BEYOND FASTER FIELDS: AUDITING DRIFTING GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Drifting generative models train a generator through an attraction–repulsion field that moves samples toward a target distribution. We ask whether cheaper fields preserve the intended update, reduce full training cost, and maintain distributional fidelity. We compare optimized dense drifting with projected, representative-based, and transport alternatives in a 240-run feature benchmark, a 24-run image screen, and a 66-model generator–encoder audit. An exact aggregation identity shows how to preserve normalization and copy-specific self deletion when compressing repeated sources. Compatible operator references change apparent approximation error. A $9.26\times$ field saving yields only $1.013\times$ inclusive wall speedup in the balanced image screen. At 15k updates in the balanced small-generator/ResNet-18 bridge, DriftXpress uses about 26.7% less recorded training-component time, with lower $\mathrm{FID}_{\mathrm{tv}}$ and higher recall in all three paired seeds under both tested Inception input paths; earlier recall gains are input-sensitive. All twelve rare-target screen runs have zero undercoverage, yet classifier-estimated rare-mass means are 5.6–$9.9\times$ the target. These results locate a positive checkpoint regime without establishing general transfer or exact matched-quality speedup. The practical takeaway is to compare native operators, declare cost scope, and measure quality, diversity, and two-sided mass error.

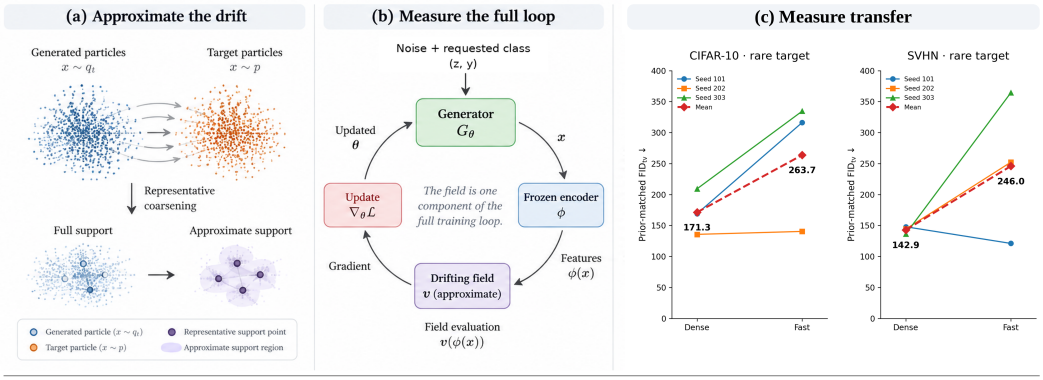


Figure 1: **From field approximation to training outcomes.** **(a)** Schematic generated particles $x \sim q_t$ approach target p; representatives and shaded supports depict coarsening, not uncertainty or model samples. **(b)** Noise z and class y enter generator G_θ; image x passes through frozen encoder ϕ. Field v defines a training target; loss $\mathcal{L}$ supplies the gradient updating θ. Preparation is omitted; sizes do not encode runtimes. **(c)** Under the historical Inception input path, *prior-matched* $\mathrm{FID}_{\mathrm{tv}}$ is the Fréchet distance between Gaussian fits to frozen torchvision Inception-v3 features (lower is better; not a percentage). Each comparison uses 9,100 generated and 4,550 real images with rare target prior $p_9 = 1/91$. Lines pair seeds 101, 202, 303; dashed diamonds are means, not uncertainty intervals. Dense is the optimized joint field; Fast is regime-adaptive. These are complete rare-target dense/Fast comparisons within the 66-run audit. Recovery and metric checks are reported in Appendix G; low recall limits the quality interpretation. This panel establishes neither time to matched quality nor a universal method ranking.

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.

1 INTRODUCTION

Drifting generative models train through attraction to target samples and repulsion from generated samples, then generate in one forward pass (Deng et al., 2026). We audit clean-room implementations with declared training protocols, not reproductions of published recipes or their best attainable quality. Cheaper fields help only insofar as their operator, workload share, and learned distribution permit.

We ask: *when do field-level savings become useful training gains, and where do they stop transferring?* The small-generator/ResNet-18 bridge within our 66-model CIFAR-10/SVHN audit provides a positive case: at 15k updates, DriftXpress uses 26.7% less recorded training-component time, with lower $\mathrm{FID}_{\mathrm{tv}}$ and higher recall in all three paired seeds under both tested Inception input paths (Table 1). At 10k, its FID advantage survives the input change, but recall gains do not. This is a localized cost–quality benefit, not high absolute recall, an exact crossing-time speedup, or evaluator-independent dominance.

Table 1: **Final-checkpoint gains survive both tested input paths.** Balanced CIFAR-10, small generator/ResNet-18; three paired seeds, fixed EMA states and images. T: recorded training-component hours, not full allocation; F: $\mathrm{FID}_{\mathrm{tv}}$; R: recall. Off/on denote `transform_input=False/True`; only this internal flag changes. Joint FID/recall gains hold in all pairs on both paths at 15k, not 10k. Appendices G.4–G.5 retain all outcomes.

	T (h)↓		$F\downarrow$		$R\uparrow$	
Updates/flag	Dense	DriftXpress	Dense	DriftXpress	Dense	DriftXpress
5k, off	2.87	2.11	195.35	194.37	0.00007	0.00013
10k, off	5.75	4.21	95.44	94.22	0.00493	0.01227
15k, off	8.62	6.31	66.13	63.44	0.02387	0.03140
5k, on	2.87	2.11	232.00	231.20	0.00000	0.00000
10k, on	5.75	4.21	106.53	104.46	0.00327	0.00300
15k, on	8.62	6.31	67.18	64.53	0.01527	0.02147

The benefit has a boundary. Across generator–encoder configurations, 41–$54\times$ field savings yield only 1.048–$1.365\times$ training speedups. Under the historical input, DriftXpress improves terminal FID in all six small-generator bridge pairs but only two of six U-Net pairs. Dense and DriftXpress also differ in normalization and positive-support access, so these are complete-configuration comparisons, not projection alone. Neither input path nor raw weights rescue recall in the six-model high-capacity diagnostic.

A 240-run feature benchmark separates compatible-field fidelity from quality and cost. A 24-run image screen exposes limited full-workload leverage and excess classifier-estimated rare mass despite zero undercoverage. Figure 1(c) retains a complementary rare-target failure example. Our contributions are exact duplicate aggregation with copy-specific self deletion and an operator-contract audit separating component savings, checkpoint gains, and evaluator-sensitive rankings. DriftXpress already measures preprocessing and wall-clock convergence (Falahati et al., 2026); our added evidence tests which conclusions survive changes in workload and evaluator input. The reverse-triangle inequality, chain rule, and Amdahl's law are standard analytical tools.

2 RELATED WORK

Drifting and distributional dynamics. Diffusion models (Ho et al., 2020), flow matching (Lipman et al., 2023), and consistency models (Song et al., 2023) offer different training and sampling trade-offs. Drifting instead evolves a generator's pushforward distribution during training (Deng et al., 2026). Kernel-based particle dynamics have antecedents in Stein variational inference (Liu & Wang, 2016). Score-difference interpretations depend on kernel assumptions (Turan et al., 2026); Wasserstein-flow analyses distinguish an idealized field from the normalized operator actually executed (Gretton et al., 2026). We study finite-support executable contracts, not a new distributional flow theorem.

Approximation versus a different operator. Nyström approximation and random features reduce kernel costs (Williams & Seeger, 2001; Rahimi & Recht, 2007); DriftXpress projects attraction in

a reproducing kernel Hilbert space (RKHS) while retaining native repulsion (Falahati et al., 2026). Sinkhorn drifting (He et al., 2026) uses entropic transport, related to regularized optimal transport and Sinkhorn divergences (Cuturi, 2013; Feydy et al., 2019). Exact tiled reductions, including KeOps and FlashAttention, illustrate a distinct memory–computation trade-off (Charlier et al., 2021; Dao et al., 2022). None licenses replacing normalization or copy-specific self deletion without checking the resulting field.

Evaluation and complete cost. Maximum mean discrepancy (MMD) compares distributions through kernels (Gretton et al., 2012). Fréchet inception distance (FID), kernel inception distance (KID), and precision/recall expose complementary aspects of generated distributions (Heusel et al., 2017; Bińkowski et al., 2018; Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020). Finite-sample effects and preprocessing can alter their interpretation (Chong & Forsyth, 2020; Parmar et al., 2022). We distinguish a balanced reference from a depleted training prior. Systems benchmarks separate throughput from time to a quality target (Mattson et al., 2020); recomputation can intentionally trade time for memory (Chen et al., 2016). Variance, leakage, and the independent unit also constrain comparisons (Bouthillier et al., 2021; Kapoor & Narayanan, 2023).

3 What Does an Acceleration Claim Preserve?

Let $X = (x_i)_{i=1}^n \in \mathbb{R}^{n\times d}$ be generated queries in feature dimension d, $P \in \mathbb{R}^{m_+\times d}$ positive-source payloads, and $N \in \mathbb{R}^{n\times d}$ indexed generated-source payloads. Let $K^+ \in \mathbb{R}^{n\times m_+}$ and $K^- \in \mathbb{R}^{n\times n}$ be nonnegative kernel blocks, with each aligned generated self entry K^-_{ii} deleted before normalization. Write $K = [K^+\ K^-]$, let $\mathbf{1}$ denote a conformable all-ones vector, and let diag form a diagonal matrix. For strictly positive row degrees r and column degrees c,

$$r = K\mathbf{1}, \qquad c = K^\top \mathbf{1}, \qquad A = \mathrm{diag}(r)^{-1/2} K\, \mathrm{diag}(c)^{-1/2}. \tag{1}$$

The joint contract used by our dense baseline forms scalar masses and vector payloads,

$$p_0 = A^+\mathbf{1}, \quad p_1 = A^+P, \quad n_0 = A^-\mathbf{1}, \quad n_1 = A^-N, \qquad V_i = (n_0)_i(p_1)_i - (p_0)_i(n_1)_i. \tag{2}$$

Here A^+ and A^- are the corresponding blocks of A. With normalized payload means $\mu_i^+ = (p_1)_i/(p_0)_i$ and $\mu_i^- = (n_1)_i/(n_0)_i$, $V_i = (p_0)_i(n_0)_i(\mu_i^+ - \mu_i^-)$ whenever both masses are nonzero. The implementation also fixes numerical floors and boundary behavior; changing them changes the executable contract.

Incompatible error contains a different-field discrepancy. For a candidate V_A, its compatible reference V_C, and an incompatible reference V_I, the reverse-triangle inequality in the Frobenius norm $\|\cdot\|_F$ gives

$$\big|\|V_A - V_I\|_F - \|V_C - V_I\|_F\big| \le \|V_A - V_C\|_F. \tag{3}$$

Consequently, even a nearly exact candidate can have substantial error against a different field. This is not permission to choose a favorable reference after observing quality: the contract is declared before comparison, and final quality is assessed independently.

Counts must respect normalization and the self mask. Replacing m identical columns by one column weighted by m *before* two-sided normalization is generally not equivalent to explicit duplication. For normalization exponents α, β, write $D_r = \mathrm{diag}(r)$ and $D_c = \mathrm{diag}(c)$. In $D_r^{-\alpha} K D_c^{-\beta}$, the grouped duplicate contribution scales as m, whereas preweighting also changes the column degree and produces $m^{1-\beta}$. At $\beta = 1/2$, the latter contributes $\sqrt{m}$, not m, with the row-degree factor held common. Copy-specific self deletion introduces a further distinction: equal representative payloads need not have equal masked column degrees.

Proposition 1 (Exact aggregation with copy-specific deletion)**.** *Fix query and representative memberships. Suppose generated column j has representative $z_{g(j)}$, so $U_{i\ell} = k(x_i, z_\ell)$ and $E_\ell = \sum_i U_{i\ell}$. Delete only the aligned entry (j, j) before computing degrees. Let $c_j = E_{g(j)} - U_{j,g(j)} > 0$, let $r_i > 0$ be the corresponding joint row degree, and set $S_\ell = \sum_{j:g(j)=\ell} c_j^{-\beta}$. Then the exact grouped negative coefficient is*

$$B^-_{i\ell} = r_i^{-\alpha} U_{i\ell} \left(S_\ell - \mathbf{1}\{g(i) = \ell\} c_i^{-\beta} \right). \tag{4}$$

Applying these coefficients to group-constant scalar or vector payloads equals explicit duplicate expansion.

Table 2: Native contracts and appropriate fidelity references. Quality can be compared across configurations. Native-operator fidelity requires a compatible reference; discrepancy to another operator is a different, well-defined quantity. Exactness of a coarsened support is not exactness to the original uncompressed support.

Route	Native operation	Compatible reference
Optimized dense	Joint geometric-mean normalization, aligned self mask	Same dense field and loss
DriftXpress	Projected attraction; exact self-masked repulsion	Exact separate-row native field
Fast / quality	Weighted representative joint field; frozen selection/teacher schedule	Full joint field, with approximation error retained
Count-expanded	Exact grouped coefficients on duplicate-coarsened support	Explicit duplicate expansion with the same copy-specific mask
Sinkhorn	Split entropic positive/generated couplings	The same transport contract; not joint geometric-mean scaling

The proof simply groups the surviving columns, but preserving their distinct masked degrees is essential. Appendix A gives the positive-source formula, proof, complexity, and counterexample to naive preweighting. The result does not certify nearest-representative approximation, changing memberships, or derivatives through discrete assignments.

The generator introduces a separate transfer question. Let g_θ be a generator with parameter vector $\theta \in \mathbb{R}^p$, let z be a fixed noise batch, and let h be the frozen field encoder (the identity for direct feature generation). Write $X_\theta = h(g_\theta(z))$. For positive drift scale η and loss scale a, held fixed during differentiation, define the detached-target regression

$$L(\theta) = \frac{a}{2}\|X_\theta - \mathrm{sg}(X_\theta + \eta V)\|_F^2, \qquad \nabla_\theta L = -a\eta J_\theta^\top \mathrm{vec}(V), \tag{5}$$

where sg stops differentiation through its entire argument, vec stacks matrix entries, and $J_\theta \in \mathbb{R}^{nd\times p}$ is the Jacobian of $\mathrm{vec}(X_\theta)$. This identity requires the same features, model state, inputs, and stochastic realization in target construction and differentiable recomputation. The extension uses sample-indexed deterministic structured masking; Appendix H specifies the two-pass contract and its numerical limits. Field error and generator-gradient error can differ because of Jacobian amplification or cancellation. Equality at one state does not establish equal trajectories without matching optimizer state and all subsequent inputs. We therefore measure both, then test—rather than assume—whether pullback diagnostics improve held-out prediction.

4 Experimental Design

Feature-space benchmark. The matrix is four datasets × two target regimes × six methods × five paired seeds, totaling 240 runs. MNIST and Fashion-MNIST use sample-normalized flattened pixels. CIFAR-10 and Street View House Numbers (SVHN) use frozen ImageNet-1K ResNet-18 pooled features, followed by sample ℓ_2 normalization (LeCun et al., 1998; Xiao et al., 2017; Krizhevsky, 2009; Netzer et al., 2011; He et al., 2016). Balanced targets contain 1,000 training examples per class; rare targets retain 100 class-9 examples and 1,000 in each other class. All methods use the same generator, optimizer, batch size 512, 1,000 updates, and checkpoints $\{0, 100, 500, 900, 1000\}$. Seeds are $\{11, 23, 37, 53, 71\}$, and method order rotates within paired blocks. All 240 rows completed with zero scientific failures or replacements.

Feature quality is measured by the preserved energy-distance V-statistic and radial-basis-function (RBF) maximum mean discrepancy (MMD). "MMD" averages four unbiased MMD-squared estimators, not their square roots (Appendix B).

The six routes are optimized dense, DriftXpress, regime-adaptive Fast, probe/teacher Quality, count-expanded aggregation, and Sinkhorn. Fast/Quality name operating points, not guarantees. Appendix B specifies supports, label use, and teacher costs.

Independent raw-image screen. A conditional residual generator trains through frozen nonlinear ResNet-18 features on CIFAR-10: four methods, two regimes, seeds $\{42, 43, 44\}$, 10,000 updates and batch 512 give 24 runs. Technical and narrow non-collapse checks pass without replacements, not proof of useful diversity. Independent torchvision Inception-v3 measures image quality; a CIFAR classifier measures class mass. $\mathrm{FID}_{\mathrm{tv}}$/$\mathrm{KID}_{\mathrm{tv}}$ and precision/recall follow their metric families (Heusel et al., 2017; Bińkowski et al., 2018; Kynkäänniemi et al., 2019), but are not interchangeable across preprocessing implementations. Final exponential-moving-average (EMA) states generate 10,000 samples for FID and 5,000 for KID/manifold metrics. Both regimes share a balanced 5,000-image real reference, so rare FID/KID are not prior-matched. Appendix B specifies architecture, optimizer, and checkpoints.

Completed generator–encoder extension. A separate 66-run audit uses 15,000 updates, batch 512, and paired seeds $\{101, 202, 303\}$: 48 U-Net/DINOv3 runs cover two datasets, two regimes and four methods; 18 additional CIFAR-10 runs complete a two-generator/two-encoder bridge. All final models have balanced-reference and prior-matched evaluations, including six evaluation-only recoveries from unchanged checkpoints (Appendix G). Labels condition generators and guide support selection, but fields match the *marginal* distribution without same-class affinities; requested priors do not certify semantic correctness. Dense uses joint normalization with a fresh positive minibatch; DriftXpress uses separate-row normalization with cached full-target information. An information-matched exact separate-row training control is absent (Appendix Table 23). A follow-up evaluates saved 5k/10k/15k EMA checkpoints for the six small-generator/ResNet-18 Dense/DriftXpress models, plus final raw and EMA states for the corresponding six U-Net/DINOv3 models. All 30 use a fixed paired protocol, without training (Appendix G.4).

Quality-evaluator scope. "Historical-input" (off) disables the pretrained Inception factory's internal input transform; "transform-on" enables it. The identical-image test toggles only `transform_input` for all 30 checkpoint/state image sets and the same real reference: 60 evaluations, no training (Appendix G.5). Both retain our custom spatial preprocessing; neither is canonical TensorFlow FID. Weights, pixels, class predictions, and training costs stay fixed. FID/KID, recall, and spread share features, so they are not independent robustness tests; class mass uses a separate classifier.

Inference and evidence levels. The feature study uses five paired seeds per cell; the two image studies use three. Joint non-inferiority (NI) requires at most 5% relative degradation in energy and MMD, no coverage loss, and at most 0.02 additional rare undercoverage. The rare deficit is bounded by its empirical evaluation prior $6/512$, making that last margin nonbinding. We therefore interpret the reported decisions as energy/MMD/coverage comparisons, not rare-calibration guarantees. Appendix D gives the component confidence bounds and separates interpretable negative predictor results from invalid targets. Historical hypothesis labels and role predicates are retained there and in Appendix F, rather than treated as new contributions. Three of the five feature seed identifiers overlap earlier screening; frozen tests do not imply wholly fresh discovery-independent data. Neither three-seed image consistency nor retrospective reanalysis supplies a population-level transfer guarantee.

5 Feature-Space Fidelity Does Not Select a Universal Winner

Compatible references materially change apparent error. The median family-level difference $e_I - e_C$ is 0.509 across four image datasets and a controlled two-dimensional family, where $e_R = \|V_A - V_R\|_F / \max(\|V_R\|_F, 10^{-12})$. This is an *absolute difference between separately normalized errors*, not a 50.9% fractional reduction or a quality gain. The family summaries and inferential scope are given in Appendix D. DriftXpress's final compatible error is approximately 0.0342, with mean cosine 0.9994 and norm ratio 1.001. Its native-contract label remains attached to the end-to-end comparison.

Quality, runtime, and memory expose different operating points. Table 3 shows descriptive equal-weight averages, not a universal ranking. DriftXpress improves average energy and MMD while reducing wall time from 4.057 to 3.681 seconds, but increases memory. The quality schedule averages 3.377 seconds with a different quality trade-off. Fast compression is particularly sensitive to

Table 3: Feature-space benchmark: equal-weight means over four datasets and two regimes, each with five seeds (240 runs). Energy and MMD are multiplied by 10^3; MMD denotes averaged unbiased MMD-squared. Memory is peak allocated CUDA memory. These descriptive averages are not a pooled confidence interval. NI counts dataset–regime cells passing all applicable frozen quality margins against optimized dense. Per-cell uncertainty is in Appendix C.1.

Method	Energy ↓	MMD ↓	Wall (s) ↓	MiB ↓	NI / 8
Optimized dense	9.17	3.48	4.057	99.7	–
DriftXpress	8.60	3.20	3.681	137.0	5
Fast	19.85	8.08	3.396	102.4	0
Quality	9.67	3.81	3.377	103.6	0
Count-expanded	9.62	3.70	4.284	99.4	0
Sinkhorn	9.46	3.74	11.904	99.2	1

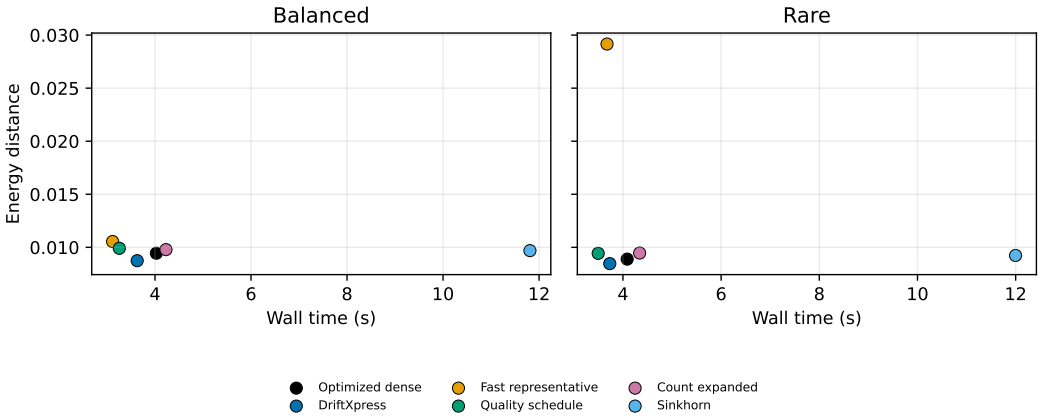


Figure 2: **Different quality–wall operating points in feature space.** Each point is one method's equal-weight mean over four datasets within a target regime, not an independent dataset observation; lower and left are preferable. These aggregate views complement, rather than replace, the paired per-dataset uncertainty and joint non-inferiority gates. Native operator contracts differ (Table 2); the shared comparison here is learned distribution quality and inclusive wall time.

the rare regime: its mean rare energy is 0.02916, compared with 0.00889 for optimized dense. Joint quality non-inferiority passes in 5/8 dataset–regime cells for DriftXpress, 1/8 for Sinkhorn, and 0/8 for fast, quality, and count-expanded. Failure of this gate means non-inferiority was not established; it is not automatically a significant inferiority result in every cell.

Each method has stable Pareto membership in at least three dataset-specific views under the fixed bootstrap rule (Appendix F). This supports conditional roles, not domination by all methods or by one preferred method. Count-expanded exactness is relative to duplicate-coarsened support, explaining why exact operator aggregation need not match original-support learning quality. Likewise, exact streamed and cached reverse paths produce different systems endpoints: at support 8,192, full caching gives only $1.065\times$ speedup over optimized dense, while streaming reduces memory by $5.47\times$ but takes approximately five times longer (Appendix C.3). A slower canonical dense implementation would overstate both advantages.

No transferable selector is established. Adding pullback diagnostics worsens four-dataset leave-one-dataset-out (LODO) energy prediction: median Spearman increment -0.00689; mean absolute error 0.0011934 to 0.0012334 (3.35% worse). This tests terminal-diagnostic association, not pre-training forecasting. The direction and capacity targets do not test the intended selector questions. Appendix D preserves their defects and unperformed holdouts; neither those invalid targets nor the negative energy test establish that a useful selector is impossible.

Table 4: Raw CIFAR-10 screen: historical-input three-seed terminal means after 10,000 updates. $\mathrm{FID}_{\mathrm{tv}}$ is mean $\pm$ seed standard deviation (SD); $\mathrm{KID}_{\mathrm{tv}}$ is multiplied by 10^3. B/R denote balanced/rare. The fixed torchvision Inception evaluator is not numerically interchangeable with other FID implementations. All 24 seed rows, including precision and recall, appear in Appendix C.2. Both regimes use a balanced real reference; rare $\mathrm{FID}_{\mathrm{tv}}$/$\mathrm{KID}_{\mathrm{tv}}$ are not prior-matched.

	$\mathrm{FID}_{\mathrm{tv}}\downarrow$		$\mathrm{KID}_{\mathrm{tv}}\downarrow$		Wall (s) $\downarrow$	
Method	B	R	B	R	B	R
Optimized dense	54.25 ± 1.45	53.33 ± 2.17	29.56	28.97	6214	6226
DriftXpress	49.75 ± 2.91	52.16 ± 2.62	27.05	28.53	6137	6139
Fast	56.03 ± 2.14	64.98 ± 5.37	30.15	33.37	6135	6601
Quality	54.75 ± 0.72	56.95 ± 4.74	29.66	30.13	6138	6365

6 What Transfers to the Raw-Image Pipeline?

Only one pre-specified role transfers stably. DriftXpress retains its broad quality–wall role across the three raw seeds: its mean $\mathrm{FID}_{\mathrm{tv}}$ against the pre-specified balanced reference is 49.75 versus 54.25 for dense in the balanced regime, and 52.16 versus 53.33 in the rare regime. The rare quality signs are mixed, so this is not an all-seed quality-superiority claim. The other three pre-specified roles remain unresolved under their joint rules. In particular, fast has rare mean $\mathrm{FID}_{\mathrm{tv}}$ 64.98 and longer wall time than dense. No stable coherent role reversal passes the screen. Thus the result is neither universal transfer nor a confirmed global inversion of the feature ranking.

6.1 Field Savings and Fixed-Update Cost Are Different

We distinguish raw historical inclusive time, extension training-component time, and extension per-model charged allocation (Appendix Table 26). Write the raw inclusive time as $T = T_F + T_R$, with field fraction $\phi = T_F/T$. Under the explicit counterfactual that only field cost changes by a factor s, Amdahl's law (Amdahl, 1967) gives

$$S(s) = \frac{1}{1-\phi+\phi/s}, \qquad S_{\max} = \frac{1}{1-\phi}. \tag{6}$$

This fixed-remainder calculation bounds fixed-update cost, not time to a specified quality target.

In balanced raw dense training, $\phi = 0.029392$, yielding a field-only ceiling of $1.0303\times$. DriftXpress reduces measured field time by $9.2629\times$, but observed inclusive speedup is only $1.0125\times$. Holding the dense remainder fixed predicts $1.0269\times$; the roughly 86-second residual shows it is not constant across methods. Their different field contracts make this a *cross-method cost substitution*, not exact-operator acceleration. Balanced fast and quality likewise achieve only about $1.013\times$ overall speedup.

Raw pullback/backward bundles recomputation and a gradient scan; construction is nested in field time. Removing the 0.60–0.71-second inclusive overlap leaves the ratio at $1.013\times$ (Appendix H). The changed pipeline does not isolate a causal representation effect.

The complete bridge identifies where leverage disappears. The 24-run balanced CIFAR-10 bridge crosses two generators with ResNet-18/DINOv3 (Siméoni et al., 2025). Although DriftXpress's field speedup stays around 41–$54\times$, training speedup falls from $1.365\times$ to $1.048\times$ as dense field share falls from 27.95% to 5.42% (Table 5). Each seed preserves this leverage ordering. Generator and encoder changes also alter parameterization, masking, preprocessing, dimension, and loss scaling, not capacity or geometry alone (Appendix H).

Saved checkpoints show where the saving is useful. At 15k, the small-generator/ResNet-18 bridge improves FID and recall in all three pairs under both input paths, at 26.7% lower recorded cost. With the internal transform enabled, mean FID is 67.18/64.53 and recall 0.01527/0.02147 for Dense/DriftXpress (Table 1). At 10k, all three FID gains persist, but recall changes become $[-0.0002, -0.0030, +0.0024]$: the historical all-seed recall advantage is input-dependent. Cost ratios remain 1.354–$1.371\times$. At 5k, recall is nearly zero and one FID pair is adverse under both paths. KID

Table 5: **Stable field savings, pipeline-dependent training and quality.** Complete balanced-target CIFAR-10 bridge, three paired seeds per cell; quality scores use the historical input path. Speeds are ratios of mean dense to DriftXpress costs; field share uses dense training time. ΔF is mean paired prior-matched $\mathrm{FID}_{\mathrm{tv}}$ (DriftXpress minus dense; negative is better); wins count negative differences. R_d/R_x are mean dense/DriftXpress recall fractions. These are complete-configuration comparisons, not matched-operator approximation tests. Small has 2.90M parameters; U-Net has 38.25M. Allocation/ranges: Appendix G; selected input sensitivity: Appendix G.5.

Pipeline	Field share	Field ↑	Train ↑	ΔF ↓	Wins	R_d/R_x ↑
Small / ResNet-18	27.95%	54.45×	1.365×	−2.70	3/3	0.0239 / 0.0314
Small / DINOv3	12.29%	40.94×	1.125×	−5.23	3/3	0.0085 / 0.0099
U-Net / ResNet-18	7.15%	53.49×	1.063×	+10.88	1/3	0.0162 / 0.0116
U-Net / DINOv3	5.42%	40.53×	1.048×	+11.28	1/3	0.0000 / 0.0000

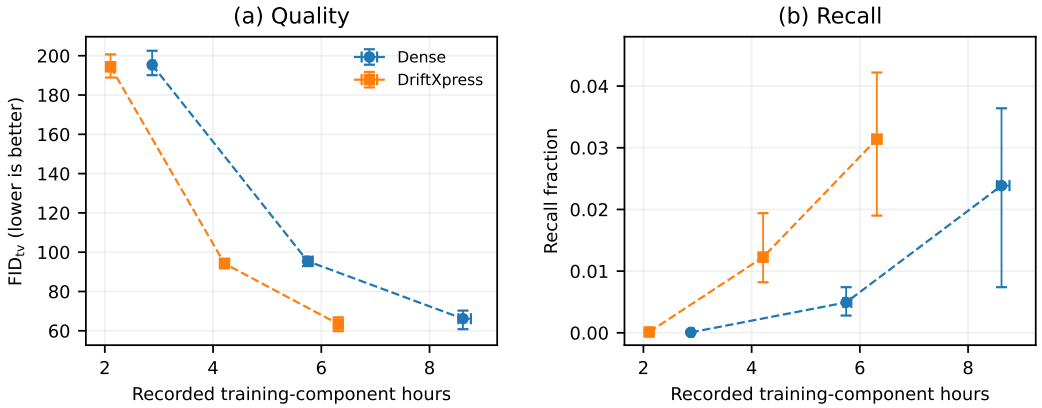


Figure 3: **Historical-input checkpoint gains.** Balanced CIFAR-10, small generator/ResNet-18. Points are three-seed means at 5k/10k/15k; bars span seed ranges, not confidence intervals. Segments connect observations, not crossings. The joint all-pair gain survives the alternate input at 15k, not 10k (Table 1). Cost excludes setup, checkpointing, and evaluation.

and precision have mixed signs; spread is not semantic coverage. Appendices G.4–G.5 retain every outcome, not a favorable-path selection or exact quality-crossing claim.

Under the historical evaluator, DriftXpress improves terminal FID in only two of six U-Net bridge pairs; 45/48 U-Net/DINOv3 secondary recalls are zero. The six-model diagnostic retains zero EMA recall under both input paths; five raw recalls are zero and one is 0.0002. Its mean EMA FID method ranking reverses with the input path, without restoring coverage. Appendix G retains all 66 outcomes and the stronger retrospective first-100-update timing baseline (0.79% median error versus 0.98% for field substitution). Neither FID rank nor execution prediction alone establishes healthy generation.

6.2 Zero Undercoverage Can Hide Substantial Mass Error

For target rare mass p_9 and estimated generated mass $\widehat{q}_9$, one-sided undercoverage is

$$U_9 = \max(p_9 - \widehat{q}_9, 0). \tag{7}$$

Whenever $\widehat{q}_9 \geq p_9$, $U_9 = 0$ regardless of the amount of overshoot. All twelve rare raw runs exhibit this: method means range from 0.0613 to 0.1085 against target 0.010989 (Figure 4). Dense and DriftXpress overshoot too; a representative safeguard alone cannot explain the shared pattern.

The first recorded excess occurs by checkpoint 1,000 in ten runs and 2,500 in the remaining two. These are interval-censored observations from 1,000-sample checkpoint diagnostics, not precise onset estimates. The classifier can also misclassify generated images. The extension's real-reference confusion matrices reveal prior-dependent class-9 precision (Appendix H); they do not calibrate

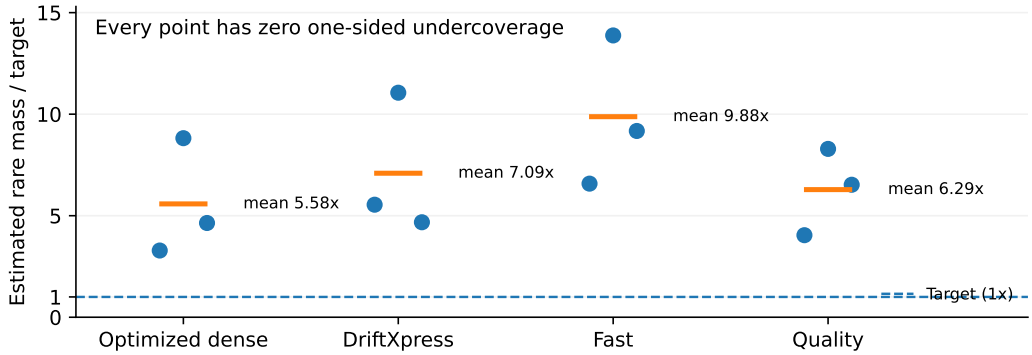


Figure 4: **Zero undercoverage is not calibrated rare mass.** Each dot is one of the twelve terminal rare raw runs (three seeds per method); short horizontal segments show means. The dashed line is the target. All dots lie above it, including dense and DriftXpress. These are classifier-estimated masses, not error-free semantic labels. Table 8 gives the means and full-vector total-variation errors.

predictions on generated images. Signed error, absolute error, rare-mass ratio, the full mass vector, and total variation expose information that undercoverage discards.

Precision and recall can move in opposite directions. In balanced raw training, fast increases mean precision by 0.0183 and decreases recall by 0.0210 relative to dense; quality changes them by $+0.0207$ and -0.0154. Both signs agree across all three seeds. $\mathrm{FID}_{\mathrm{tv}}$ and $\mathrm{KID}_{\mathrm{tv}}$ do not uniformly improve, and recall is low in this screen. These are exploratory trade-offs, not general superiority or a confirmed compression mechanism (all seeds: Appendix C.2).

7 Implications, Limitations, and Conclusion

What the audit establishes. Normalization and self deletion determine the reference operator; non-field cost limits computational leverage; one-sided undercoverage hides excess mass. The small-generator/ResNet-18 bridge pairs lower recorded cost with lower FID and higher recall at 15k under both tested input paths. The same-image input toggle reverses the 10k mean recall and high-capacity mean FID orderings without restoring high-capacity coverage. Robustness is therefore claim-specific, not a property of the evaluator as a whole.

An interpretable comparison. Check operator and update equivalence; declare cost scope; evaluate quality, diversity, and two-sided mass error (Appendix Table 21). This is not a validated selector.

Limitations. Five feature and three image seeds limit inference. Historical raw timing overlaps slightly, and its rare reference is balanced. Classifier mass is not semantic ground truth; historical predictor tests do not validate a selector. The bridge lacks information-matched exact training controls and follows a previously favorable cell, not an independent discovery. Two input paths on one Inception network do not validate all 66 models or restore a functioning high-capacity reference. Training-component curves exclude full overhead and exact quality-crossing times.

Conclusion. A cheaper field does not establish a better training method. Even in a beneficial checkpoint regime, operator fidelity, computational leverage, and mass calibration require separate evidence.

AI USE STATEMENT

Generative AI tools assisted with language and manuscript editing and with identifying relevant related work. The authors conceived the research, designed and conducted the experiments, performed and verified the analyses, and determined the interpretations and claims. All AI-assisted material, including cited related work, was reviewed and verified by the authors, who take full responsibility for the paper. The first author originated the research idea, set the scientific direction, designed and conducted the experimental program, and made the final decisions on analyses, interpretations, and claims.

REPRODUCIBILITY STATEMENT

Appendix A provides assumptions and proofs. Appendices B and F specify the completed study designs, evaluation conventions, method parameters, and decision rules; Appendices C.1–D retain detailed results and analysis limitations; Appendices G and H give all 66 extension outcomes, implementation contracts, and evaluator specifications. We intend to publicly release cleaned code and configurations after acceptance.

ETHICS STATEMENT

This study uses existing image benchmarks and generated samples and collects no new human-participant data. Artificial depletion of one semantic class is used solely as a distributional stress test and should not be interpreted as a proxy for demographic fairness. Classifier-based estimates of mass and coverage may inherit evaluator errors or biases and are therefore treated as approximate evaluation signals rather than error-free semantic labels. All datasets and pretrained models are used subject to their applicable licenses and redistribution terms.

REFERENCES

Gene M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18–20, 1967, Spring Joint Computer Conference*, pp. 483–485. ACM, 1967.

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

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021.

Benjamin Charlier, Jean Feydy, Joan Alexis Glaunès, François-David Collin, and Ghislain Durif. Kernel operations on the GPU, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6070–6079, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022.

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

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

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

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

Arthur Gretton, Li Kevin Wenliang, Alexandre Galashov, James Thornton, Valentin De Bortoli, and Arnaud Doucet. On the Wasserstein gradient flow interpretation of drifting models. *arXiv preprint arXiv:2605.05118*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 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*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, et al. MLPerf training benchmark. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 336–349, 2020.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7176–7185, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

pytorch-cifar-models contributors. PyTorch CIFAR Models. GitHub repository, n.d. URL `https://github.com/chenyaofo/pytorch-cifar-models`. CIFAR-10 ResNet-20 checkpoint variant 4118986f; accessed 22 September 2026.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 32211–32252, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

TorchVision maintainers. Inception v3 implementation. `https://github.com/pytorch/vision/blob/094e7af5/torchvision/models/inception.py`, n.d. Revision 094e7af5; Inception3 input transform and pretrained-weight factory; accessed 22 September 2026.

Erkan Turan, Nicolas Dufour, and Maks Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective. *arXiv preprint arXiv:2603.09936*, 2026.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, 2001.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

A Operator Identities and Their Limits

A.1 Exact duplicate aggregation with an aligned self mask

Let the negative source have n indexed copies, one aligned to each of the n queries. Copies need not equal their aligned query after coarsening: column j carries representative payload $z_{g(j)}$. Let $m_\ell = |\{j : g(j) = \ell\}|$. Let M_- and M_+ denote the numbers of negative and positive representative groups. Suppose positive representative w_h has q_h copies, and write

$$U^-_{i\ell} = k(x_i, z_\ell), \quad U^+_{ih} = k(x_i, w_h), \qquad E^-_\ell = \sum_i U^-_{i\ell}, \quad E^+_h = \sum_i U^+_{ih}.$$

The expanded matrix deletes only the generated entry in column i of row i. Its joint row degree and negative per-copy column degree are

$$r_i = \sum_h q_h U^+_{ih} + \sum_\ell m_\ell U^-_{i\ell} - U^-_{i,g(i)}, \tag{8}$$

$$c^-_j = E^-_{g(j)} - U^-_{j,g(j)}. \tag{9}$$

Assume all degrees used below are positive. For exponents α, β, the grouped coefficients are

$$B^+_{ih} = r_i^{-\alpha} q_h U^+_{ih} (E^+_h)^{-\beta}, \tag{10}$$

$$B^-_{i\ell} = r_i^{-\alpha} U^-_{i\ell} \left[\sum_{j:g(j)=\ell} (c^-_j)^{-\beta} - \mathbf{1}\{g(i) = \ell\}(c^-_i)^{-\beta} \right]. \tag{11}$$

The positive column degree is E^+_h for *each copy*, not $q_h E^+_h$. Negative degrees can differ within a representative group because each column loses a different aligned entry. At $\alpha = \beta = 1/2$, these are the joint geometric-mean coefficients used in Proposition 1.

Proof. For the negative block, expand the coefficient of group-constant payload z_ℓ in row i:

$$\sum_{j:g(j)=\ell,\, j\neq i} r_i^{-\alpha} U^-_{i\ell} (c^-_j)^{-\beta}.$$

Factor the common row and kernel terms, then subtract the aligned summand when its group is ℓ. This gives $B^-_{i\ell}$. In the positive block, every copy has the same kernel column, degree, and payload; summing the q_h copies gives Equation (10). Applying the identities to scalar payload one yields exactly the expanded masses; applying them to representative vectors yields exactly the expanded vector payloads. Substitution into Equation (2) therefore preserves the expanded field. □

The kernel evaluation cost is $O(n(M_+ + M_-)d)$ for M_+ positive and M_- negative representatives in dimension d, plus linear copy/group reductions and payload application. Two-pass tiled evaluation can avoid storing the entire query–representative matrix. Arithmetic order, numerical floors, and zero-degree behavior still require implementation tests. If the explicit implementation floors a degree, that *same per-copy operation* must appear before grouping; the unfloored symbolic formula is not a license to move floors across sums.

This is an identity on the specified coarsened support. It neither proves that the support approximates the original data nor eliminates selection and construction cost. Fixed memberships are an assumption. The result does not differentiate through nearest-representative assignments or discrete changes in group cardinality. Non-identical copy payloads require the corresponding degree-weighted payload sums rather than a single common payload.

A.2 Why a pre-normalization count is not a duplicate count

Consider one query and m identical positive columns, with all kernel entries equal to one and no self deletion. The expanded row degree is m, each column degree is one, and the grouped normalized coefficient is $m^{1-\alpha}$. Collapsing to one column weighted by m before normalization instead makes its column degree m, giving $m^{1-\alpha-\beta}$. For $m = 4$ and $\alpha = \beta = 1/2$, these coefficients are respectively 2 and 1. Both conventions can define legitimate algorithms, but they are not the same operator. In the joint drifting field, altered positive and negative masses also enter the cross products in Equation (2); an apparent local count convention can therefore affect the final vector field twice.

A.3 COMPATIBLE-REFERENCE DISCREPANCY

The ordinary triangle inequality gives

$$\|V_A - V_I\|_F \le \|V_A - V_C\|_F + \|V_C - V_I\|_F.$$

Interchanging V_A and V_C and combining the two inequalities proves Equation (3). No stochastic assumption is involved. A normalized relative-error version also requires a stated nonzero denominator. The bound says nothing about which native field produces better learned samples; that question requires the shared quality experiment.

A.4 DETACHED TARGETS, PULLBACKS, AND EXACT REVERSE PATHS

In Equation (5), the stop-gradient encloses the entire target $X_\theta + \eta V$, not only V. At the evaluated state, the residual is $-\eta V$, yielding the stated gradient by the chain rule. For two fields evaluated at the same X_θ, the gradient difference satisfies

$$\|\Delta \nabla_\theta L\|_2 \le |a\eta| \, \|J_\theta\|_{\mathrm{op}} \, \|\operatorname{vec}(V_A - V_C)\|_2.$$

There is no converse bound without additional restrictions: a large field discrepancy can lie in the null space of $J_\theta^\top$. Conversely, a small field discrepancy can be amplified by a large Jacobian norm. Equal cosine alone omits update magnitude and optimizer state.

Some exact systems routes exploit narrower algebra. Here $\|J_\theta\|_{\mathrm{op}}$ is the Euclidean operator norm. For a nonzero vector u, the vector–Jacobian product of $u/\|u\|_2$ with upstream v is

$$\frac{v}{\|u\|_2} - \frac{u\,(u^\top v)}{\|u\|_2^3}.$$

An affine head and compatible normalization can admit exact accumulation or streamed contraction using this identity. A generic nonlinear image generator and encoder do not become equivalent to that head by matching dimensions. For a linear projection $\Pi \in \mathbb{R}^{nd \times nd}$ on the vectorized field, with I the identity on $\mathbb{R}^{nd}$, the projected field preserves the generator gradient only when $J_\theta^\top (I - \Pi) \operatorname{vec}(V) = 0$ at that state, or when an explicitly bounded approximation is accepted. This is why rank, cancellation, and field residual diagnostics are not automatically a portable algorithm or a predictive theorem.

A.5 END-TO-END LEVERAGE AND MASS DIAGNOSTICS

Let T_F be baseline field time and let T_R contain every other charged component. If the candidate changes field time to T_F/s and changes the remainder by H, then

$$\frac{T_F + T_R}{T_F/s + T_R + H} = \frac{1}{\phi/s + 1 - \phi + H/(T_F + T_R)}.$$

Amdahl's expression follows at $H = 0$. The measured residual in the raw experiment estimates a cross-method remainder change, not the causal cost of any one code operation. Nested counters must not be summed as if they were disjoint.

For mass, $U_9 = 0$ holds for every $\widehat{q}_9 \in [p_9, 1]$. At $p_9 > 0$, the ratio $\widehat{q}_9/p_9$ ranges from 1 to $1/p_9$ on this interval and distinguishes exact matching from overshoot. For target class probabilities p_k and estimated generated probabilities $\widehat{q}_k$, total variation $\frac{1}{2}\sum_k |\widehat{q}_k - p_k|$ also detects errors outside the rare class. These are algebraic diagnostics; uncertainty and classifier calibration remain empirical questions.

B REPRODUCIBILITY AND COMPLETED PROTOCOLS

B.1 FEATURE DATA, GENERATOR, AND METRICS

Training targets use the pre-specified class-balanced or class-9-depleted training subsets. Evaluation uses held-out data, not the training target: from 500 held-out examples per class, the runner constructs a 512-example evaluation subset matched to the target class proportions, subject to integer allocation. The same held-out pool supplies fixed classifier reference and validation partitions of 250 examples

per class each. Thus the evaluation data are separate from generator training, while distribution evaluation and classifier construction are not asserted to be mutually disjoint. MNIST and Fashion-MNIST use fixed Euclidean five-nearest-neighbor voting; ties in class votes use the first class index. CIFAR-10 and SVHN use fixed linear ridge regression to one-hot class targets, with ridge coefficient 10^{-3} and an unpenalized intercept. In the retained held-out ordering, the first 250 examples of each class form the reference partition and the next 250 the validation partition. For reference design A (features with a constant column) and one-hot labels B, the ridge coefficients are $(A^\top A + 10^{-3}\,\mathrm{diag}(1,\dots,1,0))^{-1}A^\top B$; prediction is the maximum-score class. Evaluator fitting is separate from generator training.

The feature generator maps 32-dimensional Gaussian noise through three Linear–ReLU hidden blocks of width 256, followed by a linear map to feature dimension d and sample-wise ℓ_2 normalization. All linear biases start at zero. Hidden weights use Xavier-uniform initialization; the output weight is initialized from $\mathcal{N}(0, 0.02^2)$. It uses Adam (Kingma & Ba, 2015) at learning rate 10^{-3}, float32, batch 512, and temperature 0.5. Paired methods share initialization and evaluation streams, while separate random-stream offsets isolate target sampling, noise, and method-specific selection. Training comprises 1,000 updates with checkpoints at 0, 100, 500, 900, and 1,000. Checkpoint callbacks compute operator, pullback, and mechanism diagnostics; 240 runs times five checkpoints yield 1,200 records per diagnostic family, not 1,200 independent training runs.

For generated evaluation features $G = (g_i)$ and target features $Y = (y_j)$, the energy estimator is

$$\widehat{\mathcal{E}} = 2\,\overline{\|g_i - y_j\|_2} - \overline{\|g_i - g_{i'}\|_2} - \overline{\|y_j - y_{j'}\|_2}.$$

The within-sample averages here include diagonal terms: this is the preserved V-statistic implementation, not an unbiased energy estimator. For bandwidth σ, the evaluation kernel is $k_\sigma(u, v) = \exp[-\|u - v\|_2^2/(2\sigma^2)]$, with $\sigma \in \{0.1, 0.25, 0.5, 1.0\}$. These are bandwidths in normalized-feature distance units, not squared bandwidths or median-distance multipliers. This squared-distance evaluation kernel differs from the Euclidean-distance exponential used for the training field. For sample counts $n, m > 1$, the primary statistic averages

$$\widehat{\mathrm{MMD}}^2_{u,\sigma} = \frac{\sum_{i\neq i'} k_\sigma(g_i, g_{i'})}{n(n-1)} + \frac{\sum_{j\neq j'} k_\sigma(y_j, y_{j'})}{m(m-1)} - \frac{2\sum_{i,j} k_\sigma(g_i, y_j)}{nm}$$

over the four bandwidths. No square root or zero clamp is applied; finite-sample unbiased estimates can be negative. The implementation also records the diagonal-including version, but it is not the primary statistic. Manifold precision/recall and predicted class fractions complement energy/MMD rather than replacing them. All comparisons use the same frozen evaluator and sample-count policy within a cell.

B.2 Methods and what is charged

The optimized dense route evaluates the joint geometric-mean field with the aligned self mask. It is the primary dense cost reference. A less optimized canonical dense path is an engineering ablation only. DriftXpress is a clean-room projected-attraction implementation with its own separate-row native contract, exact self-masked generated repulsion, and charged landmark/cache preprocessing. It is not relabeled as an exact approximation to the joint dense operator.

The fast route uses the frozen regime-adaptive representative policy; the quality route uses the frozen probe/teacher/polish schedule. They may use target class labels for rare-support safeguards. Such labels are experimental side information and are not a claim that unsupervised rare modes would be discovered automatically. Representative selection, construction, dense teacher calls, probes, and polish are included in cost; a smaller interaction count is not itself a wall-time result. The count-expanded route groups a duplicate-coarsened support with the exact copy-specific mask formula. Sinkhorn retains its frozen split-coupling transport implementation, including its iteration count and numerical stabilization.

Appendix F records the retained method parameters and role predicates directly in the manuscript. These are historical operating points, not a newly chosen parameter sweep. Changing their representative budgets or teacher schedules would define a different transfer experiment.

B.3 RAW-IMAGE TRAINING AND INDEPENDENT EVALUATION

The low-capacity conditional residual generator uses noise dimension 128, class-embedding dimension 128, channels $(256, 256, 128, 64)$, group normalization (Wu & He, 2018), and a sigmoid RGB output at 32×32. Training uses Adam with learning rate $2 \cdot 10^{-4}$ and $(\beta_1, \beta_2) = (0.5, 0.999)$, EMA decay 0.999, drift scale 1, temperature 0.5, and float32. The frozen field encoder is ImageNet-1K ResNet-18 with 224×224 input and sample-normalized pooled features; encoder chunks contain 64 images. The generator therefore trains through a nonlinear encoder, unlike direct feature generation.

The three seeds are 42, 43, and 44. All runs use 10,000 updates and batch 512; checkpoints occur at 0, 1,000, 2,500, 5,000, 7,500, and 10,000. Training and ordinary evaluation conditioning are sampled from the empirical target class weights. Separate per-class diagnostic grids use uniform class requests; those grids do not define the $\mathrm{FID}_{\mathrm{tv}}$ or mass streams. Hence uniform grid construction is not an explanation for the observed rare overshoot.

The independent metric encoder is torchvision Inception-v3 (Szegedy et al., 2016) with its specified ImageNet weights, not the field ResNet-18. Terminal $\mathrm{FID}_{\mathrm{tv}}$ uses 10,000 generated samples. $\mathrm{KID}_{\mathrm{tv}}$ and manifold precision/recall use 5,000 generated samples, with manifold neighbor count $k = 3$. The *real* reference in the preserved raw evaluation is the same 5,000-image balanced held-out set (500 per class) for both target regimes. Unlike the feature energy/MMD reference, it is *not* reweighted to the rare training prior. Therefore rare-regime $\mathrm{FID}_{\mathrm{tv}}$/$\mathrm{KID}_{\mathrm{tv}}$ describe similarity to that balanced benchmark reference; they do not isolate fidelity to the rare target distribution. This evaluation-definition limitation is retained rather than retroactively changing the reported numbers. The extension's prior-matched evaluation is specified separately in Appendix G; it does not replace these historical raw-screen scores.

Class mass is estimated by an independent CIFAR-10 ResNet-20 classifier. Checkpoint mass diagnostics use 1,000 generated samples, so apparent early onset has sampling uncertainty and interval censoring. Classifier error on generated images is not identified by real-data validation accuracy alone. Reported mass is consequently an estimate; neither its overshoot nor a zero one-sided metric proves ground-truth semantic class proportions. Table 8 reports the rare-class fraction, its target ratio, one-sided undercoverage, and full-vector total-variation error.

B.4 COMPUTE, TIMING, AND COMPLETION ACCOUNTING

The completed local experiments use an NVIDIA RTX 5090 Laptop GPU with 24,463 MiB reported memory. The preserved raw environment records PyTorch 2.11.0+cu128; the displayed driver CUDA capability is distinct from the framework build's CUDA version. These local measurements do not establish A100 throughput or portability; the separate A100 extension is reported in Appendix G.

The raw screen retained its 24-row design while a pre-outcome compute-only amendment raised its device-hour ceiling from 32.05 to 52 and its wall ceiling to 60 hours. All 24 rows passed technical/non-collapse checks; no scientific failure or replacement was removed from this matrix. The separate 240-row feature matrix also completed without replacements. Technical validation tests are not counted as additional scientific training runs.

Inclusive times charge the preserved preprocessing, training, diagnostics, and evaluation policy. Component boundaries are stated explicitly: construction is nested within field time; the raw backward counter bundles additional forward work; and the frozen raw denominator contains a small checkpoint/evaluation overlap. The appendix retains the historical numerical comparisons and explicitly identifies these timing conventions. The historical inclusive denominator and the sum of measured components are therefore not treated as identical quantities.

C COMPLETE CELL AND SEED RESULTS

C.1 FEATURE BENCHMARK

Table 6 reports every dataset–regime–method cell. Each contains five paired seeds. The energy confidence intervals and reported standard deviations describe within-cell seed variation; the aggregate main table is an equal-weight descriptive average over the eight cells. Neither averaging metric values across representations nor averaging per-cell means creates an independent cross-dataset confidence

interval. Native field differences remain relevant to operator error, even though learned-distribution metrics are shared within each cell.

Table 6: All 48 feature cells (five seeds each). B/R denote balanced/rare targets; SD denotes standard deviation and CI a confidence interval. Energy, MMD, and the energy 95% bootstrap interval are scaled by 10^3; MMD denotes averaged unbiased MMD-squared. Wall time is in seconds. Bracketed intervals come from the per-cell analysis, not from pooling datasets.

Target	Method	Energy ± SD	Energy 95% CI	MMD ± SD	Wall ± SD
MNIST					
B	Optimized dense	8.17 ± 1.51	[7.09, 9.51]	1.71 ± 0.42	4.13 ± 0.23
	DriftXpress	6.97 ± 0.30	[6.74, 7.20]	1.37 ± 0.09	3.61 ± 0.12
	Fast	8.82 ± 1.18	[7.91, 9.73]	1.99 ± 0.29	3.22 ± 0.17
	Quality	8.20 ± 1.23	[7.29, 9.18]	1.80 ± 0.32	3.32 ± 0.12
	Count-expanded	8.39 ± 1.41	[7.48, 9.65]	1.81 ± 0.40	4.24 ± 0.25
	Sinkhorn	8.28 ± 1.48	[7.14, 9.50]	1.75 ± 0.41	11.51 ± 0.16
R	Optimized dense	6.83 ± 0.57	[6.37, 7.29]	1.32 ± 0.12	4.39 ± 0.23
	DriftXpress	6.60 ± 0.47	[6.23, 6.96]	1.25 ± 0.14	3.94 ± 0.31
	Fast	34.06 ± 3.42	[31.17, 36.82]	9.52 ± 0.94	3.81 ± 0.16
	Quality	7.79 ± 1.24	[6.94, 8.89]	1.64 ± 0.36	3.65 ± 0.19
	Count-expanded	7.61 ± 1.35	[6.88, 8.84]	1.55 ± 0.34	4.64 ± 0.20
	Sinkhorn	6.76 ± 0.41	[6.42, 7.04]	1.31 ± 0.08	12.65 ± 0.48
Fashion-MNIST					
B	Optimized dense	6.28 ± 1.07	[5.57, 7.16]	2.11 ± 0.36	4.28 ± 0.14
	DriftXpress	5.94 ± 0.73	[5.36, 6.51]	1.76 ± 0.22	3.79 ± 0.28
	Fast	7.29 ± 1.31	[6.26, 8.27]	2.59 ± 0.52	3.20 ± 0.08
	Quality	6.72 ± 1.09	[6.09, 7.70]	2.28 ± 0.25	3.39 ± 0.10
	Count-expanded	6.46 ± 0.75	[5.90, 7.14]	2.07 ± 0.24	4.45 ± 0.17
	Sinkhorn	6.04 ± 0.84	[5.50, 6.77]	2.13 ± 0.24	12.46 ± 0.20
R	Optimized dense	7.26 ± 2.86	[5.52, 9.79]	2.31 ± 1.10	4.44 ± 0.16
	DriftXpress	6.03 ± 0.62	[5.48, 6.51]	1.78 ± 0.19	4.06 ± 0.36
	Fast	30.38 ± 7.06	[25.12, 35.84]	13.11 ± 2.91	3.99 ± 0.09
	Quality	6.99 ± 1.04	[6.21, 7.79]	2.47 ± 0.49	3.77 ± 0.19
	Count-expanded	7.42 ± 2.39	[5.96, 9.53]	2.30 ± 0.87	4.74 ± 0.46
	Sinkhorn	7.71 ± 2.71	[6.21, 10.11]	2.58 ± 1.14	12.92 ± 0.74
CIFAR-10					
B	Optimized dense	6.13 ± 0.27	[5.94, 6.33]	1.79 ± 0.15	3.96 ± 0.18
	DriftXpress	5.58 ± 0.43	[5.27, 5.95]	1.52 ± 0.21	3.66 ± 0.32
	Fast	6.81 ± 0.48	[6.44, 7.19]	2.15 ± 0.22	3.08 ± 0.18
	Quality	6.52 ± 0.33	[6.22, 6.73]	2.01 ± 0.15	3.24 ± 0.17
	Count-expanded	6.43 ± 0.40	[6.16, 6.78]	1.96 ± 0.20	4.27 ± 0.40
	Sinkhorn	6.47 ± 0.35	[6.22, 6.76]	2.07 ± 0.19	12.10 ± 0.93
R	Optimized dense	6.24 ± 0.59	[5.81, 6.76]	1.83 ± 0.29	3.75 ± 0.02
	DriftXpress	5.58 ± 0.24	[5.41, 5.78]	1.49 ± 0.13	3.47 ± 0.08
	Fast	23.25 ± 1.59	[22.08, 24.42]	10.08 ± 0.71	3.45 ± 0.12
	Quality	7.06 ± 0.80	[6.43, 7.74]	2.25 ± 0.33	3.29 ± 0.07
	Count-expanded	6.56 ± 0.64	[6.15, 7.11]	2.00 ± 0.26	4.05 ± 0.09
	Sinkhorn	6.58 ± 0.57	[6.17, 7.08]	2.09 ± 0.27	11.22 ± 0.07
SVHN					
B	Optimized dense	17.20 ± 2.11	[15.67, 19.04]	8.92 ± 1.15	3.74 ± 0.04
	DriftXpress	16.48 ± 0.96	[15.84, 17.31]	8.50 ± 0.59	3.46 ± 0.13
	Fast	19.28 ± 2.95	[17.02, 21.54]	10.09 ± 1.40	2.97 ± 0.04
	Quality	18.19 ± 1.39	[17.12, 19.24]	9.61 ± 0.72	3.07 ± 0.03
	Count-expanded	17.86 ± 2.73	[15.96, 20.22]	9.40 ± 1.51	3.95 ± 0.03
	Sinkhorn	17.97 ± 1.83	[16.62, 19.51]	9.57 ± 1.04	11.17 ± 0.09
R	Optimized dense	15.22 ± 1.84	[13.98, 16.79]	7.85 ± 1.09	3.76 ± 0.06
	DriftXpress	15.65 ± 1.40	[14.72, 16.80]	7.92 ± 0.70	3.45 ± 0.07
	Fast	28.95 ± 7.20	[23.41, 34.48]	15.13 ± 3.10	3.46 ± 0.06
	Quality	15.90 ± 1.30	[14.98, 17.00]	8.42 ± 0.86	3.28 ± 0.07

Target	Method	Energy $\pm$ SD	Energy 95% CI	MMD $\pm$ SD	Wall $\pm$ SD
	Count-expanded	16.26 ± 2.80	[14.41, 18.77]	8.52 ± 1.61	3.94 ± 0.01
	Sinkhorn	15.87 ± 2.01	[14.69, 17.69]	8.45 ± 1.17	11.20 ± 0.16

C.2 RAW SCREEN: EVERY SEED

The following table preserves all 24 terminal rows. $\mathrm{FID}_{\mathrm{tv}}$ and $\mathrm{KID}_{\mathrm{tv}}$ use the balanced real reference described in Appendix B; in particular, rare rows must not be interpreted as prior-matched fidelity. Precision and recall use the same independent Inception feature space. Three seeds are enough to expose consistency or mixed signs in this screen, not to establish a population-level universal ordering.

Table 7: Every raw-image run, with historical-input quality metrics. $\mathrm{KID}_{\mathrm{tv}}$ is scaled by 10^3; precision and recall use the same frozen manifold evaluator. These rows are not independent datasets or independent checkpoints.

Method	Seed	$\mathrm{FID}_{\mathrm{tv}}$	$\mathrm{KID}_{\mathrm{tv}}$	Precision	Recall	Wall (s)
Balanced						
Optimized dense	42	55.321	29.649	0.6896	0.0434	6189.19
Optimized dense	43	54.837	28.962	0.6370	0.0546	6229.44
Optimized dense	44	52.597	30.066	0.6472	0.0538	6222.44
DriftXpress	42	52.559	28.130	0.6582	0.0356	6128.52
DriftXpress	43	46.756	23.199	0.6462	0.0712	6145.07
DriftXpress	44	49.923	29.830	0.6352	0.0696	6136.83
Fast	42	55.941	29.739	0.7060	0.0216	6131.26
Fast	43	53.930	28.127	0.6716	0.0246	6148.47
Fast	44	58.214	32.574	0.6512	0.0426	6125.84
Quality	42	54.328	28.048	0.7008	0.0256	6130.17
Quality	43	55.576	30.549	0.6854	0.0328	6156.46
Quality	44	54.337	30.392	0.6498	0.0472	6127.09
Rare						
Optimized dense	42	55.821	29.694	0.7098	0.0240	6224.61
Optimized dense	43	52.332	27.245	0.7128	0.0410	6222.11
Optimized dense	44	51.849	29.984	0.6606	0.0642	6230.51
DriftXpress	42	51.187	26.795	0.7150	0.0456	6132.90
DriftXpress	43	55.123	29.610	0.6942	0.0368	6126.42
DriftXpress	44	50.171	29.172	0.6546	0.0786	6157.51
Fast	42	68.136	33.794	0.6312	0.0296	6582.34
Fast	43	68.034	37.201	0.6192	0.0106	6595.06
Fast	44	58.777	29.101	0.6192	0.0200	6625.21
Quality	42	58.833	31.279	0.6754	0.0324	6359.39
Quality	43	60.466	32.557	0.6536	0.0202	6360.13
Quality	44	51.563	26.565	0.6674	0.0538	6374.56

Table 8: One-sided undercoverage hides rare-class overshoot in all 12 rare raw runs. Target class-9 mass is $100/9100 = 0.010989$. Entries are three-seed means; $\widehat{q}_9$ is classifier-estimated generated mass, not error-free ground truth. TV refers to the full estimated class-mass vector.

Method	$\widehat{q}_9$	$\widehat{q}_9/p_9$	Undercoverage	TV
Optimized dense	0.0613	5.58	0.0000	0.268
DriftXpress	0.0779	7.09	0.0000	0.258
Fast	0.1085	9.88	0.0000	0.306
Quality	0.0691	6.29	0.0000	0.286

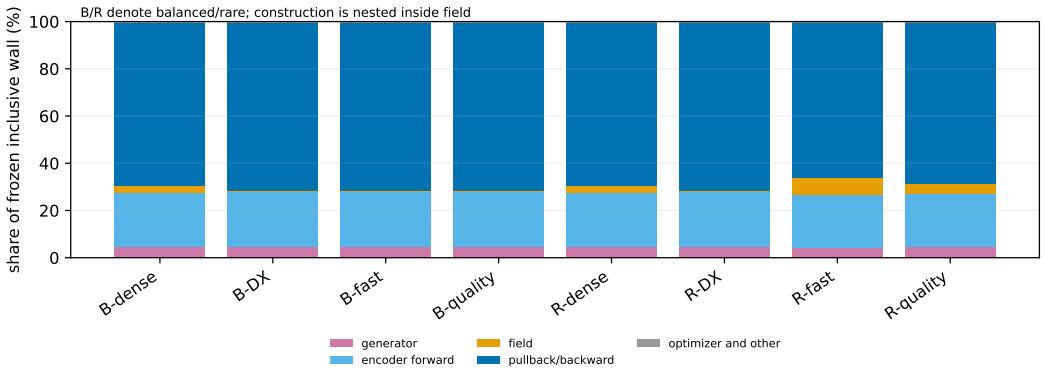


Figure 5: Preserved raw component ledger. Read construction as nested in field time, and the backward label as a bundle rather than pure encoder backward. The small historical checkpoint/evaluation overlap prevents interpreting all displayed counters as a strictly disjoint partition.

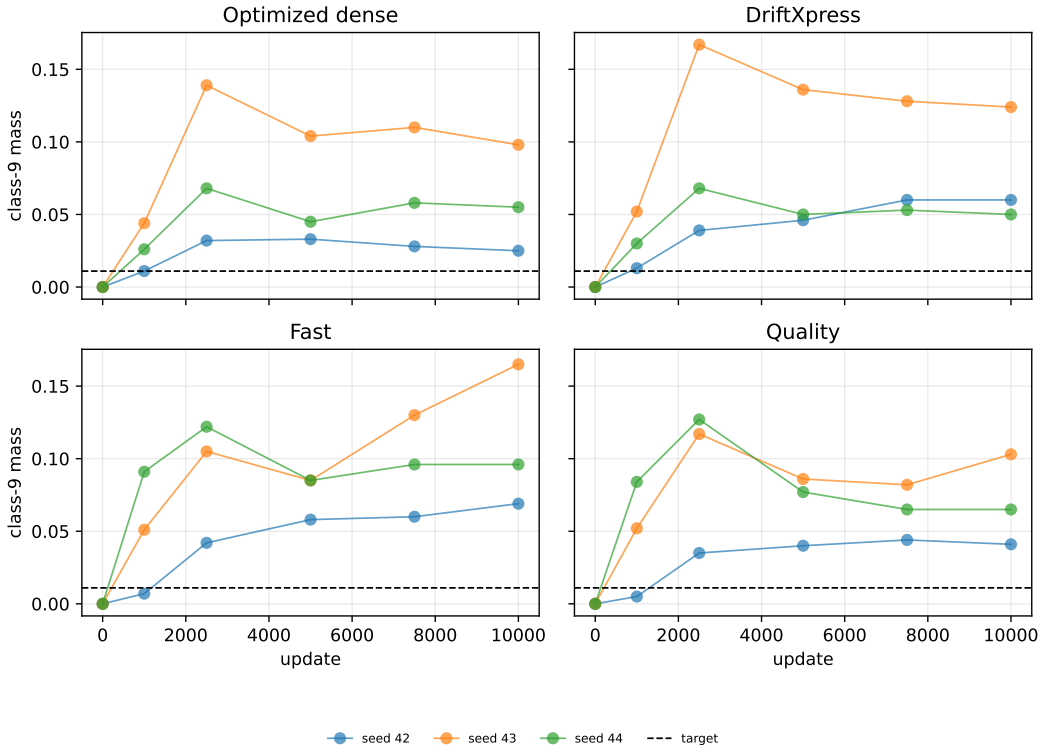


Figure 6: Estimated rare-class mass trajectories for all rare raw runs, relative to the target mass. Checkpoint estimates use 1,000 generated samples and a fixed classifier. Observed threshold crossings are interval-censored; the curves do not identify semantic class mass without classifier error. Dense and DriftXpress also overshoot.

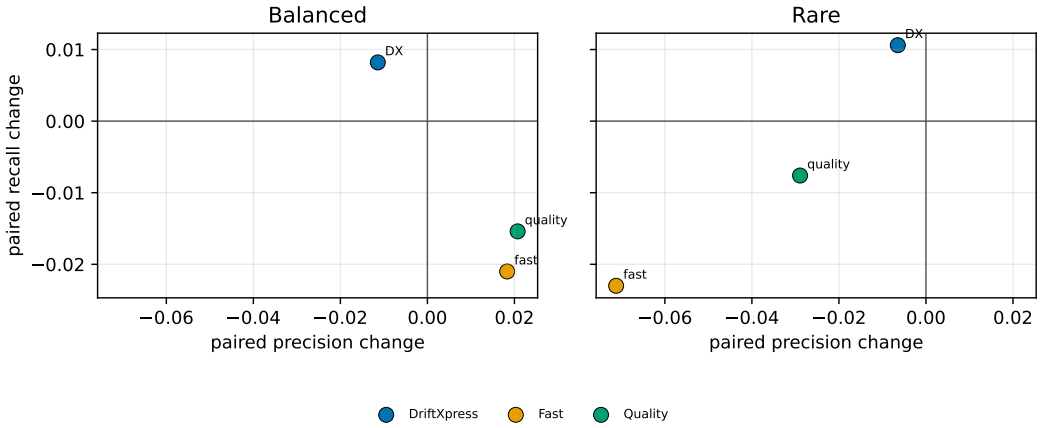


Figure 7: Raw precision–recall evidence, including paired changes relative to dense. Each point shows the mean method-minus-dense change in precision and recall across three paired seeds within one target regime. Balanced fast and quality have increased precision and decreased recall in all three seeds. This descriptive consistency is not a confirmed general transfer mechanism.

C.3 Exact systems endpoints

Table 9: Exact operator/reverse-path systems endpoints at balanced support $N = 8192$. All speed ratios use optimized dense, not the slower canonical implementation. The stream/cache variants are exact within their affine or normalized-affine architecture contract, not generic nonlinear-encoder speedups.

Implementation	Time (ms)	MiB	Speed ratio	Memory reduction
Canonical dense	59.794	3778.6	0.610	0.224
Optimized dense	36.465	845.7	1.000	1.000
Full cache	34.238	709.2	1.065	1.192
Streamed	182.240	154.6	0.200	5.470

At support 8,192, comparing full cache with optimized dense gives $36.465/34.238 = 1.065$ speedup and $845.7/709.2 = 1.192$ memory reduction. Streaming gives $845.7/154.6 = 5.470$ memory reduction but costs $182.240/36.465 = 4.998$ times as much time. These are measured engineering endpoints under the eligible reverse-path contract, not universal end-to-end generation speedups. Using canonical dense instead of optimized dense would substantially inflate the apparent improvement. Figure 8 shows these endpoints and the separate unique-state grouping measurements.

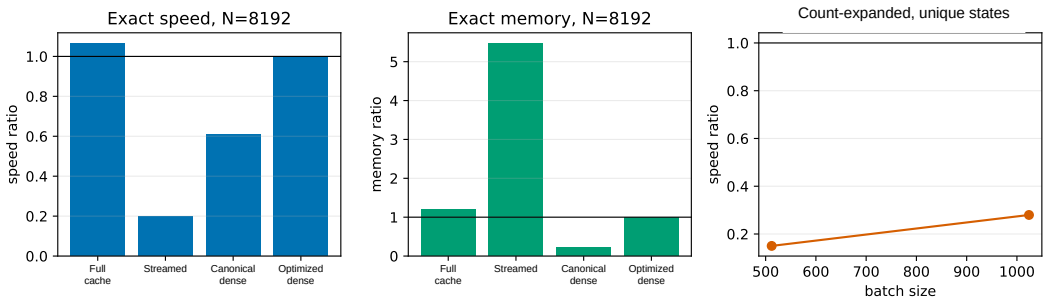


Figure 8: Exact-system measurements. Left and middle: speed and memory-reduction ratios at support size $N = 8{,}192$, with optimized dense as the reference. Canonical dense exposes implementation headroom, not an alternative baseline for the speedup claim. Right: count-expanded grouping on unique-state inputs at two batch sizes; a speed ratio below one denotes slowdown. Algebraic eligibility and the measured workload limit transfer to other architectures.

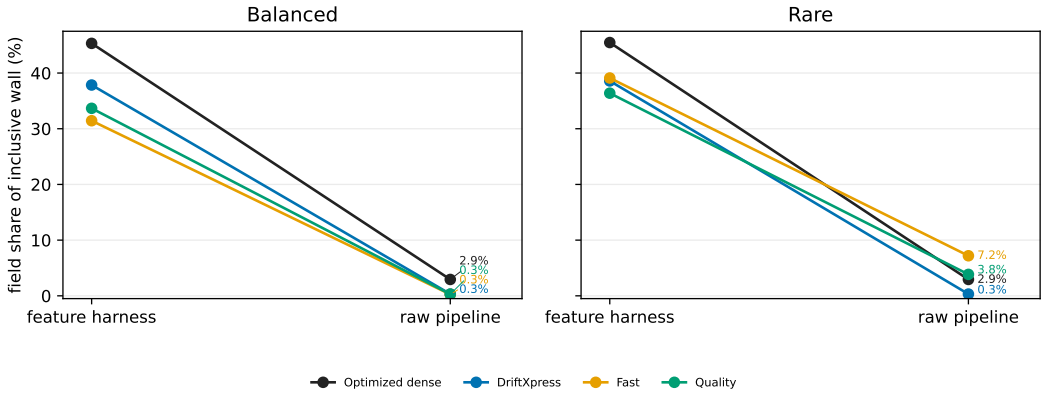


Figure 9: Historical pipeline profiles: feature points average four datasets and raw points summarize three CIFAR-10 seeds. These are not isolated representation interventions. Field fractions depend jointly on generator, encoder, support, precision, and implementation.

D Predictor Corrections, Decision Rules, and Claim Scope

D.1 Evidence status and frozen decision rules

The usable evidence is reference-dependent field discrepancy, the negative terminal-energy prediction test, and quality–cost comparisons under declared rules. The direction/capacity targets do not test the intended selector questions; the rare-deficit margin does not certify calibration. Sections below separate these conclusions from the historical analysis that produced them.

The original joint thesis required H1, H3, and H5. Its log labels H1/H5 as passes and H2/H3 as failures; these are historical statuses, not repaired inference. H1's pooled p value does not establish five-family significance, H3's targets are invalid for their intended questions, and H4's occupancy/architecture analysis gives no universal benefit. The joint thesis remains unsupported.

"Confirmatory" identifies the frozen five-seed analysis, not wholly unobserved seeds. The separate 24-row raw screen uses fixed role rules; it does not confirm every feature hypothesis. Pre-specified means fixed within the study protocol, not externally registered. The generator–encoder audit (Appendix G) does not change these decisions.

Paired seed summaries should be distinguished from cell selection. The historical joint quality non-inferiority rule requires energy and MMD within 5% relative degradation bounds, no coverage decrease, and at most 0.02 additional rare undercoverage. The last condition is nonbinding: the feature metric is $U_9 = \max(p_9^{\text{eval}} - \widehat{q}_9, 0)$, where largest-remainder allocation gives $p_9^{\text{eval}} = 6/512 =$

0.01171875. Thus $U_9^{\text{candidate}} - U_9^{\text{dense}} \leq 6/512 < 0.02$. Even complete predicted rare-class loss can pass this component. The separate coverage condition may still fail; the joint pass counts are not a calibration certificate. The component decision bounds are supplied below rather than reconstructed from rounded means. A method can lie on a quality–time–memory Pareto frontier yet fail this joint quality gate. Conversely, a non-inferiority failure alone does not establish statistically significant harm.

D.2 Compatible-reference statistic and inferential unit

For each recorded state and reference R, the relative error is $e_R = \|V_A - V_R\|_F / \max(\|V_R\|_F, 10^{-12})$. The reference norm is recomputed for each comparator. Thus $e_I - e_C$ is not a fractional reduction with a common denominator, and Equation (3) bounds unnormalized discrepancies, not this normalized difference directly. The family-level medians are:

Family	Median $e_I - e_C$
MNIST	0.500880
Fashion-MNIST	0.512246
CIFAR-10	0.499905
SVHN	0.509122
Controlled two-dimensional	0.899191

Their median is 0.509122. The ancillary two-dimensional protocol uses identity-coordinate four-Gaussian/equal-mass and imbalanced-Gaussian targets, a 2,000-update horizon, checkpoints $\{0, 200, 1000, 1800, 2000\}$, and 4,096 held-out samples. It is an existing synthetic control, not an additional image dataset or fresh five-seed expansion.

The historical H1 rule required a favorable median in at least four of five families, a family-corrected paired test at 0.05, and a positive lower bootstrap bound. The original log labels it a pass, with interval $[0.499905, 0.899191]$ and reported paired $p = 9.9999 \cdot 10^{-6}$. This historical label and pooled p value are *not current inferential support* for five independent family outcomes. Repeated states within a trajectory are not independent runs; five positive family signs alone give a two-sided exact sign-test value $2/2^5 = 0.0625$. The main claim is therefore the observed reference dependence and the algebraic distinction, not discovery-independent cross-family significance. The descriptive family medians, rather than the pooled historical p value, support the main-text comparison.

Non-inferiority decision components. Table 10 reports the upper endpoints of the recorded two-sided 95% percentile intervals, not separately computed one-sided 95% upper bounds. For paired candidate and dense values c_s, d_s at seeds $s \in \{11, 23, 37, 53, 71\}$, energy and MMD degradation is $(c_s - d_s)/\max(|d_s|, 10^{-12})$; coverage loss is $d_s - c_s$; rare-undercoverage degradation is $c_s - d_s$. Coverage is the fraction of target classes receiving at least one predicted sample. The point estimate averages the five paired degradations, not a ratio of endpoint means.

For each metric and cell, draw 10,000 bootstrap replicates of five paired degradations with replacement and average each replicate. Use NumPy's default linear-interpolated quantiles at 0.025 and 0.975; compare the *97.5th-percentile* endpoint inclusively with the margin. Restart NumPy's `default_rng` at seed $1204 + 20j + k$, where zero-based j orders DriftXpress, Fast, Quality, count-expanded, Sinkhorn, and k orders energy, MMD, coverage, rare undercoverage. There is no dataset/regime seed offset. All historical metric inputs are finite, and all dense MMD references are positive and above the denominator floor. Margins are 0.05, 0.05, 0, and 0.02, respectively; a cell passes only when every applicable upper endpoint is at or below its margin. Equality passes without a numerical tolerance, including 39 exact zero-margin coverage ties. Re-evaluating the retained routine on all 240 run records reproduces all 40 joint decisions and their component bounds without change. This verifies the historical calculation, not an additional experiment. The intervals are neither Holm-adjusted superiority tests nor simultaneous cross-cell confidence bands; the separate paired-test correction does not make the nonbinding rare margin a calibration test.

Table 10: All 40 feature candidate–cell non-inferiority decisions versus optimized dense. Entries are the 97.5th-percentile upper endpoints of two-sided 95% paired-bootstrap degradation intervals; B/R denote balanced/rare. E and M are relative energy/MMD degradation, C is coverage loss, and U is additional rare undercoverage. Margins are 0.05, 0.05, 0, and 0.02, respectively. "–" is not applicable. These are not simultaneous confidence bounds.

Dataset	Target	Method	E	M	C	U	Pass
MNIST	B	DriftXpress	-0.014401	-0.031631	+0.000000	–	Yes
MNIST	B	Fast	+0.220083	+0.350096	+0.000000	–	No
MNIST	B	Quality	+0.097603	+0.182448	+0.000000	–	No
MNIST	B	Count-expanded	+0.066131	+0.110388	+0.000000	–	No
MNIST	B	Sinkhorn	+0.062495	+0.105001	+0.000000	–	No
MNIST	R	DriftXpress	+0.036938	+0.026474	+0.000000	+0.000000	Yes
MNIST	R	Fast	+4.324718	+6.677976	+0.000000	+0.000000	No
MNIST	R	Quality	+0.302293	+0.510224	+0.000000	+0.000000	No
MNIST	R	Count-expanded	+0.216524	+0.319041	+0.000000	+0.000000	No
MNIST	R	Sinkhorn	+0.008753	+0.019471	+0.000000	+0.000000	Yes
Fashion-MNIST	B	DriftXpress	+0.097642	-0.004638	+0.000000	–	No
Fashion-MNIST	B	Fast	+0.323795	+0.446589	+0.000000	–	No
Fashion-MNIST	B	Quality	+0.152095	+0.214798	+0.000000	–	No
Fashion-MNIST	B	Count-expanded	+0.123291	+0.076277	+0.000000	–	No
Fashion-MNIST	B	Sinkhorn	+0.036165	+0.122415	+0.000000	–	No
Fashion-MNIST	R	DriftXpress	+0.120616	+0.109341	+0.000000	+0.000000	No
Fashion-MNIST	R	Fast	+4.812847	+7.245342	-0.100000	+0.000000	No
Fashion-MNIST	R	Quality	+0.200658	+0.419462	+0.000000	+0.000000	No
Fashion-MNIST	R	Count-expanded	+0.106455	+0.099387	+0.000000	+0.000000	No
Fashion-MNIST	R	Sinkhorn	+0.167439	+0.213993	+0.000000	+0.000000	No
CIFAR-10	B	DriftXpress	-0.031570	-0.070166	+0.000000	–	Yes
CIFAR-10	B	Fast	+0.147095	+0.265454	+0.000000	–	No
CIFAR-10	B	Quality	+0.112214	+0.191687	+0.000000	–	No
CIFAR-10	B	Count-expanded	+0.080642	+0.137489	+0.000000	–	No
CIFAR-10	B	Sinkhorn	+0.066004	+0.160271	+0.000000	–	No
CIFAR-10	R	DriftXpress	-0.055937	-0.117477	+0.000000	+0.001172	Yes
CIFAR-10	R	Fast	+3.107065	+5.420245	+0.000000	+0.000000	No
CIFAR-10	R	Quality	+0.178003	+0.313698	+0.000000	+0.002344	No
CIFAR-10	R	Count-expanded	+0.119986	+0.182838	+0.000000	+0.000000	No
CIFAR-10	R	Sinkhorn	+0.063321	+0.169324	+0.000000	+0.000000	No
SVHN	B	DriftXpress	+0.037184	+0.025967	+0.000000	–	Yes
SVHN	B	Fast	+0.304161	+0.284275	+0.000000	–	No
SVHN	B	Quality	+0.135605	+0.149255	+0.000000	–	No
SVHN	B	Count-expanded	+0.074966	+0.093551	+0.000000	–	No
SVHN	B	Sinkhorn	+0.070473	+0.098885	+0.000000	–	No
SVHN	R	DriftXpress	+0.107624	+0.094872	+0.000000	+0.000000	No
SVHN	R	Fast	+1.274999	+1.258074	+0.000000	+0.000000	No
SVHN	R	Quality	+0.190014	+0.249016	+0.000000	+0.000000	No
SVHN	R	Count-expanded	+0.108815	+0.124461	+0.000000	+0.000000	No
SVHN	R	Sinkhorn	+0.060742	+0.096957	+0.000000	+0.000000	No

D.3 H2: CORRELATION INCREMENTS ARE NOT MAE IMPROVEMENTS

Table 11: Corrected H2 four-dataset LODO evidence. $\Delta\rho$ is pullback minus field-only Spearman. MAE is held-out mean absolute error in energy-difference units; smaller is better. Each row summarizes one held-out-dataset fold.

Held-out dataset	$\Delta\rho$	Field MAE	Pullback MAE
MNIST	-0.033514	0.00139383	0.00138506
Fashion-MNIST	-0.006550	0.00118525	0.00118089
CIFAR-10	$+0.055871$	0.00074464	0.00092174
SVHN	-0.007238	0.00144987	0.00144585

These energy rows occur once each in the corrected unique-record analysis; the duplicated target blocks do not enter this table. The median $\Delta\rho$ is -0.00689430000034591. Averaging MAE across the same four folds gives 0.0011933958918770463 for field-only and 0.0012333845603499393 for pullback. Defining relative improvement as $(\mathrm{MAE}_{\mathrm{field}} - \mathrm{MAE}_{\mathrm{pullback}})/\mathrm{MAE}_{\mathrm{field}}$ gives -0.033508300761783766. The negative sign means worse average error. A former prose label incorrectly called the Spearman increment an MAE improvement; the correction changes the label and supporting arithmetic, not the frozen H2 failure.

For method m, dataset d, regime r, and seed s, the energy target is the signed terminal difference

$$y_{d,r,m,s} = E_{d,r,m,s} - E_{d,r,\mathrm{dense},s}, \qquad \mathrm{MAE} = |\mathcal{T}|^{-1} \sum_{i \in \mathcal{T}} |\widehat{y}_i - y_i|,$$

where E is the energy-distance statistic in Appendix B and $\mathcal{T}$ is the held-out dataset fold. This is an absolute difference, not a relative degradation or a ratio. Predictions and targets remain in these original units: no target standardization, output clipping, or additional fitted calibration map is applied. The historical export's "calibrated MAE" label refers to this ordinary held-out MAE. The four-fold means above and the H2 decision are unchanged.

D.4 H3: SOURCE-AUDITED TARGET AND DUPLICATION DEFECTS

A review of the implemented example construction reveals three problems beyond the earlier correction of statistical labels.

Direction is already an input transformation. The direction target is $y_{\mathrm{dir}} = 1 - c$, where c is the same compatible-field cosine supplied to every predictor family. An intercept and that input can reproduce the target without predicting an unseen phenomenon. The field-only exported MAE is $2.0 \cdot 10^{-15}$ to $8.8 \cdot 10^{-15}$ across dataset folds. The retrospective CPU reconstruction recovers the target identity exactly across all 240 examples and direction-prediction error below $3 \cdot 10^{-12}$. Floating-point splitting of target ties under the historical rank rule explains why rank statistics need not equal one despite negligible absolute error. This is not an independent direction-prediction task.

Selected fraction is not required capacity. The purported required-rank or budget target is the realized selected-edge fraction. It is not obtained by finding the minimum support that attains a fixed approximation or quality criterion. The realized selected fraction remains a descriptive diagnostic, but it does not identify required capacity.

Some target blocks were emitted twice. The dataset-holdout loop repeats the non-inferiority, selected-fraction, direction, and crossover targets. The exported table has 176 records but only 112 distinct dataset–target–family keys, with 64 exact duplicate records. These are not independent repetitions. The identical direction increments reported for four datasets are no longer presented as four pieces of mechanistic evidence. A retrospective CPU reconstruction from all 240 examples reproduces the 176 historical point-score and 176 bootstrap records with zero numerical error. Although each duplicate point score is identical, all 64 duplicated bootstrap pairs differ: the repeated fits consume fresh draws from a shared random-number stream. Removing the repetitions before resampling changes 80 of 112 descriptive intervals relative to their first historical occurrence; 32 remain unchanged. This reanalysis corrects the interpretation; the historical records remain unchanged and the reanalysis is not new confirmation.

The originally reported H3 decision remains unsuccessful; these issues cannot establish that no useful predictor exists. H3 capacity/direction method holdouts were not evaluated: that export contains only energy, MMD, and rare-undercoverage targets. Its historical zero reversal count came from an empty filter, not measured stability. On the separate energy target, all four predictor families have negative Spearman correlations for the held-out fast method; this is not a set of H3 direction tests.

Predictor specification and statistical unit. The field family contains relative field error, cosine, and norm ratio. The pullback family adds generator-gradient cosine, relative error and norm ratio, plus Adam-update cosine and norm ratio. The mechanism family adds effective rank, principal-component tail energy, nearest-neighbor intrinsic dimension, support overlap, distance concentration, cancellation, occupancy, and hidden/output width ratio to the field family; the combined family contains both additions. Inputs are terminal checkpoint diagnostics paired by dataset, regime, method, and seed, not five independent observations from each trajectory. Ridge choices are $\{0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$;

outer folds hold out a dataset and inner folds hold out a method. For each candidate ridge value, pool absolute validation errors from all inner method-holdout folds and minimize their mean; exact ties retain the first (smallest) grid value. Refit on all three outer-training datasets before predicting the held-out dataset. Each fit standardizes input columns with its own training means and population standard deviations; a standard deviation below 10^{-12} is replaced by one. With the intercept in the first design column, the coefficients are $(Z^\top Z + \lambda \operatorname{diag}(0, 1, \ldots, 1))^+ Z^\top y$, using the Moore–Penrose pseudoinverse. Nonfinite inputs are converted before standardization by NumPy's default `nan_to_num`: NaNs become zero and infinities become signed finite float64 extrema. Targets are not transformed. H2 concerns contemporaneous terminal-diagnostic association, not a pre-training decision rule. The historical bootstrap resamples example rows, so its intervals are not claimed to be grouped, discovery-independent validation. H2 point fits and all 72 method-holdout rows are reproduced; the separately seeded historical decision intervals remain unchanged by the duplicate-loop correction.

Operational mechanism-feature definitions. Let $P \in \mathbb{R}^{128\times d}$ be the fixed, class-mass-matched diagnostic subset of the training target, and $X \in \mathbb{R}^{128\times d}$ the model's evaluation-mode output on fixed Gaussian noise (seed $s + 12001$). These are feature-space diagnostics, not the held-out metric reference. Source spectra use the column-centered concatenation $[P; X]$, with singular values computed in float32. Write e_j for their squared values, $S = \max(\sum_j e_j, 10^{-12})$, and $p_j = e_j/S$. The recorded settings and edge cases are:

Diagnostic	Implemented convention
Effective rank	$\exp[-\sum_j p_j \log \max(p_j, 10^{-30})]$. A zero centered spectrum yields one under this convention.
Principal-component tail	$\sum_{j>32} e_j/S$, truncating the cutoff to the number of singular values when smaller.
Intrinsic dimension	From P only: let $r_{i1}, \ldots, r_{i,10}$ be the ten closest Euclidean distances, excluding only the self diagonal, each floored at 10^{-12}. Report $9/[128^{-1} \sum_i \sum_{j=1}^{9} \log(r_{i,10}/r_{ij})]$. A denominator at most 10^{-12} yields NaN (converted to zero for predictor input).
Support overlap	Average of two fractions: points of P whose nearest point in X lies within their own fifth-neighbor radius in P, and the reverse fraction for X. Radii exclude only the aligned self entry; the distance comparison is inclusive.
Distance concentration	Population standard deviation divided by mean (floored at 10^{-12}), over concatenated within-P and within-X Euclidean distances. Both directions of every off-diagonal pair are included, including zero distances between duplicate rows.
Cancellation	$(\|T_+\|_F + \|T_-\|_F)/\max(\|T_+ - T_-\|_F, 10^{-12})$, with $T_+ = W_+ Y$, $T_- = W_- X$. The exact coefficient matrices and diagnostic support are specified below; this is not cancellation between unweighted joint payload sums.
Occupancy	Number of exactly distinct rows of P, divided by 128; no tolerance, quantization, or class grouping.
Width ratio	Feature generator hidden width divided by output feature dimension d.

For the joint diagnostic, $Y = P$ and the recorded matrices are $W_+ = \operatorname{diag}(n_0)A^+$ and $W_- = \operatorname{diag}(p_0)A^-$ from Equation (2): the opposing scalar masses are included. This joint diagnostic is also used for Sinkhorn; it is not a decomposition of its transport field. For DriftXpress, Y is the full fixed target and W_+ and W_- are separate row softmaxes of $-\|x_i - y_j\|_2/\tau$ and $-\|x_i - x_j\|_2/\tau$, respectively, with aligned negative self entries masked. This dense diagnostic uses no additive denominator epsilon; the projected candidate's stabilization is a separate convention. The statistic uses these weighted payload sums, whose difference is the separate-row drift in exact arithmetic.

The mechanism predictor uses these source-spectrum and geometry quantities, not the separately logged field-spectrum or two-nearest-neighbor estimates. They specify the retained secondary analysis; none is promoted to a validated selector.

D.5 Evidence hierarchy and limits of inference

The evidence levels are distinct: a fixed-support operator identity; compatible-reference fidelity tests; 240 feature runs; 24 low-capacity raw screening runs; and the separate completed 66-run

generator–encoder audit in Appendix G. Completion and finite metrics do not by themselves validate the quality evaluator.

The paper therefore does not claim a universal accelerator, a validated selector, an isolated causal representation theorem, a state-of-the-art FID result, or a pure-encoder-backward bottleneck. Nor does the historical rare FID reference establish rare-prior fidelity. These are substantive boundaries, not merely missing implementation detail. Broader generalization requires independent validation beyond the configurations examined here, exact native-operator training controls, and quality-versus-charged-time comparisons.

E Notation and Abbreviation Reference

This table consolidates the notation and abbreviations. Local uses are distinguished explicitly; $\mathbb{R}$ denotes the real numbers. Symbols appearing only in an appendix are defined locally there.

Table 12: Notation and abbreviation reference.

Symbol	Meaning
n, m_+, d, p	Numbers of queries/generated sources and positive sources; feature dimension; number of generator parameters.
X, P, N	Generated queries and positive/generated source payload matrices.
K^+, K^-, K	Positive, generated, and concatenated kernel blocks; aligned generated self interactions are deleted before degrees are computed.
r, c, A	Row degrees, column degrees, and normalized joint kernel matrix.
$\mathbf{1}, \mathrm{diag}, \mathrm{vec}$	All-ones vector of the required size, diagonal-matrix construction, and vectorization.
p_0, p_1, n_0, n_1	Positive/generated scalar masses and vector payload sums; these are not class probabilities.
μ_i^+, μ_i^-, V_i	Normalized positive/generated payload means and row-i drifting vector.
$g(j), z_\ell, U_{i\ell}, E_\ell$	Source-to-group assignment, representative, query–representative kernel, and column sum in Proposition 1. The assignment $g(j)$ is distinct from the generator g_θ.
$\alpha, \beta, S_\ell, B_{i\ell}^-$	Normalization exponents, sum of inverse per-copy degrees, and exact grouped generated-source coefficient.
$\theta, g_\theta, h, z, X_\theta$	Generator parameters, generator, frozen field encoder, noise batch, and generated features.
a, η, J_θ	Loss scale, drift scale, and feature Jacobian; sg denotes stop-gradient.
p, q_t, G_θ, ϕ (Fig. 1)	Target and current generated distributions, generator, and frozen encoder. Here $G_\theta = g_\theta$ and $\phi = h$; this local p is not parameter count, and this ϕ is not the cost fraction.
$\|\cdot\|_F, \|\cdot\|_{\mathrm{op}}$	Frobenius and Euclidean operator norms.
T, T_F, T_R; $\phi, s, S(s)$	Baseline inclusive, field, and remainder times; field fraction; field speedup; counterfactual inclusive speedup.
$p_9, p_9^{\mathrm{eval}}, \widehat{q}_9, U_9$	Target class-9 mass, empirical evaluation prior, classifier-estimated generated mass, and one-sided undercoverage.
Recall; class coverage; undercoverage	Manifold recall measures real-feature coverage; class coverage is the fraction of labels receiving at least one classifier prediction; rare undercoverage measures only a one-sided mass deficit. They are not interchangeable.
RKHS; RBF; MMD	Reproducing kernel Hilbert space; radial basis function; maximum mean discrepancy (reported primary export is an average of unbiased squared estimators).
τ, ρ	Kernel temperature and Spearman rank correlation.
FID; KID; EMA	Fréchet inception distance; kernel inception distance; exponential moving average. Subscript tv identifies the torchvision-feature evaluator.
NI; MAE; SD; BA; CI	Non-inferiority; mean absolute error; standard deviation; balanced accuracy; confidence interval.
ViT; SiLU	Vision transformer; sigmoid linear unit.

Symbol	Meaning
SVHN; RGB; GPU	Street View House Numbers; red, green, and blue image channels; graphics processing unit.
LODO; LOMO	Leave-one-dataset-out and leave-one-method-out validation, respectively.

F Method Parameters and Decision Rules

This appendix describes the completed feature benchmark and low-capacity raw screen, not the separately reported completed generator–encoder extension. The retained settings are historical operating points; the manuscript does not interpret them as individually optimized alternatives or as a new universal accelerator.

F.1 Field and schedule specification

The joint dense and representative routes use the Euclidean-distance kernel $k(x, y) = \exp(-\|x - y\|_2/\tau)$ with $\tau = 0.5$, rather than a squared-distance Gaussian kernel. Features are sample-normalized. The aligned generated self entry is removed before the joint row/column normalization in Equation (1). The dense implementation uses log-sum-exp degree accumulation. The count-expanded route instead evaluates the grouped coefficients in Proposition 1 on its explicitly coarsened support; its degree threshold is zero. DriftXpress uses its separate-row projected-attraction/exact-repulsion contract, not these joint-normalization coefficients. Its additive denominator stabilizer is 10^{-8}, its projection ridge is 10^{-4}, and its eigenspectrum floor is 10^{-8}. Field root-mean-square rescaling is disabled in this retained configuration.

Exact separate-row reference and projected attraction. For the identity-scaled feature vectors used by DriftXpress, let p_j and n_j be positive and generated support points, respectively, with aligned generated point $n_i = x_i$. Here $\varepsilon = 10^{-8}$ is the additive denominator stabilizer. Its compatible exact field is

$$V_i^{\mathrm{sep}} = \frac{\sum_j k(x_i, p_j)p_j}{\sum_j k(x_i, p_j) + \varepsilon} - \frac{\sum_{j\neq i} k(x_i, n_j)n_j}{\sum_{j\neq i} k(x_i, n_j) + \varepsilon}. \tag{12}$$

This is a difference of separately normalized means, not Equation (1). For positive shard a with landmarks U_a, the projected feature is $\psi_a(x) = (K_{U_aU_a} + \lambda I)^{-1/2}k(U_a, x)$, where $K_{U_aU_a}$ is the landmark Gram matrix, I is the identity, $\lambda = 10^{-4}$, and the stated eigenvalue floor is applied. Cache $b_a = \sum_{p\in a} \psi_a(p)$ and $C_a = \sum_{p\in a} \psi_a(p)p^\top$. Projected attraction is $[\sum_a \psi_a(x)^\top C_a]/[\sum_a \psi_a(x)^\top b_a+\varepsilon]$; repulsion remains the exact second term of Equation (12). Shard numerators and denominators are summed *before* division. The retained denominator mode adds 10^{-8}; it does not clamp projected masses, clip negative projected entries, or fall back to exact attraction. A positive-semidefinite projected Gram construction need not be entrywise nonnegative. The nonnegative-kernel assumptions in Section 3 therefore concern the exact joint operator, not an unproved property of the projected approximation. No per-update sign/minimum-mass telemetry was retained, so finite completed losses do not establish that this boundary was never approached. These formulas describe the retained implementation of Falahati et al. (2026). Approximation error requires matching positive-support information as well as the native operator; a cached full-support projection versus a fresh dense minibatch changes more than approximation accuracy.

In their rare-support stages, both representative schedules retain a conservative latent safeguard: base-anchor fraction 0.875, uniform mixture 0.75, scoring weights 0.1/2.0/0.5 for undercoverage/geometry/diversity, minimum score entropy 4.0, and acceptance tolerance 0.25. The base uses four scrambled Sobol-normal blocks of 32 points each. The formulas below specify the weighting, proposal, acceptance, and fallback rules. Quality uses isolated, side-effect-free probes in its balanced branch; the retained rare branch disables probe teachers altogether. The threshold values alone therefore do not define a freely adaptive teacher policy.

Sinkhorn coupling-to-field map. For positive and generated supports, construct separate Gibbs matrices $\exp(-D/\tau)$ with Euclidean cost D and $\tau = 0.5$. Row and column marginals are uniform on their respective supports. The negative coupling retains its aligned diagonal: unlike the dense joint

Table 13: Retained settings that distinguish the completed method routes. Counts refer to configured supports or pools, not equal information exposure across all methods.

Route or component	Setting
Optimized dense	Feature query block 512; raw query/source blocks 64/128. Full generated self-masked support; joint degrees.
DriftXpress	Random per-class landmarks; ratio 0.2; sharded and folded positive summary; summary chunks 512; scale-landmark limit 256; absolute bandwidth and identity feature scaling. Positive summaries can expose more target information than a freshly sampled dense minibatch.
Fast, balanced branch	Geometry supports $(m_+, m_-) = (256, 384)$; uniform independent subsets with importance weights. The rare branch's class quotas and radial bins are inactive here.
Fast, rare branch	Supports $(128, 128)$, including 16 replacement latent points from 512 candidates. Real pool 4,096; four radial bins; class quota at least one, exponent 0.5. Rescore every 500 steps after 500-step warm-up; imbalance ratio at least two activates this branch.
Quality schedule	Balanced geometry and rare refinement use $(m_+, m_-) = (256, 384)$; rare first-half support uses $(128, 128)$. Balanced probes: size 128 every 25 steps; relative-error/cosine thresholds 0.18/0.985 with the slot rule below. Rare probe teachers are disabled.
Teacher and polish limits	Balanced teacher budget: at most 0.18 of the horizon and 720 slots; 25 updates between activation triggers. Rare 0.10/400-step settings are dormant. Final dense polish is 5% of the horizon, clamped to 25–200 steps. Ordinary periodic teachers are disabled.
Count-expanded	384 representatives; random current representatives with nearest assignment; refresh every step; stop-gradient assignments; query and assignment blocks 2,048; representative blocks 384.
Sinkhorn	Separate positive/generated entropic couplings; Euclidean distance; temperature 0.5; 50 configured scaling iterations. It is not an exact joint-field implementation.

route, this configured Sinkhorn field does not delete generated self entries. Scaling runs in the log domain, initializing both log scaling vectors to zero. One iteration is a complete row update followed by a column update using log-sum-exp, rather than one half-cycle; the fixed 50 cycles do not use residual-based early stopping. This is not the joint geometric-mean normalization. If Π^+ and Π^- are the resulting couplings and $R(\Pi)_{ij} = \Pi_{ij} / \sum_k \Pi_{ik}$, the implemented field is

$$V^{\mathrm{OT}} = R(\Pi^+)P - R(\Pi^-)N.$$

Thus the field uses row-stochastic barycenters of the two plans, not their unnormalized difference. The feature comparison uses the retained 50-iteration split-coupling setting; this method is not part of the four-route image extension.

Update ordering. At each update, draw the prescribed target/noise streams; compute detached generated features; select the route's support or reuse its permitted summary; evaluate its native field; and form the detached target of Equation (5). The quality route gives final polish priority, otherwise applies the stage/slot/probe constraints before choosing a representative or dense teacher field. Backpropagate the regression through the generator and, in the raw screen, through the frozen field encoder, then update Adam and EMA where applicable. Selection, probes, teachers, summaries, and evaluation are charged according to the stated timing convention. Exactness of the grouped field does not certify this entire schedule or the support approximation.

F.2 Representative weights and rare-support safeguard

Where the weights enter. For selected positive and negative sources y_j with masses w_j, form

$$A_{ij} = \max(w_j, 10^{-8}) \exp(-\|x_i - y_j\|_2/\tau), \qquad C_{ij} = \frac{A_{ij}}{\sqrt{(\sum_k A_{ik})(\sum_l A_{lj})}}. \tag{13}$$

The row sum spans both signs and the column sum spans all queries. An aligned selected generated self entry is zeroed before these sums. With $a_i^{\pm} = \sum_{j \in \pm} C_{ij}$, the representative field is $V_i =$

$a_i^- \sum_{j\in+} C_{ij}y_j - a_i^+ \sum_{j\in-} C_{ij}y_j$. Thus masses enter *before both normalizations*; they are not multipliers applied afterward. This weighted representative operator is not the exact count-expansion rule of Proposition 1.

In geometry stages, draw real and generated subsets independently, uniformly without replacement, on every update. Their masses are B_+/m_+ and B_-/m_-, respectively, for current support sizes B_+, B_- and selected sizes m_+, m_-. For Fast's balanced stage and Quality's balanced geometry and rare refinement, $B_+ = B_- = 512$ and $(m_+, m_-) = (256, 384)$. The setting "geometry budget 384" names only the generated subset, not both supports. Dense teacher/polish updates instead use $(512, 512)$. Negative indices retain their original query identities for self deletion. Rare-support negatives are instead generator images of a separate fixed/adapted latent set: each carries mass $512/128$, and no aligned query identity is removed.

Static rare-support positives. Use normalized class-mean prototypes, target class frequencies p_c, and nearest-prototype assignments in feature space. Let N_c count the 4,096 real candidates assigned to prototype c, and let f_c be the current generated-minibatch assignment fraction. Allocate the 128 positive representatives using scores $q_c = \sqrt{p_c} + 2(p_c - f_c)_+ + 10^{-8}$, setting $q_c = 0$ for empty candidate classes. First reserve one point per nonempty class in decreasing-score order. Distribute the remaining budget by repeatedly flooring allocations proportional to q_c over classes with remaining capacity, capping by that capacity; when all floor allocations are zero, assign one point to the largest fractional remainder. Within each class, sort by distance to its prototype and put zero-based rank r in bin $\min(\lfloor 4r/\max(N_c - 1, 1)\rfloor, 3)$. Split the class quota evenly over nonempty bins, give remainders to earlier bins, and draw without replacement; fill capacity shortfalls from remaining class points. For selected point j in class $c(j)$, with n_c selected points in that class, use

$$\widetilde{w}_j = N_{c(j)}/n_{c(j)}, \qquad w_j = 512\,\widetilde{w}_j / \sum_{k \text{ selected}} \widetilde{w}_k.$$

The pool, selection, and masses are then reused, not rebuilt at each latent refresh. These routing assignments are distinct from the independent evaluation classifier.

Latent proposal and refresh. The negative budget is *128 in total*, not 128 plus 16: keep 112 base points and propose replacing 16. Build the 128-point base from four scrambled Sobol blocks of 32 points, with seeds $s + 1000 + 104729e$, $e = 0, \ldots, 3$, for run seed s. Apply the standard-normal inverse CDF coordinatewise after clipping uniforms to $[10^{-7}, 1 - 10^{-7}]$. The separate cached 512-candidate Sobol set uses seed $s + 1811$. The retained anchors are the ordered base points with zero-based indices $0, \ldots, 111$ (the first three 32-point blocks and the first 16 points of the fourth); indices $112, \ldots, 127$ are proposed for replacement. There is no anchor shuffle, score ranking, or rotating subset at refresh. Base latent values and this prefix are cached; their generated images and candidate scores are recomputed using the current generator. Unavailable ensemble construction first falls back to a single Sobol-normal set; unavailable Sobol construction falls back to standard-normal draws from the method-specific pseudorandom-number generator (RNG). Before update 500, retain the full base. From update 500, refresh on the first eligible rare-support call and then after at least 500 updates since the last refresh; cache the accepted or fallback set between refreshes. In particular, the 1,000-update Quality schedule leaves its rare-support stage at update 500, before any such refresh is eligible during training.

For a candidate with nearest-prototype label c and distance r, let f_c^{cand} be candidate assignment frequencies, $u_c = (p_c - f_c^{\text{cand}})_+$, and let μ_0, σ_0 be the base set's mean and population standard deviation of nearest-prototype distances. Define $\mathcal{N}(v) = v/\max(\text{mean}(v), 10^{-12})$ over candidates. The balance, deficit, geometry, and diversity components are

$$b_j = \mathcal{N}\Big(p_{c(j)}/\max(f_{c(j)}^{\text{cand}}, 1/512)\Big), \qquad u'_j = u_{c(j)}/\max(\max_c u_c, 10^{-12}),$$
$$g_j = \mathcal{N}\big(\exp\{-|r_j - \mu_0|/\max(\sigma_0, |\mu_0|/4, 10^{-6})\}\big),$$
$$v_j = \mathcal{N}\big(\min\{d_j/\max(\text{median}(d), 10^{-6}), 2\}\big).$$

Here d_j is distance to the closest of the 112 retained generated anchors; each median uses the lower central order statistic when its input count is even. Start with scores $h_j = 0.75 + 0.25(b_j + 0.1u'_j + 2g_j + 0.5v_j)$, floored at 10^{-12}. Make 16 sequential single-candidate draws. At each

draw, recompute distance δ_j to the anchors and already selected candidates, and multiply h_j by $1 + 0.125 \min\{\delta_j / \max(\text{median}_{\text{unselected}}(\delta), 10^{-6}), 2\}$. Zero already selected scores, then cast to float32, floor every score at 10^{-12}, normalize, and draw. Thus selected indices regain a tiny positive probability: the implementation does not strictly enforce sampling without replacement across the 16 calls. This numerical qualification does not establish that any historical draw repeated an index. A failed multinomial draw falls back to the maximum adjusted score before the normalization floor.

Acceptance, not an unconstrained resampling step. For a generated set S, write $f_c(S)$ for its prototype-assignment fractions and define

$$M(S) = \sum_c |f_c(S) - p_c|, \qquad U(S) = \max_c (p_c - f_c(S))_+,$$
$$D(S) = \frac{1}{|S|} \sum_{x \in S} \frac{1}{\max(\min_{y \neq x} \|x - y\|_2, 10^{-6})}.$$

The implementation excludes self distances by adding 10^6 on the pairwise diagonal. Let S_0 be the current generator's images of the full base and S_1 the proposed 112+16 set. Using their nearest-prototype distance means and population standard deviations, define

$$G = \frac{|\mu_1 - \mu_0|}{\max(|\mu_0|, 10^{-6})} + \frac{|\sigma_1 - \sigma_0|}{\max(|\sigma_0|, 10^{-6})}, \qquad D_{\text{rel}} = D(S_1) / \max(D(S_0), 10^{-6}).$$

Accept only when *all* of the following hold: natural-log entropy of normalized h is at least 4; the number of represented prototypes does not decrease; $M(S_1) \leq M(S_0)$ *or* $U(S_1) \leq U(S_0)$; $G \leq 0.25$; $D_{\text{rel}} \leq 1.25$; and

$$M(S_1) + 0.1U(S_1) + 2G + 0.5(D_{\text{rel}} - 1)_+ \leq M(S_0) + 0.1U(S_0) + 0.25.$$

On rejection, revert to the full base, *not* the previously accepted adaptation. On either outcome, cache the chosen 128 points until the next eligible refresh. These are prototype-based safeguards, not guarantees of semantic coverage or improved sample quality.

F.3 Executable quality schedule

Updates are numbered $t = 1, \ldots, T$. Let R denote round-to-nearest with ties to even. Final dense polish has length

$$P = \min\{T, 200, \max(25, R(0.05T))\},$$

and takes precedence for every $t > T - P$. The balanced branch's last eligible teacher slot is $L = \min\{T - P, R(0.95T)\}$ and its slot budget is $K = \min\{R(0.18T), 720, L\}$. For the retained horizons ($K > 1$), the exact eligible set is

$$\mathcal{S} = \left\{ 1 + R\left(\frac{j(L-1)}{K-1} \right) : j = 0, \ldots, K-1 \right\}. \tag{14}$$

These positions are unique at all three horizons below. They are *eligible* slots, not a claim that threshold tests must activate all of them.

Balanced branch: activation and update order. Start with no active or consumed slots. Outside final polish, consume an already active slot at t first and use its dense teacher field. Otherwise, at $t = 1 + 25j \leq L$, evaluate a separate geometry representative proposal using a fresh isolated RNG with seed $(1000003s + 9176t + 315940) \bmod (2^{31} - 1)$ for run seed s. For $B = 512$ queries, probe the 128 indices

$$I_t = \{[R(j(B-1)/127) + 37t] \bmod B : j = 0, \ldots, 127\}.$$

The dense reference still computes column degrees over *all* queries before restricting to I_t. For representative and dense vectors $\widehat{V}_i, V_i$ on these indices, compute

$$e = \frac{\text{mean}_{I_t} \|\widehat{V}_i - V_i\|_2}{\max(\text{mean}_{I_t} \|V_i\|_2, 10^{-8})}, \qquad c = \text{mean}_{I_t} \frac{\langle \widehat{V}_i, V_i \rangle}{\max(\|\widehat{V}_i\|_2, 10^{-8}) \max(\|V_i\|_2, 10^{-8})}.$$

If $e > 0.18$ or $c < 0.985$, at least 25 updates have elapsed since the previous activation trigger (or none exists), and fewer than K slots have been consumed, activate every as-yet-unused slot in $\mathcal{S} \cap [t, \min(t + 24, L)]$. Consume t immediately if it has just become active; otherwise perform the ordinary geometry update using the training method's random stream, not the discarded probe proposal. A failed threshold leaves that block's unused slots inactive. Each teacher slot is consumed at most once, with at most K teachers. The 25-step minimum gap is between *activation triggers*, not between individual teacher updates; the one-step burst parameter does not restrict a triggered block to one slot. There is no additional periodic-teacher mechanism.

Rare branch. The retained regime switch disables both probe eligibility and teacher-slot activation, not merely the balanced slot restriction. Use rare-support representatives for $t/T < 0.5$, geometry refinement for $t/T \geq 0.5$, and the same final-polish priority. The configured rare thresholds and 400-step cap are dormant; they do not specify an active rare teacher policy.

Derived schedule quantities. The following are implied by the retained settings, not additional experiments. "All-active dense" includes polish and applies only if every balanced slot is activated; rare-stage counts exclude polish.

T	P	L	K	All-active dense	Rare support	Refinement
1,000	50	950	180	230	499	451
10,000	200	9,500	720	920	4,999	4,801
15,000	200	14,250	720	920	7,499	7,301

The resulting 15,000-update counts agree with the twelve-model schedule summary in Appendix H. For the six balanced runs, 720 teachers and the at-most-once consumption rule imply full slot use, conditional on the stated implementation. The recorded slot fingerprint identifies the configured eligible set, not an independently saved consumed-index trace. CPU fixtures reproduce the all-trigger and no-trigger schedules at each horizon; they do not replay the historical field-dependent probe trajectory. This specification does not establish a causal benefit of the schedule or replace a matched-operator control.

F.4 Raw role predicates

The four raw roles are evaluated separately at seed means and at every one of the three paired seeds. Lower scores and costs are preferred. A point is Pareto-dominated only if another point is no worse in both coordinates and strictly better in at least one; no extra practical tolerance is applied. The four quality–cost views combine the two regimes with $\mathrm{FID}_{\mathrm{tv}}$ or $\mathrm{KID}_{\mathrm{tv}}$.

DriftXpress: broad quality–wall role. It must belong to at least three of the four quality–wall Pareto views both at the aggregate and at each seed, and must not be slower than dense in either regime at any seed.

Dense: quality–memory role. Dense must remain on the quality–memory frontier in all four views, both at the aggregate and at each seed.

Fast: balanced speed with rare risk. It must have minimum balanced wall time at the aggregate and each seed. For both quality scores, its excess over dense must be positive in the rare regime and larger there than in the balanced regime, consistently across the aggregate and seeds.

Quality schedule: near-dense quality–wall role. In each regime, its $\mathrm{FID}_{\mathrm{tv}}$ may be at most 1.10 times dense, its $\mathrm{KID}_{\mathrm{tv}}$ excess at most 0.002, and its wall time no greater than dense; the rule is checked at the aggregate and every seed.

A role not meeting all conditions is unresolved unless its separately specified reversal holds coherently at all three seeds. The reversal tests require both quality scores: worse quality and wall for DriftXpress, quality–memory domination of dense, reversal of the fast route's rare-risk direction, or the quality route being outside its near-dense margins and slower. In the completed screen, only the broad DriftXpress role transfers, and none has a stable coherent reversal. This does not mean three methods

are proved inferior. The rare-reference mismatch also limits the target-fidelity interpretation of these historical predicates.

F.5 Feature hypothesis and Pareto definitions

H1 compares the compatible and incompatible relative field errors on the same recorded states; its directional aggregate is the median of incompatible minus compatible relative error. Checkpoints within a trajectory are not independent experimental replications. H2/H3 predictor construction, targets, ridge selection, and defects are specified in Appendix D; no prospective learned decision rule is validated.

For H5, each dataset/regime has the views energy–wall, MMD–wall, energy–peak memory, MMD–peak memory, and energy–selected-edge fraction. Rare regimes add undercoverage–wall. Pareto membership uses coordinatewise dominance without an additional practical tolerance. Ten thousand bootstrap resamples draw the same seed indices for all methods in a cell. A membership probability of at least 0.70 is labeled stable. The observed multi-view memberships are descriptive operating roles, not uniform non-inferiority or universal dominance. A stable undercoverage–wall position does not independently demonstrate calibration, especially given the one-sided metric's limitation.

G Completed Generator–Encoder Audit

This appendix gives selected endpoint metrics for every final model, not every recorded metric or diagnostic vector. The design and aggregate comparisons precede the checkpoint analysis (Section G.4) and identical-image input intervention (Section G.5); complete model coverage is distinct from exhaustive metric release.

G.1 Design, completion, and recovery

This extension tests the fidelity–cost–calibration question across generator and encoder configurations, complementing the 240-run feature benchmark and 24-run low-capacity image screen. All 66 planned runs finished 15,000 updates at batch 512, with seeds 101, 202, and 303. The main matrix comprises CIFAR-10/SVHN $\times$ balanced/rare targets $\times$ four methods $\times$ three seeds (48 runs). Eighteen additional balanced CIFAR-10 runs complete the generator–encoder bridge. The six U-Net/DINOv3 dense/DriftXpress runs are shared with the main matrix, so the bridge's 24 rows are not 24 additional training runs.

The bridge cells are the 2,898,435-parameter residual generator with ResNet-18 (B_0) or DINOv3 (B_1), and the 38,245,251-parameter conditional U-Net with ResNet-18 (B_2) or DINOv3 (B_3). DINOv3 is frozen ViT-B/16, with 112-pixel inputs and spatially pooled 12,288-dimensional features. Unit-range RGB images are resized by antialiased bilinear interpolation and normalized with ImageNet channel statistics. The normalized $7 \times 7 \times 768$ patch-token map is adaptively average-pooled to 4×4, flattened, and ℓ_2-normalized per sample. Optimized dense uses its joint operator and DriftXpress its native separate-row projected field; the latter has 2,000 landmarks in the balanced bridge (1,820 in rare-target main-matrix runs). Thus this is a pipeline comparison, not an isolated projection ablation. The encoder intervention also changes preprocessing, feature dimension, computation, and the elementwise-mean loss prefactor; the generator intervention changes architecture, noise parameterization, and masking. Initial generator-state hashes match across methods and across encoders at each fixed generator/seed. No seed, update budget, or endpoint was selected using terminal scores.

Architecture and optimization. The U-Net receives $3 \times 32 \times 32$ Gaussian noise and class conditioning of dimension 512. Its four resolution levels have widths $(128, 256, 256, 256)$, with two residual blocks per encoder level, three per decoder level, group normalization and sigmoid linear unit (SiLU) activations, four-head attention at resolution 16, deterministic structured masking at nominal drop rate 0.1, and a sigmoid RGB output. The small residual generator uses the 128-dimensional noise input described in Appendix B. All extension runs use AdamW (Loshchilov & Hutter, 2019) with learning rate 2×10^{-4}, coefficients $(0.9, 0.999)$ and zero weight decay, gradient-norm clipping at 2, and exponential-moving-average decay 0.9999. Generator and encoder forwards use bfloat16 autocasting; parameters, optimizer state, moving averages, fields and losses use float32. Gradients

are accumulated in chunks of eight over each 512-sample update, with a global-mean loss. Each run uses one A100. These optimization settings differ from the earlier low-capacity screen and are held fixed across the extension's method comparisons. Field and schedule settings are inherited only where explicitly listed in Appendix H, which also specifies mask replay, marginal-field conditioning, numerical guards, and evaluator preprocessing.

Adaptation, not published-recipe reproduction. These are clean-room configurations, not a reproduction of the attainable quality of a published drifting recipe. For example, the DriftXpress paper's experimental setup specifies temperature 0.05, global batch 12,000, and an eight-hour main-benchmark budget on eight H100 GPUs (Falahati et al., 2026, Appendix C.1); this extension uses temperature 0.5, batch 512, and 15,000 updates on one A100. Our terminal $\mathrm{FID}_{\mathrm{tv}}$ protocol also differs from its 50,000-sample FID evaluation. The comparison here is between our declared configurations; poor quality under this protocol is not evidence that the original published model cannot generate well.

EMA convention and saved-state diagnostic. The retained implementation initializes the exponential moving average (EMA) from the initial model and updates it once after each optimizer step with fixed decay $\beta = 0.9999$, without warm-up. Thus, in exact arithmetic,

$$\bar{\theta}_T = \beta^T \theta_0 + (1-\beta) \sum_{t=1}^{T} \beta^{T-t} \theta_t.$$

At $T = 15{,}000$, $\beta^T \simeq 0.2231$ and the half-life is about 6,931 updates. This is a parameter-space coefficient, not a mixture weight on generated images. Final evaluation uses EMA parameters; the checkpoint-writing convention retains the unaveraged generator and the EMA state separately. Saved counters and both parameter states have now been checked for a predefined six-model balanced CIFAR-10 subset. Appendix G.4 reports the complete raw-versus-EMA evaluation: some FID values improve, but raw recall remains zero in five models and 0.0002 in the sixth. The coefficient alone does not establish evaluation lag or explain poor generation.

Sixty jobs published both evaluation tracks directly. Six rare-regime Quality jobs completed training and wrote final checkpoints, then exceeded an application watchdog during evaluation. Their original failed statuses and costs are retained. A separate evaluation-only recovery verified the final checkpoint and EMA identities, performed no training or resume, and supplied both tracks for all six. An already-completed dense model was reevaluated as a parity control; it is not counted as a new model. Therefore the scientific table contains **66 distinct final models and 132 evaluations**, not 60 successes with six omitted outcomes.

Training-component time is distinct from allocation time. The former sums non-overlapping training counters; field construction is nested in field time, never added twice. Reported per-model allocation includes the original job, including any evaluation failure, and its successful evaluation-only recovery. It excludes queue delay, external preparation, and diagnostic or failed recovery attempts. Original scientific allocations used 2,621.60 GPU-hours, including failures. Successful recovery plus the dense parity control used 0.355 GPU-hours. These are experiment-execution costs, not a total for development and recovery diagnostics.

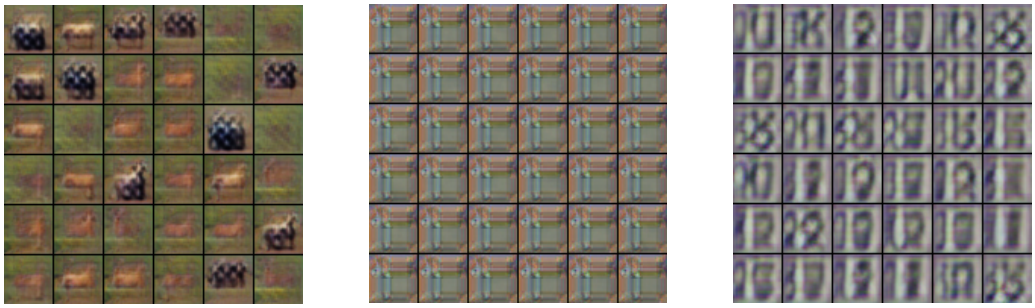
G.2 Quality evaluation and what the controls establish

Both tracks use frozen torchvision Inception-v3 features, independently of the training encoder. Primary evaluation draws 10,000 generated images and uses the balanced 5,000-image held-out reference. Secondary evaluation uses a distinct, fixed generation stream: 10,000/5,000 generated/real images for balanced targets, and 9,100/4,550 for rare targets. For rare secondary evaluation, the generation stream requests 1,000 images per majority conditioning label and 100 for class 9; the real reference contains 500 per majority class and 50 for class 9. Both the request prior and the real-reference prior therefore have $p_9 = 1/91$. These are requested labels, not verified semantic classes of generated images. Scores are $\mathrm{FID}_{\mathrm{tv}}$ and $\mathrm{KID}_{\mathrm{tv}}$, not externally interchangeable leaderboard values. Manifold precision/recall uses $k = 3$. The secondary prior-matched track is the target-fidelity analysis; primary balanced-reference results remain visible rather than being silently relabeled.

The recovery dense control reproduced all eleven metric values exactly on both tracks, with matching generated-image and feature identities. Synthetic controls checked identical and disjoint feature

sets, duplicates, rank deficiency, self masks, distance units and block-size invariance. All returned input-range checks passed. Fixed disjoint real-feature subsets, 512 versus 512, gave recall 0.684–0.814. These subset calculations use the deployed metric convention with CPU distances; they are not full-size A100 validation, recalibration or independent replication.

The checks argue against an always-zero implementation, but *do not establish healthy generation*: secondary recall is zero in 45 of 48 U-Net/DINOv3 rows (primary: 43 of 48). All 18 additional bridge rows have nonzero recall, showing that the encoder/generator operating point matters. Figure 10 preserves the three preselected pilot models, including repetitive or unrecognizable images. Classifier assignments on these outputs cannot be assumed to be semantic ground truth. Neither a finite score nor the narrow recorded non-collapse flag establishes adequate diversity.



CIFAR-10, Dense | CIFAR-10, Quality | SVHN, Quality

Figure 10: **Actual generated images, not conceptual artwork.** The preselected recovery pilot uses rare-target seed-101 U-Net/DINOv3 models: two failed-evaluation models and a completed dense control. Each grid selects 36 primary-stream images by a fixed uniform index permutation (selection seed 20261017), without outcome filtering. Raw 32×32 pixels are clipped only for display and rounded to 8-bit values; grids are enlarged here. These examples support caution about diversity, not a quantified or blinded human quality assessment.

G.3 Complete bridge: cost leverage and terminal quality

Table 5 reports all four cost–quality cells. The ratios of mean per-job charged allocation are $1.357\times$, $1.122\times$, $1.062\times$, and $1.047\times$ in its row order; they exclude external preparation and queueing, as defined above. With the encoder fixed, changing to U-Net leaves field speedup approximately unchanged, but greatly reduces training leverage. For ResNet-18, field speedup is $54.45\times$ versus $53.49\times$, while training speedup falls from $1.365\times$ to $1.063\times$; with DINOv3 the corresponding values are $40.94\times$ versus $40.53\times$ and $1.125\times$ versus $1.048\times$. Dense field time stays near 2.4 hours while total dense training grows from 8.62 to 44.92 hours across the four cells. Absolute training savings remain 2.01–2.30 hours. This locates the change in leverage in the surrounding workload rather than loss of the local field saving. It does not isolate hardware scheduling noise or representation geometry from execution.

Under the historical input path, the low-capacity bridge gives six of six paired secondary FID improvements for DriftXpress. With U-Net the signs become mixed: one of three improvements per encoder, and worse means. Table 14 retains every sign. The larger generator configuration is therefore not evidence of better generation under this fixed protocol; the experiment does not identify why its optimization or learned distribution deteriorates. Native fields and support exposure also differ between dense and DriftXpress, precluding attribution to projection alone.

G.4 Saved-checkpoint cost–quality and raw-versus-EMA evaluation

Scope and selection. This follow-up evaluates existing checkpoints without training, optimizer steps, EMA updates, or changes to historical results. The bridge block contains the six balanced CIFAR-10 small-generator/ResNet-18 Dense/DriftXpress models (seeds 101, 202, 303), each at 5k, 10k, and 15k updates using EMA parameters: 18 evaluations. The final-state block contains the six corresponding balanced CIFAR-10 U-Net/DINOv3 models, each with raw and EMA parameters at 15k: 12 evaluations. All 30 cells and evaluation settings were fixed before these new outcomes.

Table 14: Complete bridge quality and seed sensitivity under the historical input path. ΔF is DriftXpress minus dense prior-matched $\mathrm{FID}_{\mathrm{tv}}$; negative is better. Ranges are minimum/maximum paired effects over three seeds, not confidence intervals. Means include unfavorable outcomes; the selected input-path comparison is in Appendix G.5.

Generator	Encoder	Mean ΔF	Paired range	FID wins
Small	ResNet-18	-2.70	$[-3.67, -1.00]$	3/3
Small	DINOv3	-5.23	$[-8.69, -0.66]$	3/3
U-Net	ResNet-18	$+10.88$	$[-3.31, +28.89]$	1/3
U-Net	DINOv3	$+11.28$	$[-37.65, +60.57]$	1/3

The bridge cell was chosen to examine an already-observed favorable endpoint, and the larger configuration to test the EMA concern. This is therefore a targeted descriptive follow-up, not a fresh independent discovery, a random configuration sample, or a complete trajectory comparison over all four bridge cells. All three seeds and all specified states/steps are retained, including unfavorable values. The optional bridge block was admitted using measured execution cost and memory only, not quality. Both blocks completed.

The 30 evaluations use 24 distinct checkpoint files and do not increase the 66-model training count. Twelve final-EMA evaluations reproduce existing outcomes; the other 18 evaluate previously unreported intermediate or raw states. Checkpoint bytes, parameter-state identities, saved update counters, and recorded timer fields were authenticated before evaluation. Every inspected EMA has decay 0.9999 and update count equal to the saved step. The final-EMA reruns match all 132 historical metric values (12 endpoints, 11 metrics each) exactly, with identical generated-image and Inception-feature identities. A tolerance of $10^{-7} + 10^{-6}|m_{\mathrm{old}}|$ was fixed beforehand; no discrepancy used that tolerance. These are same-implementation consistency checks, not independent replication or production gradient/update parity. The separate input-path intervention is reported in Appendix G.5; results in this subsection retain the historical input path.

Fixed evaluation protocol. Each cell generates 10,000 images and uses the same 5,000 balanced native CIFAR-10 test examples as the extension's secondary reference (500 per class, split seed 903). At a fixed scientific seed and generator configuration, both methods, states, and checkpoint steps share exactly the same ordered labels and Gaussian noise. Generate the label vector by concatenating 1,000 copies of each integer 0–9 and applying one full CPU `randperm` with the label seed below. Independently generate one full CPU float32 standard-normal tensor with the indicated noise seed and shape $10{,}000 \times 128 \times 1 \times 1$ for the small generator, or $10{,}000 \times 3 \times 32 \times 32$ for U-Net. Do not reseed per chunk.

Scientific seed	Label seed	Small-generator noise seed	U-Net noise seed
101	7497565167798192925	9014115170944622485	5136915333843343240
202	8986018996384044376	8876013952629036963	3789653012105501377
303	6722263838839714394	881596130590817862	8453059047370529288

Both states use evaluation mode and inference mode; "raw" means saved unaveraged weights, not training-mode stochastic masks. Generator forwards use bfloat16 autocasting in eight-image chunks, returning float32 images. The classifier uses float32 chunks of 128, and Inception uses float32 chunks of 32 without generator autocasting. TF32 is disabled and deterministic algorithms are enabled. The retained runtime uses Python 3.12.3, PyTorch 2.9.0a0+145a3a7bda.nv25.10, torchvision 0.24.0a0+094e7af5, CUDA 13.0, and one A100-SXM4-40GB. FID/KID, precision/recall, preprocessing, and classifier definitions are those of Appendix H, not a new leaderboard metric. In particular, KID uses all samples rather than random subsets, and manifold precision/recall uses $k = 3$.

For generated Inception features $f_1, \ldots, f_n$, the spread statistic is

$$D_g = \frac{2}{n(n-1)} \sum_{i<j} \|f_i - f_j\|_2.$$

Distances use float32 blocks of 256, with float64 accumulation. The narrow collapse statistic is the fraction of these pairs with distance at most 10^{-8}; all 30 values are zero. Spread and this narrow flag

do not by themselves certify useful coverage, semantic class correctness, or human-perceived image quality. Class-mass diagnostics remain classifier estimates.

What the time coordinate measures. The horizontal coordinate is the cumulative sum recorded *at that saved checkpoint* of target sampling, detached generator and encoder forwards, field work including construction, differentiable generator and encoder forwards, encoder image pullback, generator backward, clipping/optimizer work, and EMA updates. Construction is nested in field time and is not added again. This non-overlapping training-component sum excludes one-time setup, checkpoint writes, evaluation, and unrecorded wall overhead. No intermediate time is inferred by multiplying the final time by a step fraction. Full original per-model charged allocations remain reported separately; this evaluation adds 970 seconds on one A100 (0.2694 GPU-hours), not new training time or an adjustment of earlier endpoints.

Historical-input checkpoint results and limits. The historical-input panel of Table 1 reports the means, Figure 3 brings the FID/recall curves into the main text, and Table 15 retains all 18 individual points. At 10k and 15k, FID differences (DriftXpress minus dense) are respectively $[-1.6569, -1.0183, -0.9848]$ and $[-3.6719, -0.9997, -3.4202]$ for seeds 101/202/303. Recall differences are $[0.0046, 0.0008, 0.0166]$ and $[0.0058, 0.0052, 0.0116]$. Thus all three pairs improve both metrics at those two checkpoints under the historical input while reducing recorded cost. The internal-transform comparison retains this joint result at 15k, but not 10k (Appendix G.5). Ratios of mean dense to DriftXpress training-component costs are 1.3647, 1.3646, and 1.3649 at 5k/10k/15k; the full per-seed range is 1.3538–1.3709. At 15k, mean savings are 8,293.45 recorded seconds, or 2.3037 hours. This is a fixed-checkpoint cost reduction, not an estimate of the ratio of first-passage times to a chosen quality threshold.

At 5k, one paired FID difference is unfavorable ($+0.1061$), and mean recalls are only 0.000067/0.000133 for dense/DriftXpress. Mean spread changes are $+0.038\%$, $+0.451\%$, and $+0.294\%$ at the three checkpoints, but two of three final spread differences are slightly negative. KID and precision also have mixed per-seed signs. Classifier-estimated total-variation error decreases in all nine pairs, yet its final means remain 0.3271/0.3101; neither method matches the intended balanced class mass well. Hence these data support a local checkpoint cost–quality benefit, not dominance in every metric, a formal non-inferiority result, or high absolute generative quality.

Table 15: All 18 historical-input small-generator/ResNet-18 checkpoint evaluations, balanced CIFAR-10. T is cumulative recorded training-component hours; F, K are $\mathrm{FID}_{\mathrm{tv}}, \mathrm{KID}_{\mathrm{tv}}$; P, R are precision/recall fractions; D_g is mean pairwise generated Inception-feature distance; V is classifier-estimated total-variation mass error. No values are significance tests. All final EMA endpoints are reevaluated, not reused.

Step	Seed	Method	T (h)	$F\downarrow$	$K\downarrow$	$P\uparrow$	$R\uparrow$	D_g	$V\downarrow$
5k	101	Dense	2.8494	190.0952	0.188767	0.5097	0.0000	13.1326	0.6661
5k	101	DriftXpress	2.1032	188.9009	0.187309	0.5011	0.0000	13.1173	0.6598
5k	202	Dense	2.8533	202.5316	0.198446	0.5435	0.0002	12.8176	0.6882
5k	202	DriftXpress	2.0831	200.6689	0.197636	0.5371	0.0004	12.8996	0.6849
5k	303	Dense	2.9216	193.4213	0.194561	0.4978	0.0000	12.8699	0.6566
5k	303	DriftXpress	2.1332	193.5274	0.195636	0.5083	0.0000	12.8178	0.6540
10k	101	Dense	5.6947	95.7718	0.077544	0.6978	0.0046	16.4577	0.5398
10k	101	DriftXpress	4.2065	94.1149	0.076023	0.7020	0.0092	16.4584	0.5144
10k	202	Dense	5.7030	92.9194	0.069280	0.7420	0.0074	16.5080	0.5320
10k	202	DriftXpress	4.1632	91.9011	0.068929	0.7292	0.0082	16.6201	0.5148
10k	303	Dense	5.8417	97.6249	0.076493	0.8056	0.0028	15.4493	0.4488
10k	303	DriftXpress	4.2640	96.6401	0.077106	0.7985	0.0194	15.5546	0.4257
15k	101	Dense	8.5332	67.3195	0.045757	0.6649	0.0364	17.6840	0.3628
15k	101	DriftXpress	6.3030	63.6476	0.043425	0.6919	0.0422	17.6573	0.3475
15k	202	Dense	8.5551	60.8119	0.033756	0.6708	0.0278	17.9063	0.3259
15k	202	DriftXpress	6.2406	59.8122	0.033788	0.6821	0.0330	17.9052	0.3229
15k	303	Dense	8.7649	70.2686	0.045823	0.7939	0.0074	16.7300	0.2927
15k	303	DriftXpress	6.3984	66.8484	0.044395	0.7729	0.0190	16.9117	0.2600

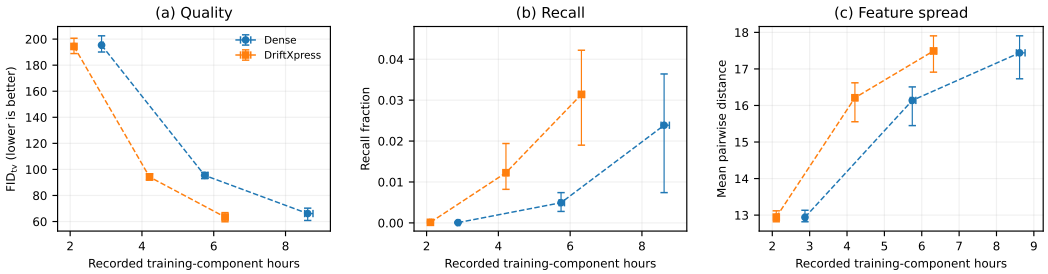


Figure 11: **Historical-input quality and spread versus recorded training cost.** Small-generator/ResNet-18 balanced CIFAR-10 bridge. Each method's points are the three-seed means at 5k, 10k, and 15k updates, in left-to-right order. Horizontal and vertical bars span seed minima/maxima, not confidence intervals. Dashed segments only connect observed checkpoints; they do not supply intermediate quality values or crossing times. Training-component cost excludes the overheads specified above. The three checkpoints are repeated observations of the same six models, not 18 independent training runs.

Three sparse points do not identify an exact time to matched quality. We report neither a fitted crossing-time speedup nor an outcome-selected quality threshold, and do not assume monotonic behavior between checkpoints. The configuration difference in normalization and source access persists. This follow-up tests the learning behavior of the declared Dense/DriftXpress configurations, not projection in isolation.

Historical-input raw/EMA comparison. Table 16 retains both states for every selected U-Net/DINOv3 model. Raw-minus-EMA FID is negative in four of six pairs and positive in two. Dense mean FID changes from 154.37 to 139.27; DriftXpress from 165.65 to 160.94. Mean precision decreases for both methods (0.7154 to 0.6525, and 0.7067 to 0.5669), while mean spread increases. All six EMA recalls are zero; five raw recalls are zero and DriftXpress seed 202 is 0.0002 (one of 5,000 real reference points). Thus replacing EMA by raw parameters can move FID but does not produce a functioning high-capacity reference in this subset. It neither proves EMA irrelevant to all quality measures nor isolates the cause of poor generation. Historical EMA endpoints are not replaced by the better state, and no claim is made about the 42 other main-matrix models.

Table 16: All 12 historical-input final-state evaluations for six balanced CIFAR-10 U-Net/DINOv3 models. Raw means unaveraged saved parameters in evaluation mode, not training-mode masking. Metrics are defined as in Table 15; no best-state replacement is made. All EMA values match their historical endpoints exactly.

Seed	Method	State	$F \downarrow$	$K \downarrow$	$P \uparrow$	$R \uparrow$	D_g	$V \downarrow$
101	Dense	EMA	161.2573	0.130692	0.6706	0.0000	13.9123	0.5602
101	Dense	Raw	148.1632	0.099058	0.5785	0.0000	15.3106	0.4578
101	DriftXpress	EMA	221.8285	0.188123	0.7205	0.0000	11.6331	0.7393
101	DriftXpress	Raw	235.6541	0.208337	0.6280	0.0000	12.2773	0.6017
202	Dense	EMA	130.1109	0.082433	0.7161	0.0000	16.1315	0.3011
202	Dense	Raw	96.4695	0.053891	0.6835	0.0000	17.5726	0.2368
202	DriftXpress	EMA	141.0381	0.103599	0.6873	0.0000	15.3824	0.3341
202	DriftXpress	Raw	127.0594	0.093014	0.5419	0.0002	16.9220	0.2411
303	Dense	EMA	171.7337	0.140064	0.7594	0.0000	12.8364	0.7460
303	Dense	Raw	173.1771	0.136574	0.6955	0.0000	13.1991	0.6468
303	DriftXpress	EMA	134.0791	0.092490	0.7123	0.0000	15.7906	0.3363
303	DriftXpress	Raw	120.1005	0.078811	0.5308	0.0000	17.2541	0.3394

Table 17: Remaining class diagnostics for all 30 checkpoint-state evaluations. C is the fraction of classes receiving at least one classifier prediction; $\widehat{q}_9$ is the classifier-estimated class-9 fraction. All targets here are balanced ($p_9 = 0.1$), so signed error, absolute error, undercoverage, and target ratio are $\widehat{q}_9 - 0.1$, $|\widehat{q}_9 - 0.1|$, $\max(0.1 - \widehat{q}_9, 0)$, and $\widehat{q}_9/0.1$. Full-vector mass error V is in Tables 15–16. All 30 narrow collapse statistics are zero; this does not establish healthy recall.

Generator/encoder	Method	Seed	Step	State	C	$\widehat{q}_9$
Small/ResNet-18	Dense	101	5k	EMA	0.8	0.0128
Small/ResNet-18	DriftXpress	101	5k	EMA	0.7	0.0147
Small/ResNet-18	Dense	202	5k	EMA	0.7	0.0043
Small/ResNet-18	DriftXpress	202	5k	EMA	0.7	0.0063
Small/ResNet-18	Dense	303	5k	EMA	0.8	0.0282
Small/ResNet-18	DriftXpress	303	5k	EMA	0.8	0.0303
Small/ResNet-18	Dense	101	10k	EMA	0.9	0.0640
Small/ResNet-18	DriftXpress	101	10k	EMA	1.0	0.0756
Small/ResNet-18	Dense	202	10k	EMA	1.0	0.0813
Small/ResNet-18	DriftXpress	202	10k	EMA	0.9	0.0965
Small/ResNet-18	Dense	303	10k	EMA	1.0	0.1430
Small/ResNet-18	DriftXpress	303	10k	EMA	1.0	0.1487
Small/ResNet-18	Dense	101	15k	EMA	1.0	0.0956
Small/ResNet-18	DriftXpress	101	15k	EMA	1.0	0.1011
Small/ResNet-18	Dense	202	15k	EMA	1.0	0.1347
Small/ResNet-18	DriftXpress	202	15k	EMA	1.0	0.1419
Small/ResNet-18	Dense	303	15k	EMA	1.0	0.1540
Small/ResNet-18	DriftXpress	303	15k	EMA	1.0	0.1539
U-Net/DINOv3	Dense	101	15k	EMA	0.9	0.2042
U-Net/DINOv3	Dense	101	15k	Raw	0.9	0.1908
U-Net/DINOv3	DriftXpress	101	15k	EMA	0.7	0.0025
U-Net/DINOv3	DriftXpress	101	15k	Raw	0.9	0.0009
U-Net/DINOv3	Dense	202	15k	EMA	1.0	0.1068
U-Net/DINOv3	Dense	202	15k	Raw	1.0	0.1246
U-Net/DINOv3	DriftXpress	202	15k	EMA	1.0	0.1071
U-Net/DINOv3	DriftXpress	202	15k	Raw	1.0	0.1026
U-Net/DINOv3	Dense	303	15k	EMA	0.9	0.0899
U-Net/DINOv3	Dense	303	15k	Raw	1.0	0.1777
U-Net/DINOv3	DriftXpress	303	15k	EMA	1.0	0.1382
U-Net/DINOv3	DriftXpress	303	15k	Raw	1.0	0.1388

G.5 Identical-image Inception input-path sensitivity

One controlled intervention, not replacement endpoints. We reevaluate all 30 checkpoint/state image sets from Appendix G.4 with the historical Inception internal transform off and with it on: 60 complete metric records. The 18 bridge endpoints are evaluated first, followed by all 12 U-Net raw/EMA endpoints. The matrix, input-path order, sample streams, and metric definitions were fixed before the new outcomes; no training, state selection, optimizer step, or EMA update occurs. This test reuses the same selected configurations and is not an independent discovery sample. It adds 1,502 seconds of single-A100 allocation (0.4172 GPU-hours), kept separate from the training-cost axis.

Each checkpoint is loaded read-only and its saved state authenticated. Generate its 10,000 images once using the exact labels/noise and numerical conventions above; both feature paths consume this same tensor and the same ordered 5,000 real images. The only model change is `transform_input=False` to `True`; pretrained weights, external standardization, resize/crop, precision, batching, and metric routines remain fixed. Real-reference feature caches are keyed separately by input path, weight identity, preprocessing, and ordered reference identity. Both real and generated features therefore use the corresponding input path. Neither result is asserted to be canonical TensorFlow FID or to match another spatial preprocessing convention.

A forward pre-hook at the first Inception convolution checks the actual inputs for a fixed two-image fixture in the evaluation runtime. The off/on/off sequence gives $(u - \mu)/\sigma$, $2u - 1$, and the original

input again, with unchanged parameters and buffers. The maximum absolute difference from $2u - 1$ is $1.20 \cdot 10^{-7}$; the off-path round trip is exact. This validates the executed intervention, not the scientific quality of the generated images.

All 30 historical-input reevaluations reproduce all 330 preceding metric values exactly; generated-image, conditioning-label, reference-image, and historical feature identities agree. The transformed generated and reference feature identities differ for every endpoint. Classifier predictions, all five classifier-based mass/coverage diagnostics, and the recorded training-cost coordinates are exactly invariant across paths. No new mass estimate or training time is inferred from the changed features. Table 18 retains every transform-on outcome alongside the historical tables; no favorable input path replaces an unfavorable one.

A surviving final-point benefit, but not a robust whole-trajectory ordering. At 15k, both paths give lower FID and higher recall for DriftXpress in all three bridge pairs at the same 26.7% reduction in recorded training-component cost (Table 1). With the internal transform on, paired FID differences are $[-3.6591, -1.1140, -3.1935]$ and recall differences are $[0.0130, 0.0026, 0.0030]$ for seeds 101/202/303. Mean FID is 67.1832/64.5277 and recall 0.01527/0.02147 for Dense/DriftXpress. All three transformed final spread differences are positive, with only a 0.443% mean increase; one KID difference and two precision differences remain unfavorable. A wider feature cloud alone is not better semantic coverage.

At 10k, the FID differences remain favorable in all three pairs: $[-2.1339, -2.5170, -1.5737]$. Recall differences instead become $[-0.0002, -0.0030, +0.0024]$, versus all-positive historical values. Consequently the mean recall ordering reverses (Dense 0.003267; DriftXpress 0.003000). This prevents an input-robust claim of jointly better FID and recall throughout the later trajectory, even though the final-point joint benefit survives. At 5k, all transformed recalls are zero, and the adverse seed-303 FID difference remains adverse ($+0.5142$). The three points are repeated observations of six models, not extra independent seeds or exact time-to-quality crossings.

High-capacity metric ranks can change without recovered coverage. Under the internal transform, U-Net/DINOv3 mean EMA FID is 182.37/176.96 for Dense/DriftXpress, reversing the historical mean ordering of 154.37/165.65. At the paired-seed level, however, the EMA FID signs remain unchanged: DriftXpress is better in only one of three seeds on each path. The mean reversal reflects changed magnitudes, not a replicated all-seed gain. Raw weights improve FID over EMA in five of six cases under the transformed input, versus four historically. Nevertheless all six EMA recalls remain zero on both paths, and five raw recalls remain zero; the same DriftXpress seed-202 raw state has recall 0.0002. Neither this input change nor substituting raw weights establishes a functioning high-capacity reference.

Table 18: All 30 transform-on outcomes on identical checkpoint/state image sets. Paired effects use full-precision values before display rounding. S/R: small generator/ResNet-18; U/D: U-Net/DINOv3; E: EMA parameters. F, K, P, R, D_g are $\mathrm{FID}_{\mathrm{tv}}$, $\mathrm{KID}_{\mathrm{tv}}$, precision, recall, and feature spread. All narrow collapse values are zero. Classifier diagnostics and recorded training costs are exactly unchanged from Tables 15–17; the corresponding historical-input quality scores remain in Tables 15 and 16.

Model	Method	Seed	Step/state	$F \downarrow$	$K \downarrow$	$P \uparrow$	$R \uparrow$	D_g
S/R	Dense	101	5k/E	222.08	0.24266	0.6920	0.0000	11.751
S/R	Dense	101	10k/E	104.81	0.09540	0.7650	0.0030	15.645
S/R	Dense	101	15k/E	68.54	0.05153	0.7061	0.0194	17.121
S/R	Dense	202	5k/E	239.18	0.26529	0.6493	0.0000	11.337
S/R	Dense	202	10k/E	104.20	0.08933	0.8050	0.0058	15.485
S/R	Dense	202	15k/E	61.82	0.03806	0.7316	0.0200	17.158
S/R	Dense	303	5k/E	234.76	0.26082	0.6362	0.0000	11.226
S/R	Dense	303	10k/E	110.59	0.09671	0.8578	0.0010	14.506
S/R	Dense	303	15k/E	71.19	0.05006	0.8111	0.0064	16.151
S/R	DriftXpress	101	5k/E	221.40	0.24242	0.6773	0.0000	11.768
S/R	DriftXpress	101	10k/E	102.68	0.09314	0.7590	0.0028	15.698
S/R	DriftXpress	101	15k/E	64.88	0.04853	0.7110	0.0324	17.132

Model	Method	Seed	Step/state	$F \downarrow$	$K \downarrow$	$P \uparrow$	$R \uparrow$	D_g
S/R	DriftXpress	202	5k/E	236.92	0.26378	0.6438	0.0000	11.475
S/R	DriftXpress	202	10k/E	101.68	0.08699	0.7903	0.0028	15.634
S/R	DriftXpress	202	15k/E	60.70	0.03814	0.7257	0.0226	17.184
S/R	DriftXpress	303	5k/E	235.27	0.26312	0.6457	0.0000	11.202
S/R	DriftXpress	303	10k/E	109.01	0.09720	0.8542	0.0034	14.592
S/R	DriftXpress	303	15k/E	68.00	0.04872	0.7966	0.0094	16.337
U/D	Dense	101	15k/E	183.22	0.16449	0.7793	0.0000	12.587
U/D	Dense	101	15k/raw	157.20	0.11012	0.6086	0.0000	14.448
U/D	Dense	202	15k/E	153.06	0.11857	0.8156	0.0000	14.215
U/D	Dense	202	15k/raw	108.58	0.06719	0.7925	0.0000	16.149
U/D	Dense	303	15k/E	210.85	0.19237	0.7308	0.0000	11.220
U/D	Dense	303	15k/raw	199.08	0.16705	0.7429	0.0000	11.764
U/D	DriftXpress	101	15k/E	221.57	0.17836	0.8153	0.0000	10.970
U/D	DriftXpress	101	15k/raw	230.68	0.19000	0.7160	0.0000	11.679
U/D	DriftXpress	202	15k/E	154.16	0.11966	0.8168	0.0000	13.723
U/D	DriftXpress	202	15k/raw	130.86	0.09905	0.6492	0.0002	15.307
U/D	DriftXpress	303	15k/E	155.15	0.12039	0.8366	0.0000	13.729
U/D	DriftXpress	303	15k/raw	130.08	0.09462	0.6580	0.0000	15.499

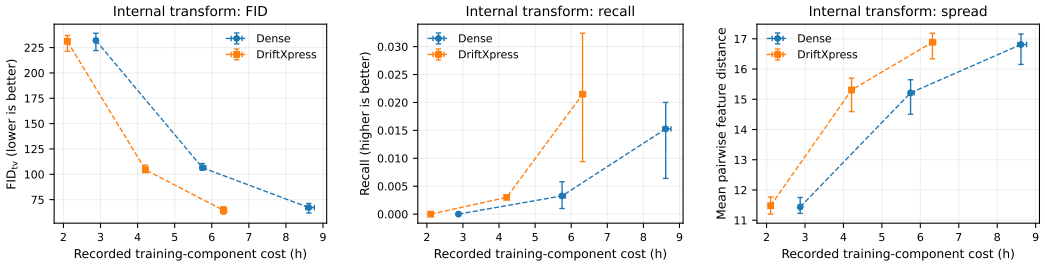


Figure 12: **The same bridge images with the Inception internal transform enabled.** Means and seed ranges at 5k/10k/15k; compare the historical-input curves in Figure 11. Cost is unchanged and excludes setup, checkpointing, and evaluation. The 15k paired FID/recall advantage survives, while 10k recall gains do not. Segments connect measured points, not observed intermediate outcomes; ranges are not confidence intervals.

Fixed visual evidence and remaining scope. Figure 13 shows all six final bridge models, using the first eight original stream indices for each requested conditioning class. The selection uses neither a quality score nor predicted class. The stored float32 pixels are converted only to rounded uint8 for display, without enhancement or per-image normalization; metric computation uses the original image tensors. Requested labels are not verified semantic classes. These grids are visual context, not a human-rating study or evidence of high absolute generative quality.

The intervention closes the specific untested-input question for these 30 endpoints. It does not validate all 66 final models, the older raw-image screen, another feature network, or semantic class correctness. FID, KID, manifold metrics, and spread still share Inception features; their agreement cannot be counted as independent corroboration. The information-access and normalization confounds in the training comparison remain. The supported result is thus a final-checkpoint cost–quality benefit that survives this input toggle, with explicitly input-sensitive earlier recall and high-capacity mean FID ordering.

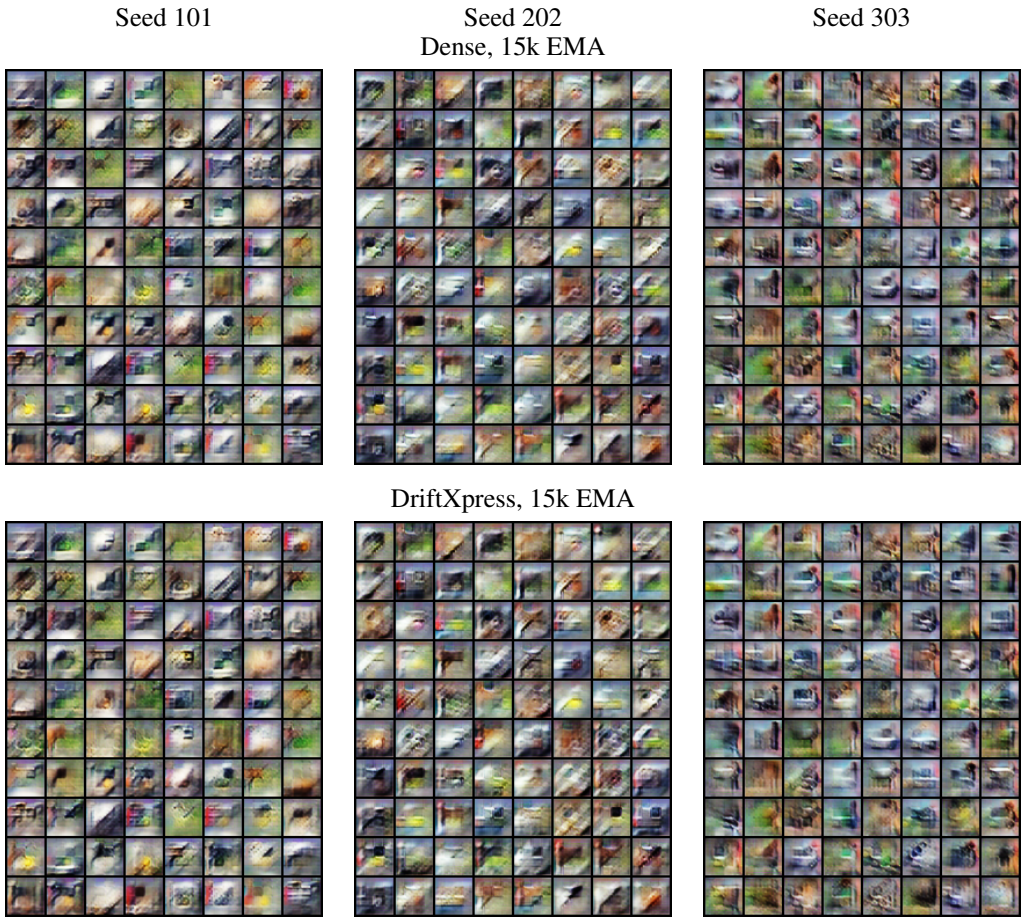


Figure 13: **Fixed final bridge samples, not quality-selected, from all three paired seeds.** Balanced CIFAR-10, small generator/ResNet-18. Within each grid, rows request classes 0–9 and columns show the first eight corresponding images in the fixed stream, not favorable classifier predictions. The 480 displayed samples are a deterministic subset of the already evaluated images. Labels/noise are paired between methods. Enabling the Inception transform does not change these pixels. Limited fidelity and conditional ambiguity remain visible; class requests are not semantic annotations.

G.6 A RETROSPECTIVE EARLY-TO-FINAL COST CHECK

The first 100 updates were recorded before terminal results. For paired dense and candidate runs, let $T_{d,t}, T_{c,t}$ be cumulative training-component costs through update t, and $F_{d,t}, F_{c,t}$ their field costs, including construction. Without fitting to final outcomes, compare field substitution with the directly measured early total-cost ratio:

$$\widehat{S}_{\text{field}} = \frac{T_{d,100}}{T_{d,100} - F_{d,100} + F_{c,100}}, \qquad \widehat{S}_{\text{total}} = \frac{T_{d,100}}{T_{c,100}}. \tag{15}$$

The second baseline includes the candidate's measured non-field overhead. We evaluate both against the full-run ratio and the ratio for the remaining updates alone:

$$S_{\text{final}} = \frac{T_{d,15000}}{T_{c,15000}}, \qquad S_{\text{rem}} = \frac{T_{d,15000} - T_{d,100}}{T_{c,15000} - T_{c,100}}. \tag{16}$$

All 45 candidate–dense pairs are retained (21 DriftXpress, 12 Fast, 12 Quality), sharing 21 distinct dense controls. Paired runs match dataset, target regime, seed, generator, and encoder. Early records are checked against the corresponding final training records; evaluation-only recoveries do not change training timers. Table 19 reports the median relative error $100|\widehat{S}/S - 1|$. Field substitution retains its 0.98% full-run median (maximum 4.15%) and beats a constant-one prediction in 36/45 pairs by absolute log error. However, the early total-cost baseline is more accurate overall: 0.79% median, 3.08% maximum, and smaller absolute log error in 26/45 pairs. This is not uniform across methods: field substitution has the lower median for Quality.

Table 19: Retrospective cost-prediction controls. Median relative errors (%) use every paired comparison; lower is better. "Field" substitutes only the field cost, "Total" uses the measured first-100-update total-cost ratio, and "One" predicts no speedup. Remaining-update targets exclude the first 100 updates. Shared dense controls preclude treating the 45 pairs as independent replicates.

		Full 15,000 updates			Remaining 14,900 updates		
Candidate	Pairs	Field	Total	One	Field	Total	One
DriftXpress	21	0.789	0.722	5.900	0.780	0.727	5.901
Fast	12	1.476	0.583	1.742	1.472	0.587	1.752
Quality	12	1.511	1.574	2.159	1.519	1.584	2.173
All	45	0.981	0.789	3.766	0.998	0.794	3.763

Removing the first 100 updates from the target yields median errors of 1.00% and 0.79%, respectively, with the early total-cost ratio again better in 26/45 pairs. This removes direct early-period overlap, not retrospective design or shared-control dependence. The analyses are neither preregistered prospective prediction tests nor independent holdout evaluations. They forecast fixed-update execution, not quality or time to a common quality target; the data do not show that field substitution improves on the stronger overall timing baseline.

G.7 COMPLETE MAIN MATRIX AND CLAIM BOUNDARY

Table 20 summarizes the entire 48-run U-Net/DINOv3 matrix under the historical input path. For rare targets, Fast mean prior-matched FID_{tv} is 263.75 versus 171.33 for dense on CIFAR-10, and 245.95 versus 142.88 on SVHN; mean training savings are below 1%. Recovered Quality means are 289.71 and 226.03, respectively. These seed-variable failures do not establish universal inferiority. On balanced SVHN, Quality improves secondary FID in all three paired seeds (mean 126.89 versus 140.88), but mean recall is only 0.00033. DriftXpress's main-matrix quality signs are mixed. A favorable FID difference is not sufficient evidence of useful support coverage.

Paired differences, means, and ranges are computed from full-precision scores before display rounding; subtracting printed component scores can therefore differ by 0.01. Three paired seeds support descriptive sensitivity, not a new confirmatory significance or non-inferiority test. The subsequent saved-checkpoint study supplies quality/diversity versus recorded training-component cost for one complete six-model bridge cell (Appendix G.4), not full-allocation curves for all 66 models or exact quality-crossing times. No exact separate-row training arm isolates projection from native-field

Table 20: Complete U-Net/DINOv3 matrix: three-seed means; quality scores use the historical input path. F is prior-matched $\mathrm{FID}_{\mathrm{tv}}$ (lower is better), R is secondary recall, and V is classifier-estimated total-variation mass error. Training hours exclude evaluation; allocation includes the original job and successful recovery, but excludes separate recovery diagnostics. All methods/regimes are retained; selected input sensitivity: Appendix G.5.

Dataset	Target	Method	$F\downarrow$	$R\uparrow$	$V\downarrow$	Train (h)	Alloc. (h)
CIFAR-10	Bal.	Dense	154.37	0.00000	0.536	44.92	45.84
CIFAR-10	Bal.	DriftXpress	165.65	0.00000	0.470	42.87	43.79
CIFAR-10	Bal.	Fast	165.09	0.00000	0.583	43.88	44.81
CIFAR-10	Bal.	Quality	161.45	0.00007	0.663	43.80	44.73
CIFAR-10	Rare	Dense	171.33	0.00000	0.668	45.85	46.80
CIFAR-10	Rare	DriftXpress	160.16	0.00007	0.508	43.45	44.37
CIFAR-10	Rare	Fast	263.75	0.00000	0.738	45.51	46.44
CIFAR-10	Rare	Quality	289.71	0.00000	0.817	44.68	61.39
SVHN	Bal.	Dense	140.88	0.00000	0.307	45.12	46.04
SVHN	Bal.	DriftXpress	131.03	0.00000	0.366	42.72	43.63
SVHN	Bal.	Fast	135.01	0.00000	0.342	43.20	44.13
SVHN	Bal.	Quality	126.89	0.00033	0.394	43.75	44.69
SVHN	Rare	Dense	142.88	0.00000	0.413	45.24	46.18
SVHN	Rare	DriftXpress	135.47	0.00000	0.338	42.97	43.88
SVHN	Rare	Fast	245.95	0.00000	0.623	44.81	45.76
SVHN	Rare	Quality	226.03	0.00000	0.639	44.35	61.47

changes. These results support a conditional cost boundary, rather than improved high-capacity generation. Poor high-capacity diversity and unfavorable outcomes remain part of the scientific evidence.

Table 21: What each observed signal does—and does not—establish in this study. These are evidence distinctions, not a trained selection rule.

Observed signal	Supported interpretation	Remaining question
Small compatible field error	Approximation to the declared native operator	Same generator update and useful learned distribution?
41–54× field speedup in the bridge	A cheaper component across four pipelines	Training speedup ranges from 1.048 to 1.365×; terminal quality need not improve.
At 15k, lower FID and higher recall at lower training-component cost on both input paths; the joint 10k gain is historical-input only	A paired final-checkpoint benefit for the selected small-generator/ResNet-18 configuration that survives the tested input toggle	Exact crossing times, full allocation cost, other configurations, and isolated projection effects remain unestablished.
Identical-image input toggle changes 10k recall and high-capacity mean FID ordering	Evaluator-sensitive comparisons, without changing generated images or restoring high-capacity coverage	Robustness to another feature network or semantic assessment remains untested.
Zero rare undercoverage	No deficit under that one-sided classifier-based metric	Calibrated mass? Every rare raw row has estimated excess.
Stable feature Pareto membership	A conditional quality–cost operating point	Does its role survive the raw pipeline? Only one of four does stably.

G.8 Teaser legend and selected metrics for every final model

Figure 1. (a) Generated particles $x \sim q_t$ move toward target p; representatives and shaded supports depict coarsening, not measured uncertainty or generated images. **(b)** Noise z and class y enter G_θ, followed by frozen encoder ϕ; field v defines a detached target and $\nabla_\theta \mathcal{L}$ updates the generator. Preparation is omitted and diagram sizes are not times. **(c)** Under the historical input path, prior-matched $\mathrm{FID}_{\mathrm{tv}}$ is the Fréchet distance between Gaussian fits to frozen torchvision Inception-v3 features (lower is better), not a percentage or DINOv3 score. The six rare-target Dense/Fast seed pairs use 9,100 generated and 4,550 real images per evaluation, with target $p_9 = 1/91$. Lines join paired seeds; dashed diamonds are arithmetic means, not intervals or trajectories. This panel presents a complete two-method comparison, not the complete method ranking. The metric checks and low-recall limitations above apply; no unseen time-to-quality trajectory is implied.

Table 22: Historical-input endpoint metrics for all 66 distinct final models, 15,000 updates each. "C10" is CIFAR-10; $B_0/B_1/B_2$ are the additional bridge cells, and Main is U-Net/DINOv3. F_b, F_m denote balanced-reference and prior-matched $\mathrm{FID}_{\mathrm{tv}}$; R_p, R_s are recall fractions. Zero means an observed zero, not missingness. † marks evaluation-only recovery from the original final checkpoint. Training costs are unchanged by recovery; selected input sensitivity: Appendix G.5.

Pipeline/data	Target	Method	Seed	F_b	F_m	R_p	R_s	Train (h)
B_0/C10	Bal.	DriftXpress	101	63.93	63.65	0.0306	0.0422	6.30
B_0/C10	Bal.	DriftXpress	202	59.75	59.81	0.0372	0.033	6.24
B_0/C10	Bal.	DriftXpress	303	67.02	66.85	0.0138	0.019	6.40
B_0/C10	Bal.	Dense	101	67.59	67.32	0.0276	0.0364	8.53
B_0/C10	Bal.	Dense	202	60.84	60.81	0.029	0.0278	8.56
B_0/C10	Bal.	Dense	303	70.34	70.27	0.014	0.0074	8.76
B_1/C10	Bal.	DriftXpress	101	58.89	58.89	0.008	0.0122	17.67
B_1/C10	Bal.	DriftXpress	202	58.65	58.87	0.0062	0.0096	17.94
B_1/C10	Bal.	DriftXpress	303	64.07	64.42	0.0092	0.008	17.68
B_1/C10	Bal.	Dense	101	64.36	65.22	0.0098	0.0074	19.77
B_1/C10	Bal.	Dense	202	66.59	67.56	0.01	0.0076	19.73
B_1/C10	Bal.	Dense	303	65.16	65.09	0.011	0.0104	20.45
B_2/C10	Bal.	DriftXpress	101	98.66	99.37	0.0012	0.0004	32.18
B_2/C10	Bal.	DriftXpress	202	67.15	67.35	0.0168	0.017	31.80

Continued on next page.

Table 22 continued.

Pipeline/data	Target	Method	Seed	F_b	F_m	R_p	R_s	Train (h)
B_2/C10	Bal.	DriftXpress	303	73.01	74.78	0.0174	0.0174	31.61
B_2/C10	Bal.	Dense	101	70.23	70.48	0.011	0.0194	33.76
B_2/C10	Bal.	Dense	202	71.03	70.65	0.0134	0.017	33.96
B_2/C10	Bal.	Dense	303	68.19	67.72	0.0168	0.0122	33.91
Main/C10	Bal.	DriftXpress	101	222.46	221.83	0	0	42.71
Main/C10	Bal.	DriftXpress	202	140.49	141.04	0	0	43.13
Main/C10	Bal.	DriftXpress	303	132.90	134.08	0	0	42.79
Main/C10	Bal.	Dense	101	161.01	161.26	0	0	44.24
Main/C10	Bal.	Dense	202	130.13	130.11	0	0	45.14
Main/C10	Bal.	Dense	303	171.38	171.73	0	0	45.40
Main/C10	Bal.	Fast	101	199.93	200.97	0	0	43.60
Main/C10	Bal.	Fast	202	120.41	119.65	0	0	44.15
Main/C10	Bal.	Fast	303	174.83	174.65	0	0	43.88
Main/C10	Bal.	Quality	101	175.84	176.58	0	0	43.24
Main/C10	Bal.	Quality	202	116.15	115.40	0.0002	0.0002	44.46
Main/C10	Bal.	Quality	303	192.72	192.38	0	0	43.69
Main/C10	Rare	DriftXpress	101	189.25	190.68	0	0	43.41
Main/C10	Rare	DriftXpress	202	137.48	136.97	0.0008	0.00022	43.70
Main/C10	Rare	DriftXpress	303	153.30	152.82	0	0	43.23
Main/C10	Rare	Dense	101	169.17	169.05	0	0	45.75
Main/C10	Rare	Dense	202	137.92	135.70	0.0002	0	45.85
Main/C10	Rare	Dense	303	205.71	209.23	0	0	45.95
Main/C10	Rare	Fast	101	315.57	316.23	0	0	46.04
Main/C10	Rare	Fast	202	140.83	140.45	0	0	44.92
Main/C10	Rare	Fast	303	329.89	334.56	0	0	45.57
Main/C10	Rare	Quality†	101	421.62	427.79	0	0	44.55
Main/C10	Rare	Quality†	202	138.35	138.39	0	0	44.90
Main/C10	Rare	Quality†	303	301.27	302.95	0	0	44.59
Main/SVHN	Bal.	DriftXpress	101	148.20	147.57	0	0	42.80
Main/SVHN	Bal.	DriftXpress	202	138.34	138.48	0	0	42.20
Main/SVHN	Bal.	DriftXpress	303	108.83	107.05	0.0002	0	43.16
Main/SVHN	Bal.	Dense	101	126.44	126.50	0	0	45.55
Main/SVHN	Bal.	Dense	202	146.52	146.46	0	0	44.30
Main/SVHN	Bal.	Dense	303	149.49	149.69	0	0	45.51
Main/SVHN	Bal.	Fast	101	157.91	158.84	0	0	42.95
Main/SVHN	Bal.	Fast	202	119.92	120.45	0	0	42.89
Main/SVHN	Bal.	Fast	303	125.13	125.73	0	0	43.76
Main/SVHN	Bal.	Quality	101	116.49	116.38	0	0	43.02
Main/SVHN	Bal.	Quality	202	138.58	138.22	0	0	43.38
Main/SVHN	Bal.	Quality	303	127.38	126.06	0.0006	0.001	44.86
Main/SVHN	Rare	DriftXpress	101	142.32	141.60	0	0	43.30
Main/SVHN	Rare	DriftXpress	202	161.63	160.41	0	0	43.48
Main/SVHN	Rare	DriftXpress	303	105.05	104.39	0	0	42.14
Main/SVHN	Rare	Dense	101	150.17	147.96	0	0	44.58
Main/SVHN	Rare	Dense	202	147.24	144.25	0	0	45.55
Main/SVHN	Rare	Dense	303	138.12	136.42	0	0	45.59
Main/SVHN	Rare	Fast	101	121.51	121.08	0	0	44.03
Main/SVHN	Rare	Fast	202	253.68	252.28	0	0	44.97
Main/SVHN	Rare	Fast	303	364.96	364.50	0	0	45.44
Main/SVHN	Rare	Quality†	101	105.49	104.61	0	0	44.09
Main/SVHN	Rare	Quality†	202	251.60	250.97	0	0	44.12
Main/SVHN	Rare	Quality†	303	323.70	322.51	0	0	44.84

H Executable Contracts and Evaluator Details

This appendix describes the image-study implementations and the changes made for the 66-model extension. Source-level descriptions, recorded numerical checks, and unperformed controls are distinguished throughout.

H.1 EXTENSION FIELD, SUPPORT, AND UPDATE CONTRACT

The extension uses $k(x, y) = \exp(-\|x - y\|_2/0.5)$ on sample-normalized features, drift scale $\eta = 1$, and no field RMS rescaling. Dense uses the joint operator with aligned generated self deletion; its exact blocked evaluation uses query/source blocks 16/64, replacing the earlier raw screen's 64/128. "Optimized dense" denotes this implementation, not a hardware-wide optimum. The affine/normalized-affine microbenchmark in Appendix C.3 does not establish optimality of this different image workload; no block-size sensitivity comparison on the extension hardware is reported. A 512-query update uses all 512 current generated sources and a fresh 512-sample target batch. Encoder chunking does not create independent eight-sample fields: the complete 512-query field is computed before the regression pass.

DriftXpress uses the separate-row operator and additive denominator behavior in Appendix F. It caches sums over the entire fixed target support, with class-sharded random landmarks, identity feature scaling, absolute bandwidth 0.5, ridge 10^{-4}, eigenvalue floor 10^{-8}, folded summaries, and summary chunks of 512. The landmark budget is 20% of target count: 2,000 for 10,000 balanced targets and 1,820 for 9,100 rare targets. Equal per-class allocation is capped by class availability, then remaining slots are redistributed. Table 23 separates the measured configurations from the unperformed controls.

Table 23: Normalization, positive-source access, and approximation in the extension. The two exact separate-row controls are *not measured training arms*. All rows specify the same current generated support and aligned self deletion; the difference in native normalization remains explicit.

Configuration	Normalization	Positive-source access	Training evidence
Dense	Joint	Fresh 512-sample batch	Reported
Exact separate-row / batch	Separate-row	Same fresh batch	Not run
Exact separate-row / full	Separate-row	Full fixed target	Not run
DriftXpress	Separate-row	Projected full-target cache	Reported

What conditioning does and does not do. Class labels condition the image generator and determine target-prior sampling. They also guide landmark sharding and representative quotas. They do *not* impose a same-class query–source affinity mask: the deployed field acts on the marginal feature distribution, summing attraction from all positive shards. Thus matching the distribution of requested conditioning labels does not enforce correct semantic class assignment or a diagonal request–prediction table. Representative selection is run with an evaluation-mode image-generator view with an isolated pseudorandom-number stream; this auxiliary view does not replace the training-mode generated-query features used in the two regression passes.

Fast and Quality: inheritance and realized schedules. The representative budgets, radial/class quotas, importance weighting, rare latent safeguard, candidate distribution, and probe thresholds are those in Appendix F; the horizon is changed to 15,000 updates. Fast uses 256 real/384 generated representatives in the balanced branch and rare budgets $(m_+, m_-) = (128, 128)$, with 16 latent replacement points within the 128-point negative budget. The fixed target candidate pool contains 4,096 examples and is not periodically rebuilt. Selection and construction remain charged.

Quality's polish setting is 5% *clamped to at most 200 updates*, not 750 updates at this horizon. Appendix F.3 specifies slot positions, block activation, and polish precedence. The 25-step gap constrains activation triggers rather than individual teacher calls: all six balanced runs used 720 teacher slots, 14,080 geometry updates, and 200 final dense-polish updates, totaling 15,000. Their total dense fraction is $920/15000 = 6.133\%$, not the nominal 18% cap. The six rare runs used 7,499 rare-support updates, 7,301 refinement updates, no probe-triggered teacher updates, and 200 polish updates. Rare probe teachers are disabled by the retained regime rule; the nominal 400-step cap is dormant. These are realized schedule counts, not new adaptive-policy tuning or evidence that the teacher mechanism improved quality.

Regression normalization and mixed precision. At update t, generate $X^{(1)} = h(g_\theta(z, y))$ without retaining its graph, compute the complete field V, and detach $Y = X^{(1)} + V$. Recompute $X^{(2)}$

differentiably using the same parameters, labels, noise, per-sample indices, and eight-sample chunk schedule. For $n = 512$ and feature dimension d, the regression is

$$L = \frac{1}{nd}\|X^{(2)} - Y\|_F^2.$$

Each chunk's elementwise mean squared loss is multiplied by its sample count divided by 512 before gradient accumulation. Thus Equation (5) uses $a = 2/(nd)$ and $\eta = 1$; summing unweighted chunk means would change the update. Parameters, loss, field, optimizer state and EMA remain float32; generator/encoder forwards use bfloat16 autocasting, with no GradScaler and with TF32 disabled. The encoder is frozen and in evaluation mode. The image gradient is first obtained through the encoder and then propagated through the generator; optimizer clipping and AdamW occur only after all chunks. Initial-state equality across paired methods does not imply trajectory equality once fields differ. ResNet-18 has $d = 512$, whereas DINOv3 has $d = 12{,}288$: the explicit $1/d$ prefactor is therefore 24 times smaller for DINOv3. This is not a claim of 24-fold smaller gradients or AdamW updates, since features, field magnitudes, Jacobians, clipping and optimizer state also matter.

H.2 Same-state recomputation and structured mask replay

The U-Net uses sample-indexed deterministic structured masking at nominal drop rate 0.1, not independent Bernoulli dropout. Masks are *not* independently resampled between detached and differentiable passes. Each residual block assigns a fixed integer stream index ℓ; its mask is a deterministic function of the within-sample flattened activation index q, global batch index i, and update t. With $M = 2{,}147{,}483{,}647$, the retained integer construction is

$$\begin{aligned} u_0 &= (48271q + 69621i + 104729t + 130363\ell + 15485863) \bmod M, \\ u_1 &= (48271u_0 + 1) \bmod M, \qquad u_2 = (69621u_1 + 1) \bmod M, \\ \mathrm{drop}(x) &= x\,\mathbf{1}\{u_2 \geq \mathrm{round}(0.1M)\}/0.9. \end{aligned}$$

The production caller supplies the same i, t to both passes; chunk slicing preserves these indices. The small residual generator has no stochastic dropout. This specifies mask replay, not independence of all mask bits or a proof of production floating-point equality. The generic dropout fallback used when indices and update are both absent is not the production training call.

The modular construction remains affine in q: adjacent indices add $94{,}793{,}834 \bmod M$ to u_2. A CPU check of the retained int64 masking module on $q = 0, \ldots, 499{,}999$, with $i = 7$, $t = 100$, and $\ell = 3$, gives drop fraction 0.1000 and adjacent keep-bit correlation 0.50956. The module agrees exactly with the printed formula and replays under sample-batch chunking in this check. These observations quantify mask structure; they do not constitute a production full-model gradient/update check, an independent-dropout ablation, or evidence that masking caused poor generation. Changing the masking rule would define a new experiment, not a reproduction of these runs.

If the two feature realizations differ, the actual gradient is instead

$$\nabla_\theta L = a(J_\theta^{(2)})^\top \mathrm{vec}\Big(X^{(2)} - X^{(1)} - \eta V\Big).$$

The additional residual cannot be discarded merely because the target is detached. Matching indexed masks rules out independent-mask resampling as the specified mechanism, but does not identify the cause of poor generation. The recorded bfloat16-versus-float32 forward check and dense evaluator repeatability check are different controls; neither is a saved same-state full-model gradient-parity measurement. No new production-CUDA parity or training run is claimed here.

H.3 Frozen feature weights and preprocessing

The extension and its successful evaluation recoveries used Python 3.12.3, PyTorch `2.9.0a0+145a3a7bda.nv25.10` with CUDA 13.0, and torchvision `0.24.0a0+094e7af5`, on one A100-SXM4-40GB per run. This is distinct from the local software environment in Appendix B.

The training ResNet-18 uses torchvision's `ResNet18_Weights.IMAGENET1K_V1` release (checkpoint variant `f37072fd`). Raw images are resized to 224×224 and normalized by ImageNet means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$; its 512-dimensional

pooled outputs are sample-normalized. The DINOv3 encoder is the released ViT-B/16 LVD-1689M variant `73cec8be`. Its 112×112, antialiased-resize and $4 \times 4 \times 768$ patch-pooling convention is given in Appendix G. Replacing either preprocessing path or spatial pooling changes this experiment.

The independent quality encoder is torchvision Inception-v3 with ImageNet-1K V1 weights (release variant `0cc3c7bd`), in evaluation mode, with the final classifier replaced by identity. It emits the 2,048-dimensional pooled vector, without sample ℓ_2 normalization. The retained implementation has `transform_input=False`: unit-range RGB is bilinearly resized to 342×342 with `align_corners=False` and no antialiasing, cropped to 299×299 starting at pixel $(21, 21)$, and normalized by the ImageNet statistics above. This is the executed preprocessing, not a claim of equivalence to other FID packages or their resizing conventions (Parmar et al., 2022). Evaluation extracts features in chunks of 32.

Internal-input transform and scope of validation. The evaluator constructs the architecture with `weights=None`, explicitly sets `transform_input=False`, and then loads the pretrained state dictionary. In the cited torchvision revision, the pretrained-weight factory instead enables the internal transform unless overridden (TorchVision maintainers, n.d.). For resized/cropped unit-range pixels u_c, external standardization gives $z_c = (u_c - \mu_c)/\sigma_c$. The two first-convolution inputs are

$$x_c^{\text{retained}} = z_c, \qquad x_c^{\text{internal}} = z_c \frac{\sigma_c}{0.5} + \frac{\mu_c - 0.5}{0.5} = 2u_c - 1.$$

Thus this is a defined custom feature path, not the pretrained factory's default input path. FID, KID, manifold precision/recall, and spread all use its features; classifier-based mass does not. Appendix G.5 reports the completed identical-image comparison for all 30 selected checkpoint/state endpoints. A first-convolution hook in the evaluation job verifies both formulas on a fixed fixture and exact restoration of the historical path, with unchanged parameter/buffer identities. Real and generated features are recomputed with the same flag and path-specific reference caches. The 15k bridge FID/recall signs survive; the 10k recall signs and the high-capacity mean EMA FID ordering do not. This is a controlled input-path sensitivity test, not an independent evaluator or a causal explanation of poor generation. Spatial resizing/cropping is unchanged, so enabling the transform does not establish equivalence to another FID package. All historical scores keep their original definition.

Statistic definitions and sample counts. For feature sets of sizes n, m, FID uses float64 empirical means and unbiased sample covariances (denominators $n - 1, m - 1$). The extension computes the covariance cross-term by thin QR followed by a singular-value sum, rather than a sample-size-square matrix square root; only the final FID is clamped below at zero. KID is the unbiased squared MMD with $k(u, v) = (u^\top v/2048 + 1)^3$, diagonal terms excluded. The extension evaluates all within- and cross-sample pairs in float64 blocks of 256; it does not average randomly chosen 1,000-sample KID subsets. In the original raw screen, KID and manifold metrics use 5,000 generated features, while FID uses 10,000. In the extension, every metric uses the full track sample counts: primary 10,000 generated/5,000 real; secondary 10,000/5,000 balanced or 9,100/4,550 rare.

Manifold precision is the fraction of generated features inside at least one real-feature ball; recall reverses these roles. Ball radii are third-nearest-neighbor *Euclidean* distances within each set, excluding only the aligned self entry, not all equal-valued duplicates. Membership uses $\leq$ with the same distance units, float32 distances and 256-row blocks. Extension diversity is mean distance over distinct unordered generated-feature pairs; its narrow collapse statistic is the fraction at distance at most 10^{-8}. A zero collapse statistic can therefore coexist with zero recall. Both extension real references use native test data, never the generator's training target. Secondary class counts and independent generation streams are fixed as described in Appendix G.

H.4 Class-mass evaluator identity and measured calibration

CIFAR-10 mass uses the independently pretrained ResNet-20 from `chenyaofo/pytorch-cifar-models` (pytorch-cifar-models contributors, n.d.), checkpoint variant `4118986f`, not the training ResNet-18 or Inception. Inputs remain unit-range 32×32 RGB and use means $(0.4914, 0.4822, 0.4465)$ and standard deviations $(0.2023, 0.1994, 0.2010)$, without test augmentation. SVHN mass uses a separately trained CIFAR-style ResNet-20: official 73,257-image training split only (no extra split), 80 epochs, batch 256, seed 12808, cross-entropy, SGD with initial learning rate 0.1, momentum 0.9, weight decay $5 \cdot 10^{-4}$, no Nesterov,

and epoch-wise cosine decay to zero. Augmentation is random crop with four-pixel constant padding, without horizontal flips. Means are $(0.4377, 0.4438, 0.4728)$ and standard deviations $(0.1980, 0.2010, 0.1970)$. The 26,032-image official test split was not used for checkpoint selection. Its recorded test accuracy is 96.6195%, negative log likelihood 0.15183, and 15-bin equal-width expected calibration error 0.01642; these are real-image checks, not generated-image error estimates.

Table 24 is a reanalysis of the already recorded predictions on the exact secondary real references, not a new classifier fit. Each reference/confusion matrix is identical across all applicable methods and seeds. Class-9 precision drops under the rare prior even when overall real-image accuracy remains high. No confusion correction, recalibration, or relabeling has been applied to any generated-image metric.

Table 24: Classifier performance on the exact prior-matched *real* references. Accuracy is in percent; class-9 recall and precision are fractions. The last two columns concern classifier class 9, not generative manifold recall/precision. True/predicted class-9 counts expose prior-dependent false positives. They do not estimate generated-image classification error.

Dataset	Prior	Images	Acc. (%)	True / pred. 9	Recall_9	Prec._9
CIFAR-10	Balanced	5,000	92.80	500 / 497	0.9540	0.9598
CIFAR-10	Rare	4,550	92.53	50 / 67	0.9400	0.7015
SVHN	Balanced	5,000	96.54	500 / 499	0.9560	0.9579
SVHN	Rare	4,550	96.64	50 / 69	0.9600	0.6957

Table 25: Complete rare-reference confusion counts: true classes are rows, predicted classes columns. Classes 0–8 each have 500 examples and class 9 has 50. These are real-reference, not generated request–prediction matrices.

CIFAR-10

True \ pred.	0	1	2	3	4	5	6	7	8	9
0	470	3	7	4	3	1	1	0	7	4
1	1	484	0	1	0	0	0	0	4	10
2	9	0	444	14	8	7	14	2	2	0
3	4	0	12	428	10	30	9	2	3	2
4	2	0	11	9	464	3	6	5	0	0
5	2	0	10	27	6	449	1	4	1	0
6	1	0	8	8	4	2	477	0	0	0
7	5	0	2	4	7	8	1	472	1	0
8	10	4	1	6	0	0	0	0	475	4
9	2	1	0	0	0	0	0	0	0	47

SVHN

True \ pred.	0	1	2	3	4	5	6	7	8	9
0	487	0	2	1	0	0	6	2	1	1
1	1	488	2	2	3	0	1	3	0	0
2	0	1	490	4	1	1	0	3	0	0
3	0	6	4	472	0	6	0	2	1	9
4	1	5	0	1	488	1	1	2	0	1
5	0	0	0	4	0	486	4	1	2	3
6	6	1	0	1	1	2	481	0	8	0
7	1	18	3	1	1	0	0	473	1	2
8	2	3	0	1	1	0	4	0	484	5
9	0	1	1	0	0	0	0	0	0	48

The requested generated-label counts and predicted-class marginals are both known, but their joint request–prediction contingency is not determined by those marginals. No generated-image semantic labeling study or such joint table is supplied here. In particular, real-reference class-9 precision is not an estimate of precision on these generated images. Reported class mass, absolute/signed rare error, and total variation remain classifier-based diagnostics, not identified semantic probabilities.

H.5 TIMING SENSITIVITY AND REMAINING CONTROLS

The historical raw "encoder pullback/backward" counter occupies about 69.1% of dense time and 71.2–71.3% for the balanced candidates. It bundles a second generator/encoder forward, backward work, and a gradient scan, and cannot be attributed entirely to encoder backward. In rare Fast runs, construction accounts for 458.10 of 474.79 field seconds; construction is already nested within field time and must not be added again. These historical component shares are distinct from the extension's non-overlapping training-component timers.

The historical balanced raw dense and DriftXpress inclusive means are approximately 6,213.69 and 6,136.8067 seconds. If their only correction is removal of between 0.60 and 0.71 seconds of overlap from each mean, the ratio lies between

$$\frac{6213.69 - 0.71}{6136.8067 - 0.60} \quad \text{and} \quad \frac{6213.69 - 0.60}{6136.8067 - 0.71},$$

which both round to $1.013\times$. This is an overlap-only sensitivity bound, not a reconstructed disjoint timer or a correction for bundled backward operations. Original costs and speed ratios are retained.

Table 26 defines the three distinct cost scopes. Cached-summary construction is charged when performed in the model's job, not amortized across an invented deployment count. Table 5 uses extension training-component time, not per-model allocation or time to matched quality.

Table 26: Cost scopes used in the image studies. None is a total for all project development. Historical numbers are preserved under their original denominator.

Scope	Included	Exclusions or boundary
Raw historical inclusive time	Preserved preparation, training, diagnostics and evaluation policy	Small checkpoint/evaluation overlap; backward counter bundles recomputation
Extension training-component time	Sum of disjoint update components; construction nested in field time	Setup outside updates, checkpointing and evaluation excluded
Extension per-model charged allocation	Original job's setup, training, checkpointing and evaluation, including original failure; successful evaluation-only recovery	Queueing, external data/model preparation, separate diagnostics and failed recovery attempts excluded

The saved-checkpoint evaluation in Appendix G.4 adds fixed-protocol quality/diversity versus recorded training-component cost for the six small-generator/ResNet-18 bridge models, and raw-versus-EMA evaluation for six U-Net/DINOv3 models. Appendix G.5 evaluates both Inception input paths on those same 30 image sets. These studies do not provide full-allocation quality curves for every configuration, production gradient/update parity, an information-matched exact separate-row training arm, or a functioning high-capacity reference. These remain empirical limitations; sparse checkpoints do not turn unobserved quality crossings into measured speedups.