step16K paired ConvNeXt and LPIPS score 2.06 and 3.40 after 1.0656 and 1.0325 hours. All six paired evaluations use the same 50K bank; the detailed raw reports and full precision costs are included in the supplement.

The B/2 paired controls use the primary eight-step teacher and one-step student settings in Table 5, batch 16, 6K updates, and 96K independently labelled noise samples. They retain the primary initialization, parameter mask, optimizer, and zero endpoint/anchor/hidden weights, with the same paired feature objectives described above. Paired ConvNeXt reaches 3.84 FID-50K in 2104 seconds (0.5844 H200-hour); LPIPS-VGG reaches 5.43 in 2082 seconds (0.5783 H200-hour). Each completes training in one process, including setup and compilation, then evaluates raw checkpoint 6000 on the common 50K bank with batch 8. All logged losses and gradient norms are finite. These are fixed-budget controls rather than an objective-specific hyperparameter search.

Table 5: Primary sampler settings. Scales and interval endpoints are retained at the precision required for reproduction. $K = 1$ students have one conditional branch and one network evaluation.

| Model/mode | $K$ | Guidance | Interval | Noise scale | Output scale |
|---|---|---|---|---|---|
| B/2 baseline/student | 1 | 7.75 | $[0.38, 0.65]$ | 0.975 | 0.975 |
| B/2 self-teacher | 8 | 4.60 | $[0.43, 0.65]$ | 1.00 | 1.00 |
| L/2 matched baseline/student | 1 | 5.00 | $[0.40, 0.60]$ | 1.00 | 1.00 |
| L/2 matched self-teacher | 4 | 5.00 | $[0.40, 0.60]$ | 1.00 | 1.00 |

Table 6: Frozen ResNet-50V2 diagnostics on 50K generated images. B/2 teacher/student sampler settings differ; L/2 settings are shared.

| Arm | Top-1 (%) | Top-5 (%) | Within-class distance |
|---|---|---|---|
| B/2 calibrated baseline | 87.21 | 97.16 | 0.47 |
| B/2 primary student | 87.86 | 97.43 | 0.46 |
| B/2 selected teacher | 86.98 | 97.09 | 0.48 |
| B/2 same-class student | 86.95 | 97.14 | 0.47 |
| L/2 same-CFG baseline | 84.91 | 96.51 | 0.51 |
| L/2 primary $K = 4$ student | 87.08 | 97.31 | 0.49 |
| L/2 paired ConvNeXt 4K student | 87.87 | 97.72 | 0.48 |
| L/2 $K = 4$ teacher | 88.07 | 97.88 | 0.49 |
| L/2 same-class student | 87.66 | 97.61 | 0.49 |

**Evaluation separation.** Sampler selection uses a 5,000-sample development bank with seed 20260831 and ID offset 500000. Every FID comparison measured in this work uses 50,000 generated samples, the same seed, offset 700000, and strictly balanced labels $y = \mathrm{sample_id} \bmod 1000$. Evaluation uses the same real-reference moments and CPU Inception implementation with float32 feature export. The latent transform is $z_{\mathrm{VAE}} = z_{\mathrm{normalized}} \odot \mathrm{std} + \mathrm{mean}$, with $\mathrm{mean} = (0.86488, -0.27787, 0.21617, 0.37384)$ and $\mathrm{std} = (4.85504, 5.31922, 3.93725, 3.98700)$. Reference moments are used for post-training evaluation and sampler/model selection, so dataset-free refers to adaptation rather than the entire selection pipeline. Repeated development comparisons on the 50K bank do not constitute an untouched test set. Reports record seeds, ID offsets, codec, branch count, and NFE. The final and independent 22-file rerun archives are SHA-256 verified; raw export verifies parameter count and serialization.

**Cost boundary.** Adaptation time includes initialization, compilation, online teacher sampling, VAE decoding, feature extraction, optimization, and checkpointing on one H200. It excludes checkpoint pretraining, sampler selection, and final evaluation. The recorded L/2 eight-step sampler search costs 5.17 H200-hours; a five-point four-step follow-up costs 1.07 H200-hours. These are additional evaluation costs, not included in the 0.93-hour adaptation claim. The B/2 discovery process includes earlier experiments whose full search cost is not summarized by the primary training figure. Therefore we claim low-cost adaptation once a teacher setting is chosen. The independent B/2 host rerun uses the same training seed and reaches 3.00 FID-50K in 0.59 H200-hour. Intermediate L/2 checkpoint times in Figure 3 are approximate.

**Class fidelity and coverage.** A frozen ResNet-50V2 checks 50K generated images per arm. Requested-class top-1 rises from 87.21% to 87.86% for B/2 and from 84.91% to 87.08% for fixed-guidance L/2. All 1,000 predicted classes remain covered; within-class feature distances contract slightly. Table 6 reports the full statistics.

## B Additional controls and interpretation

**Endpoint and initialization controls.** The B/2 endpoint-only arm uses latent MSE weight 0.1 and reaches 17.69 FID-50K in 0.23 H200-hour; adding a one-step anchor of weight 1.0 reaches 3.35 in 0.22 hour. The L/2 $K = 4$ endpoint-only arm instead uses weight 1.0 and reaches 13.33 in 0.67 hour. The endpoint-plus- distribution arms use endpoint weight 0.1. Randomizing only the trainable B/2 leaves while retaining the frozen pretrained leaves yields 346.03 FID-50K in 0.62 hour with endpoint-plus-distribution training. This failure supports the importance of the pretrained initialization in this short adaptation regime.

**Hidden alignment.** The B/2 matched pair augments endpoint-plus-distribution training with sparse readouts at student layers 2, 5, and 8 and teacher rollout states 2, 4, and 6. A normalized latent MSE coefficient of 0.001 reduces the final hidden loss from 0.96 to 0.77; both aligned and zero-weight arms score 3.05 FID-50K. The zero-weight arm computes the same readouts. Both arms use the endpoint-plus-distribution recipe.

**Exploratory class grouping.** The primary independent-label batches mainly compare images across classes. A subsequent same-class minibatch variant reaches B/2 FID-50K 3.05, 2.94, and 2.87 after 1K, 3K, and 6K updates, with costs 0.16, 0.33, and 0.59 H200-hour. A same-host independent-label 6K run reaches 3.00 in 0.59 hour. For L/2 at fixed $K = 4$, same-class training reaches 1.69 in 0.92 hour versus the primary 1.72. A round-robin cross-class arm reaches 1.75, and regrouping the same pair set into cross-class batches reaches 1.72, both in 0.92 hour. Batch composition changes the finite-batch objective and optimization order. Independent-label runs remain primary.

**Paired endpoint objectives.** Let $G_0$ be the frozen released one-step endpoint. The B/2 paired baseline is

$$\mathcal{L}_{\text{endpoint}} = 0.1\,\mathbb{E}\|G_\theta - T_K\|_2^2 + \lambda_0\,\mathbb{E}\|G_\theta - G_0\|_2^2, \tag{3}$$

using elementwise mean reductions, with $\lambda_0 = 0$ or 1. The L/2 endpoint-only control uses unit endpoint weight as specified above. The primary M2O objective contains neither endpoint nor anchor terms.

**Additional sampler controls.** The B/2 two-step control uses the primary teacher settings (CFG=4.60, interval $[0.43, 0.65]$) and student settings (CFG=7.75, interval $[0.38, 0.65]$, input/output scales 0.975). It retains the 6K-update, batch-16 budget and distribution-only objective. The teacher scores 2.89237 FID-50K; the student scores 3.06520 after 1783 seconds (0.4953 H200-hour), including setup and compilation. Reports and the verified source snapshot are in `evidence/b2_k2/` in the supplement.

In an exploratory B/2 same-CFG comparison, the released one-step baseline scores 4.06, one-step self-matching scores 4.04, and eight-step-to-one adaptation reaches 2.86 after 0.61 H200-hour. The student configs differ only in teacher step count. Extending the latter to 12K updates gives 2.89 in 1.10 hours. The main tables use the primary selected-sampler settings. The L/2 step-count comparison fixes guidance and all other sampling settings as reported in Table 2.

## C M2O OBJECTIVE DETAILS

For feature dimension $d_\phi$ and batch size $n$, the common teacher RMS is

$$c = \mathrm{sg}\left[\left(\frac{1}{nd_\phi}\sum_{i,j} t_{ij}^2 + 10^{-6}\right)^{1/2}\right], \qquad \tilde{s}_i = s_i/c, \quad \tilde{t}_i = t_i/c. \tag{4}$$

The primary recipe uses Equation 2 with unit weight and zero endpoint, anchor, and hidden-state weights. Its KID-style objective uses ConvNeXt features, batch-dependent teacher normalization, and exclusion of the paired cross diagonal. The finite-batch surrogate can be negative; the batch-dependent normalization differs from a fixed-population MMD estimator. Independent-label batches predominantly contain cross-class off-diagonal pairs. The pretrained initialization carries the class structure, which is evaluated by the classifier diagnostics in Table 6.

Figure 6 illustrates the generated-distribution objective in Section 3. It uses $u_i = \tilde{s}_i = s_i/c$ and $v_i = \tilde{t}_i = t_i/c$ to keep the normalized-feature notation compact. The paired teacher and student endpoints share sample IDs, and their cross-term contributions use different IDs. The student receives gradients through the frozen VAE and visual encoder; the teacher features and common RMS scale are stop-gradient.

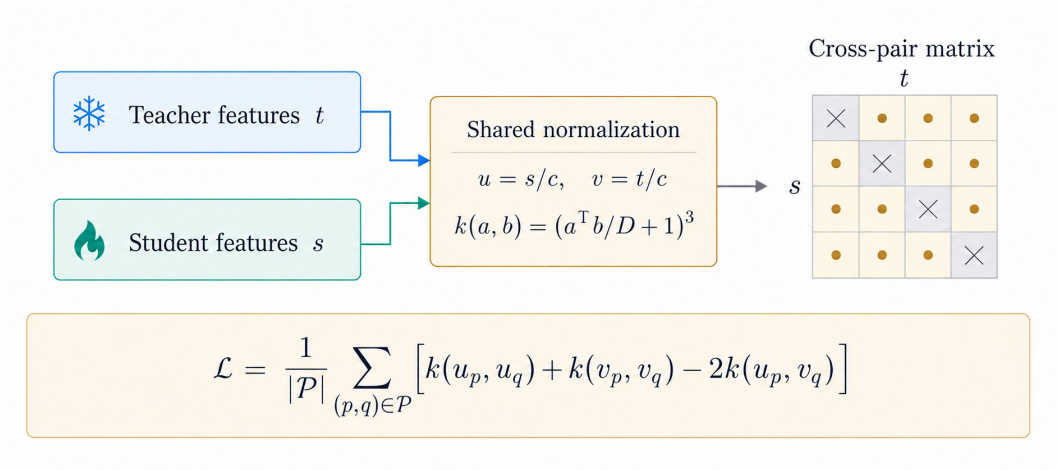



Figure 6: M2O's KID-style distribution objective. The cubic kernel compares normalized global features across a minibatch, excluding self diagonals and the paired teacher–student cross diagonal. Here $\mathcal{P} = \{(p,q) : 1 \le p,q \le n,\ p \ne q\}$, so $|\mathcal{P}| = n(n-1)$. The $4 \times 4$ matrix illustrates the cross-term index mask, with student rows and teacher columns; crosses are excluded and dots retained. B/2 uses batch 16 and L/2 batch 4. The common RMS normalizer is computed from the teacher batch; $D$ in this schematic denotes the feature dimension $d_\phi$, not the evaluation penalty in Equation 1. Both branches use the same frozen encoder; the student branch backpropagates through it to selected generator parameters. The primary recipe uses only this distribution term.

## D LIMITATIONS AND FURTHER DIRECTIONS

The experiments cover two iMF sizes on ImageNet-256, with one training seed per key arm. B/2 uses selected teacher/student guidance; L/2 uses fixed guidance. The ConvNeXtV2 encoder is pretrained on ImageNet-22K; its visual prior is therefore derived from the same broader ImageNet collection as the ImageNet-1K benchmark. Pretrained iMF and VAE weights also supply image-derived prior knowledge. Dataset-free describes access during adaptation. Real-reference moments enter evaluation and model selection, with compute accounting given in Appendix A. Further directions include broader model families, full-model adaptation, independent training seeds, and real-reference diversity evaluation. Classifier agreement and predicted-class coverage measure conditional fidelity but

do not establish real-distribution coverage. The L/2 recipe uses four images per update; an equal-budget batch-size study and repeated-seed estimates of the finite-batch objective remain open. Direct comparisons with alternative distillation methods and continued iMF training would further characterize how to allocate a small adaptation budget; original iMF training would require a different, real-data-access protocol.

## E Complete shared-noise qualitative comparisons

Figures 7 and 8 show all eight examples declared before rendering. Every column shares the same noise and class label across baseline, teacher, and student. The raw checkpoint, config, source, and image hashes are archived in `evidence/fixed_id_grids/`. The original images are preserved without retouching; the main-text panels use identical crop coordinates across rows, recorded in `figures/qualitative_selection.json`. The main text selects four full-image cases and three detail cases after inspecting these grids. Full-image panels retain the complete image extent; the detail panels use the matched regions recorded in the selection file. B/2 uses its primary eight-step teacher and student settings. L/2 uses the primary four-step-teacher student trained for 0.93 H200-hour, with CFG=5.00 for all three rows.

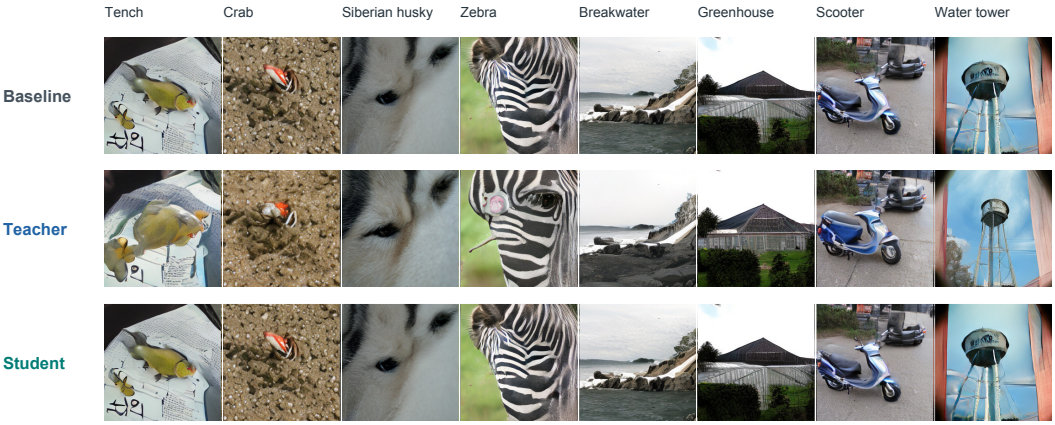



Figure 7: Complete predeclared B/2 bank. Rows: released one-step baseline, frozen eight-step self-teacher, and primary one-step M2O student. Teacher and student use their respective primary sampling settings.




Figure 8: Complete predeclared L/2 bank. Rows: released one-step baseline, frozen four-step self-teacher, and primary one-step M2O student (0.93 H200-hour). All rows use fixed CFG=5.00, the same guidance interval, and unit input/output scales.

## F UNCURATED ONE-STEP SAMPLES

Figures 9 and 10 show 128 uncurated samples from the primary M2O students. Each row contains the first eight evaluation samples for a class chosen before generation. All samples are shown in order, using the primary settings in Table 5.

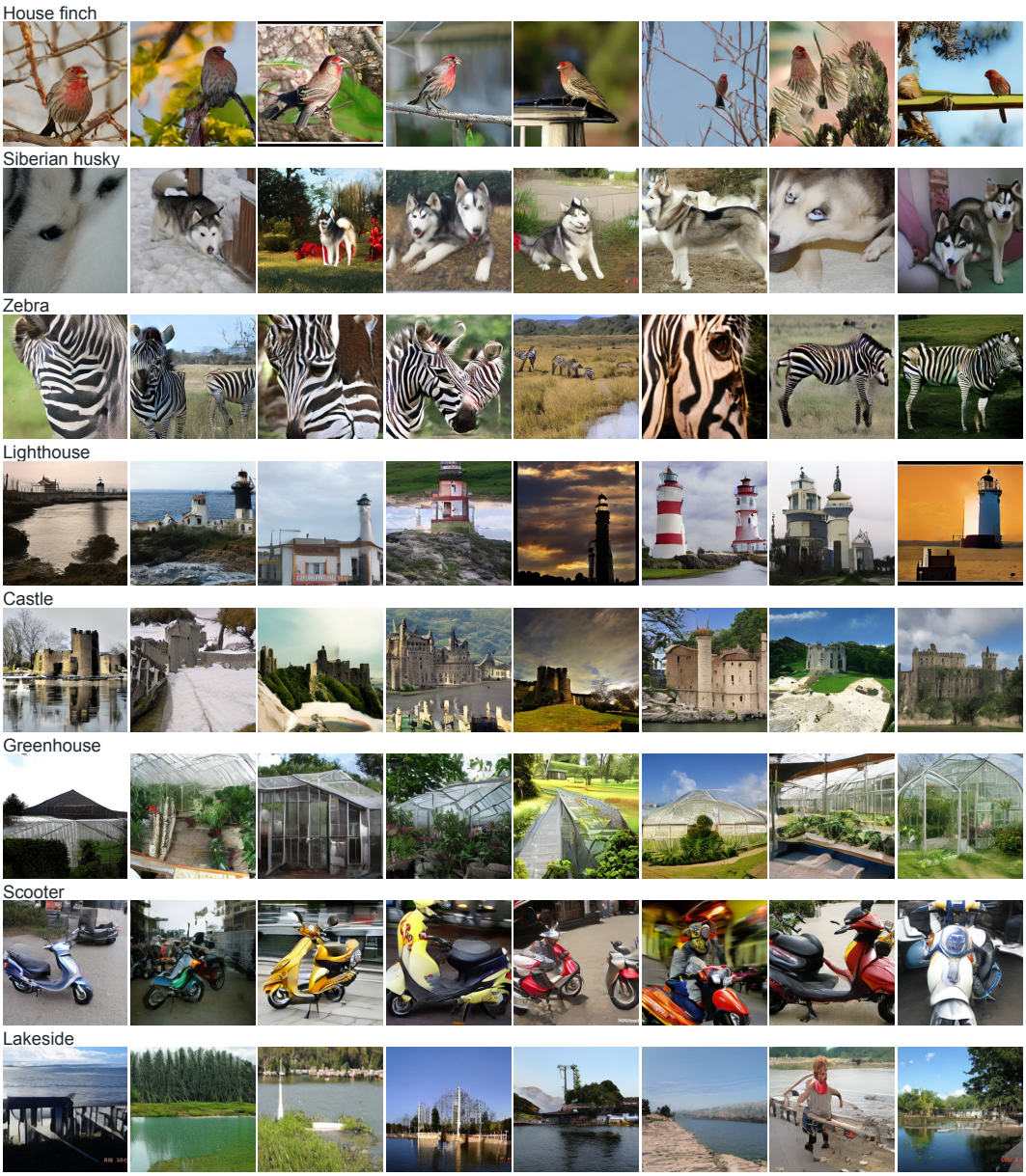



Figure 9: Uncurated one-step generation from iMF-B/2 + M2O (FID-50K 3.00; 0.60 H200-hour adaptation). Each row shows eight samples of the indicated class, with complete images and no sample filtering.

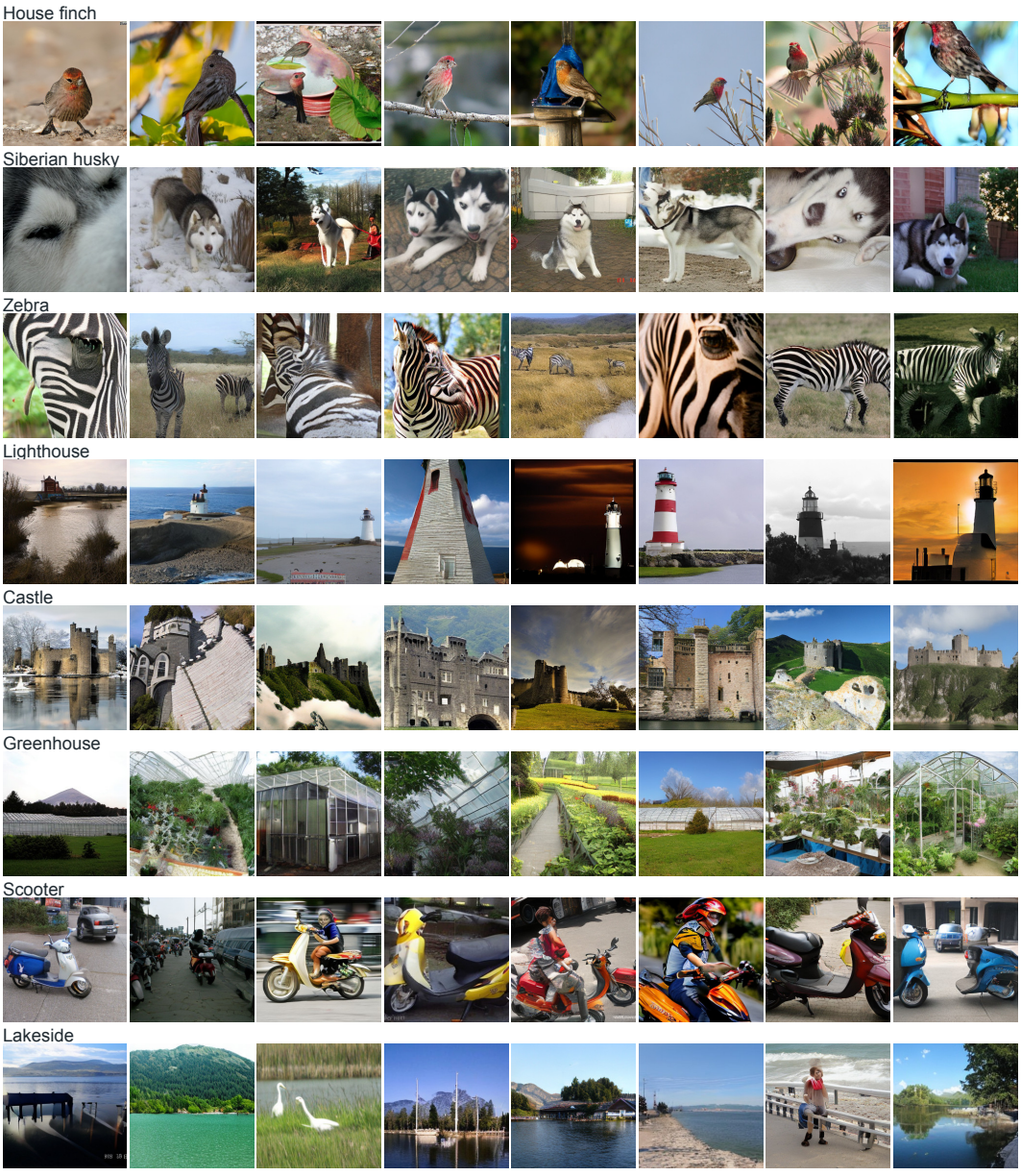



Figure 10: Uncurated one-step generation from iMF-L/2 + M2O (FID-50K 1.72; 0.93 H200-hour adaptation). Class order and noise IDs match Figure 9. The primary four-step-teacher student uses CFG=5.00; all eight samples per class are shown without filtering.