MASKDD: SCULPTING DISCRIMINATIVE FEATURES FOR DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based dataset distillation uses pretrained generative models to construct compact training sets, but visual realism alone does not ensure that synthetic images are useful for classification. We introduce `MaskDD`, a framework that adapts the feature pathways of a frozen diffusion transformer through class-aware channel masks. Temperature-annealed sigmoid gates are learned with a denoising objective and regularizers for sparsity, binary polarization, and cross-class overlap. A rollback step selects a channel support, restores surviving mask logits to an earlier checkpoint, and removes the sparsity penalty for subsequent refinement. This separates support discovery from gate refinement without updating the pretrained backbone. Experiments on ImageNet subsets and ImageNet-1K show improvements in many evaluated settings. At 50 images per class, the reported ImageWoof accuracy with ResNet-18 is 59.0%, compared with 56.3% for MGD^3; on ImageNet-1K, the reported accuracy is 59.7%, compared with 58.6% for Minimax Diffusion. These results support class-aware masking as a useful adaptation mechanism for generative dataset distillation, while revealing sensitivity to the masking and refinement schedule.

1 INTRODUCTION

Dataset distillation aims to replace a large training set with a compact synthetic dataset that remains useful for training new models (Wang et al., 2018; Liu et al., 2022). Its value depends on downstream generalization: a distilled dataset should retain information needed to learn the task despite a small storage budget. Matching-based approaches optimize synthetic samples through gradients, feature distributions, or training trajectories (Zhao et al., 2021; Zhao & Bilen, 2023; Cazenavette et al., 2022). These approaches can be effective, but optimizing high-resolution images through repeated network evaluations can be expensive.

Pretrained diffusion models offer a complementary approach. Their generative priors enable synthesis of realistic images without learning an image distribution from a small distilled set alone (Ho et al., 2020; Peebles & Xie, 2023; Gu et al., 2024). However, a generator is trained to model images, whereas dataset distillation is evaluated by the classifiers trained on those images. A low denoising loss or realistic appearance therefore need not imply high classification accuracy. This distinction matters when a small synthetic set must represent subtle differences between similar classes. We use *generative–discriminative mismatch* to describe this difference in objectives. Visual fidelity is therefore an incomplete proxy for distilled-set utility.

This motivates a structural question: can the features already available in a pretrained generator be routed differently for each class to improve the resulting training set? We propose `MaskDD` (Figure 1), which learns channel masks inside a frozen diffusion transformer (DiT). Rather than updating backbone weights or directly optimizing each synthetic image, the method adapts the generator through class-specific gate parameters. A denoising term preserves class-conditioned reconstruction, while structural regularizers encourage sparse, nearly binary masks with reduced cross-class overlap. The masks define the feature pathways used during synthesis.

Mask learning also introduces an optimization difficulty. Sparsity pressure lowers gate values, and temperature annealing can leave saturated logits with small gradients. Continuing the same penalty throughout training may limit the ability of useful channels to recover. We address this with a rollback schedule inspired by weight rewinding (Frankle & Carbin, 2019; Frankle et al., 2020).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

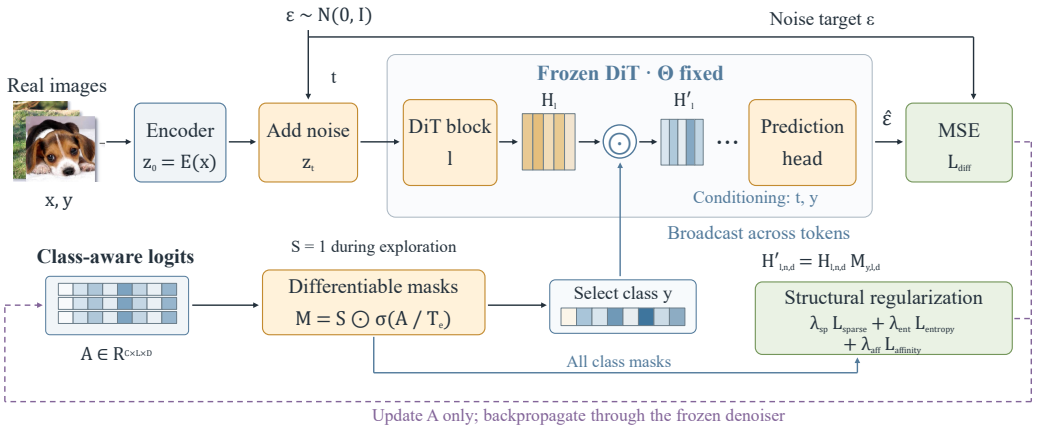

Figure 1: MaskDD adapts a frozen DiT using class-aware channel gates. Real images are encoded and corrupted before denoising. The selected class mask is broadcast across tokens and applied channel-wise; structural penalties act on the masks. Dashed arrows denote gradient updates to logits only. The internal gate placement is schematic. Synthetic images are subsequently sampled with the learned masks.

After exploration, we select the surviving support, restore its logits to an earlier checkpoint, and refine the gates without the sparsity penalty. Rollback acts on mask logits, not on the frozen DiT weights, and does not assume that the selected support is globally optimal.

Our contributions are threefold. First, we develop class-aware masking for adapting a frozen DiT to dataset distillation. Second, we separate support discovery from gate refinement through rollback and analyze mask gradients and hardening error under explicit assumptions. Third, we evaluate synthetic datasets across image budgets and classifier architectures, identifying improvements and settings in which competing methods remain stronger.

2 Related Work

Dataset distillation. Early methods optimize compact training sets through differentiable learning objectives (Wang et al., 2018), distribution matching (Zhao & Bilen, 2023), and trajectory matching (Cazenavette et al., 2022). For larger datasets, SRe^{2}L separates model training, image recovery, and relabeling (Yin et al., 2023), whereas RDED constructs compact datasets from selected real-image content (Sun et al., 2024). Neither relies on a diffusion generator. Diffusion-based approaches include Minimax Diffusion (Gu et al., 2024), D^{4}M (Su et al., 2024), and MGD^3 (Chan Santiago et al., 2025), which combine pretrained image priors with mechanisms for representation or diversity. CaO_2 studies inconsistencies between synthesis and evaluation (Wang et al., 2025). Aligning synthesis with downstream utility is therefore an existing research concern. Our focus is class-specific internal channel gates and their optimization schedule.

Intervening in diffusion generation. IGD uses trajectory-influence signals to guide reverse diffusion toward training-effective samples (Chen et al., 2025). IGDS instead balances label-relevant information and within-class variation through an information-guided sampling objective (Ye et al., 2025). CaO_2 optimizes latent representations during inference to address feature- and class-level inconsistencies (Wang et al., 2025). These approaches motivate a distinction between modifying a sample's generation trajectory and learning a reusable internal routing rule. MaskDD learns its class-specific gates before synthesis and reuses them across images and timesteps. It does not optimize a trajectory-influence objective or claim an information-theoretic equivalence to IGDS. Whether channel routing complements these sampling interventions requires a direct combined experiment.

Sparse adaptation and rewinding. The lottery ticket hypothesis and weight-rewinding studies investigate trainable sparse subnetworks within larger networks (Frankle & Carbin, 2019; Frankle et al., 2020). Differentiable relaxations enable optimization of discrete architectural decisions (Jang

et al., 2017; Louizos et al., 2018). MaskDD uses a different setting: the pretrained backbone remains fixed, and only channel-mask logits are learned and rewound. The selected support is not certified as a winning ticket. Appendix E discusses the distinction from other generative-model adaptations.

3 PRELIMINARIES

Dataset distillation. Let $\mathcal{D}_{\text{real}} = \{(x_i, y_i)\}_{i=1}^{n}$ contain images and labels from C classes. With a budget of K images per class (IPC), distillation produces $\mathcal{D}_{\text{syn}}$ with $CK \ll n$ examples. Utility is measured by training a classifier on $\mathcal{D}_{\text{syn}}$ and evaluating it on held-out real images. The generator and evaluation classifier are distinct. In MaskDD, the synthetic set is parameterized implicitly by a masked generator.

Latent diffusion. Let $z_0 = \mathcal{E}(x)$ be an image latent under a pretrained encoder. At a discrete diffusion timestep $t \in \{1, \dots, T_{\text{diff}}\}$, the forward corruption is

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t$ is the cumulative noise-schedule coefficient. A conditional denoiser $\epsilon_\Theta(z_t, t, y)$ predicts the noise. Squared noise prediction is related to a variational likelihood bound, with details depending on timestep weighting (Ho et al., 2020); it is not a classification objective. We freeze Θ and optimize only the masks.

4 METHOD

MaskDD learns class-aware masks on real training images and then uses them during reverse diffusion. Let L be the number of masked blocks and D their common channel dimension. The trainable logits are $A \in \mathbb{R}^{C \times L \times D}$; all backbone weights remain fixed.

4.1 CLASS-AWARE DIFFERENTIABLE MASKING

At training epoch e, the soft gate for class c and block l is

$$P_{c,l}^{(e)} = \sigma(A_{c,l}/T_e), \qquad T_e = \max\{T_{\min}, T_0 \gamma^e\}, \quad 0 < \gamma < 1. \tag{2}$$

Here $T_e > 0$ is the gate temperature, distinct from diffusion timestep t. For a batch with labels y_b and hidden states $H_l \in \mathbb{R}^{B \times N \times D}$,

$$H'_{l,b,n,d} = H_{l,b,n,d}\, \mathcal{M}_{y_b,l,d}. \tag{3}$$

The mask is shared across tokens of an image; different labels within a batch select different masks. During exploration, $\mathcal{M} = P^{(e)}$. After rollback, a fixed binary support gates $P^{(e)}$, as defined in Section 4.3.

This parameterization is a multiplicative class-conditional adapter: it rescales existing channels without introducing an additive shift or modifying the backbone weights. Its structural interpretation comes from support selection and hardening, rather than from class conditioning alone. Because masks can overlap, they define a shared generative core with class-dependent routing variations, not mutually exclusive subnetworks. The continuous-scaling control in Section 5.3 tests the contribution of the complete structural schedule beyond this additional conditioning capacity.

For fixed logits and temperature, gates are deterministic. Annealing encourages polarization but does not make all gates binary: a zero logit remains at $1/2$. The mask is held fixed across denoising timesteps during synthesis. This removes mask-sampling randomness, without eliminating randomness from data, diffusion noise, or optimization. Appendix C compares deterministic gates with stochastic perturbations; no universal variance ordering is assumed.

4.2 TRAINING OBJECTIVE

We optimize a denoising term and three structural penalties:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{sp}} \mathcal{L}_{\text{sparse}} + \lambda_{\text{ent}} \mathcal{L}_{\text{entropy}} + \lambda_{\text{aff}} \mathcal{L}_{\text{affinity}}. \tag{4}$$

The coefficients are nonnegative; λ_{sp} becomes zero after rollback.

Denoising. With the noisy latent in Eq. (1), we use

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{(x,y),t,\epsilon}\left[\|\epsilon - \epsilon_\Theta(z_t, t, y; \mathcal{M}_{y,1:L})\|_2^2\right]. \tag{5}$$

Although Θ is frozen, differentiation through the denoiser is needed to update A. This encourages the masked network to retain class-conditioned reconstruction ability. Its contribution to classification is assessed experimentally rather than assumed from the loss.

Sparsity and polarization. Write $Q = CLD$. Average gate mass and binary entropy are

$$\mathcal{L}_{\text{sparse}} = \frac{1}{Q}\sum_{c,l,d} \mathcal{M}_{c,l,d}, \tag{6}$$

$$\mathcal{L}_{\text{entropy}} = \frac{1}{Q}\sum_{c,l,d} h(\mathcal{M}_{c,l,d}), \qquad h(p) = -p\log p - (1-p)\log(1-p). \tag{7}$$

We use $0 \log 0 = 0$. Sparsity discourages large gate mass; entropy encourages values near either endpoint. A gate near one has low entropy but contributes to sparsity. For nonnegative gates, the mass penalty equals their normalized ℓ_1 norm. Both penalties are smooth functions of finite logits; Appendix A.2 derives their gradients.

Cross-class overlap. For $C > 1$, define

$$\mathcal{L}_{\text{affinity}} = \frac{2}{C(C-1)L}\sum_{i<j}\sum_{l=1}^{L} \frac{\langle \mathcal{M}_{i,l}, \mathcal{M}_{j,l}\rangle}{\max\{\|\mathcal{M}_{i,l}\|_2\|\mathcal{M}_{j,l}\|_2, \delta\}}, \quad \delta > 0. \tag{8}$$

The denominator floor handles zero masks. Nonnegative masks make this term lie in $[0, 1]$. It discourages overlap, but shared channels may remain useful, and zero overlap need not be attainable or desirable. For large class counts, we average within fixed groups; this changes the penalized pairs (Appendix B).

4.3 Support Discovery and Rollback

The derivative of an exploratory gate is $\partial P/\partial A = P(1-P)/T_e$. Saturated gates can have small gradients when $|A|/T_e$ is large. Our schedule separates support selection from optimization of surviving gates.

Exploration and checkpointing. We optimize Eq. (4) and save a detached copy A^{anc} at the start of epoch E_{anc}. At a later epoch $E_{\text{trg}} > E_{\text{anc}}$, we select

$$S_{c,l,d} = \mathbb{1}[P_{c,l,d}^{(E_{\text{trg}})} > 1/2] = \mathbb{1}[A_{c,l,d} > 0]. \tag{9}$$

This equivalence holds for every positive temperature. The trigger epoch is a schedule parameter; the support is selected from the learned state.

Rewind and restricted refinement. We reset surviving logits to the anchor and assign a large negative constant to inactive coordinates:

$$A_{c,l,d} \leftarrow \begin{cases} A^{\text{anc}}_{c,l,d}, & S_{c,l,d} = 1, \\ -\Omega, & S_{c,l,d} = 0, \end{cases} \qquad \mathcal{M}_{c,l,d} = S_{c,l,d}\,\sigma(A_{c,l,d}/T_e). \tag{10}$$

The support multiplier makes inactive channels exactly zero; a finite $-\Omega$ alone only approximates zero. We reset optimizer moments, set $\lambda_{\text{sp}} = 0$, and update active logits. Inactive coordinates are restored after each update to prevent decoupled weight decay from moving them. Surviving gates remain differentiable; an ordinary binary indicator would block their gradients. The support cannot expand, but surviving gates can still shrink.

Algorithm 1 `MaskDD` mask learning with support-restricted rollback

Require: Frozen denoiser Θ, training data, logits A, optimizer;
epochs $0 \le E_{\text{anc}} < E_{\text{trg}} < E$, temperatures and loss weights.

Initialize $S \leftarrow \mathbf{1}$.
for $e = 0, \dots, E-1$ **do**
  Set T_e using Eq. (2).
  if $e = E_{\text{anc}}$ **then**
    $A^{\text{anc}} \leftarrow \text{copy}(A)$, detached from gradients.
  end if
  if $e = E_{\text{trg}}$ **then**
    Select S with Eq. (9); rewind A with Eq. (10).
    Reset optimizer moments; set $\lambda_{\text{sp}} \leftarrow 0$.
  end if
  for each training minibatch (x, y) **do**
    Sample t, ϵ and form z_t using Eq. (1).
    Form $\mathcal{M} = S \odot \sigma(A/T_e)$ and compute Eq. (4).
    Differentiate with respect to A; update active coordinates only.
    Restore $A_{c,l,d} = -\Omega$ wherever $S_{c,l,d} = 0$.
  end for
end for
return hard masks from Eq. (11).

Hardening and synthesis. After the last update, we threshold the refined gates:

$$\mathcal{M}^*_{c,l,d} = S_{c,l,d} \mathbb{1}[\sigma(A_{c,l,d}/T_{\text{last}}) > 1/2]. \tag{11}$$

The final support is contained in the rollback support. For the flattened final soft mask $m \in [0,1]^Q$ and its thresholded version $\widehat{m}$, entropy gives the gate-level guarantee

$$\frac{\|m - \widehat{m}\|_1}{Q} \le \frac{\mathcal{L}_{\text{entropy}}(m)}{2 \log 2}. \tag{12}$$

The bound follows from concavity of binary entropy and applies to inactive coordinates as well, since their error is zero. It explains why polarization is useful before replacing soft gates with hard routing. It does not identify an optimal entropy coefficient: excessive early polarization can impede support exploration. Nor does it guarantee small image distortion, which additionally depends on the denoiser's sensitivity to masks and error propagation through the sampler. Appendix A.3 gives the proof and a conditional Lipschitz bound.

We sample K images per class and decode the resulting latents. Classifier-free guidance uses

$$\widehat{\epsilon} = (1 - s)\epsilon_\Theta(z_t, t, \emptyset; \mathcal{U}_y) + s\epsilon_\Theta(z_t, t, y; \mathcal{M}^*_y), \tag{13}$$

where s is the guidance scale and $\mathcal{U}_y$ is the unconditional-branch routing. This routing must be specified separately: class masks alone do not define a learned null-class mask. Algorithm 1 summarizes training. Appendix B records the available settings and remaining reproducibility requirements.

5 Experiments

5.1 Experimental Setup

We evaluate 256×256 synthetic images on ImageNette, ImageWoof (Howard, 2019a;b), the ImageNet-Dogs subset denoted IDC in the source experiment tables, ImageNet-100 (Tian et al., 2020), and ImageNet-1K (Russakovsky et al., 2015). The principal budgets are 10, 20, and 50 IPC; additional ImageWoof results cover 70 and 100 IPC. Evaluated classifiers include ConvNet-6, ResNetAP-10, and ResNet-18. We retain the reported means and standard deviations where available. The record refers to hard- and soft-label protocols, but does not fully specify their assignment, downstream settings, or run counts; Appendix B details these reporting limitations.

Comparisons include random selection, K-Center (Sener & Savarese, 2018), herding (Welling, 2009), distribution matching (DM) (Zhao & Bilen, 2023), IDC (Kim et al., 2022), the pretrained

Table 1: Full ImageWoof Top-1 accuracy (%) at 256×256 resolution. CN: ConvNet-6; AP: ResNetAP-10; RN: ResNet-18. Subscripts give standard deviations. DiT, DM, and IDC-1 are quoted from Gu et al. (2024); Minimax and MGD^3 are our reproductions. Bold/underline mark the highest/second-highest distinct means, excluding Full; tied means share the same rank.

IPC (ratio)	Net	Random	K-Center	Herding	DiT	DM	IDC-1	Minimax	MGD^3	MaskDD	Full
10 (0.8%)	CN	$24.3_{\pm 1.1}$	$19.4_{\pm 0.9}$	$26.7_{\pm 0.5}$	$\mathbf{34.2_{\pm 1.1}}$	$26.9_{\pm 1.2}$	$33.3_{\pm 1.1}$	$\underline{34.1_{\pm 0.4}}$	$33.5_{\pm 1.9}$	$34.0_{\pm 0.2}$	$86.4_{\pm 0.2}$
	AP	$29.4_{\pm 0.8}$	$22.1_{\pm 0.1}$	$32.0_{\pm 0.3}$	$34.7_{\pm 0.5}$	$30.3_{\pm 1.2}$	$\underline{39.1_{\pm 0.5}}$	$35.7_{\pm 0.3}$	$36.6_{\pm 0.9}$	$\mathbf{40.2_{\pm 0.2}}$	$87.5_{\pm 0.5}$
	RN	$27.7_{\pm 0.9}$	$21.1_{\pm 0.4}$	$30.2_{\pm 1.2}$	$34.7_{\pm 0.4}$	$33.4_{\pm 0.7}$	$37.3_{\pm 0.2}$	$\underline{37.6_{\pm 0.9}}$	$35.1_{\pm 1.8}$	$\mathbf{40.2_{\pm 0.4}}$	$89.3_{\pm 1.2}$
20 (1.6%)	CN	$29.1_{\pm 0.7}$	$21.5_{\pm 0.8}$	$29.5_{\pm 0.3}$	$36.1_{\pm 0.8}$	$29.9_{\pm 1.0}$	$35.5_{\pm 0.8}$	$\underline{36.8_{\pm 0.2}}$	$36.2_{\pm 1.6}$	$\mathbf{36.9_{\pm 0.2}}$	$86.4_{\pm 0.2}$
	AP	$32.7_{\pm 0.4}$	$25.1_{\pm 0.7}$	$34.9_{\pm 0.1}$	$41.1_{\pm 0.8}$	$35.2_{\pm 0.6}$	$43.4_{\pm 0.3}$	$43.3_{\pm 0.3}$	$\mathbf{44.5_{\pm 2.8}}$	$\underline{44.2_{\pm 0.2}}$	$87.5_{\pm 0.5}$
	RN	$29.7_{\pm 0.5}$	$23.6_{\pm 0.3}$	$32.2_{\pm 0.7}$	$40.5_{\pm 0.5}$	$29.8_{\pm 1.7}$	$38.6_{\pm 0.2}$	$\underline{40.9_{\pm 0.6}}$	$40.3_{\pm 2.5}$	$\mathbf{44.5_{\pm 0.5}}$	$89.3_{\pm 1.2}$
50 (3.8%)	CN	$41.3_{\pm 0.6}$	$36.5_{\pm 1.0}$	$40.3_{\pm 0.7}$	$46.5_{\pm 0.8}$	$44.4_{\pm 1.0}$	$43.9_{\pm 1.2}$	$51.4_{\pm 0.4}$	$\underline{51.9_{\pm 0.4}}$	$\mathbf{54.2_{\pm 0.5}}$	$86.4_{\pm 0.2}$
	AP	$47.2_{\pm 1.3}$	$40.6_{\pm 1.0}$	$49.1_{\pm 0.7}$	$49.3_{\pm 0.2}$	$47.1_{\pm 1.1}$	$48.3_{\pm 1.0}$	$54.4_{\pm 0.6}$	$\underline{55.6_{\pm 1.0}}$	$\mathbf{57.5_{\pm 0.1}}$	$87.5_{\pm 0.5}$
	RN	$47.9_{\pm 1.8}$	$39.6_{\pm 1.0}$	$48.3_{\pm 1.2}$	$50.1_{\pm 0.5}$	$46.2_{\pm 0.6}$	$48.3_{\pm 0.8}$	$53.9_{\pm 0.6}$	$\underline{56.3_{\pm 0.5}}$	$\mathbf{59.0_{\pm 0.4}}$	$89.3_{\pm 1.2}$
70 (5.4%)	CN	$46.3_{\pm 0.6}$	$38.6_{\pm 0.7}$	$46.2_{\pm 0.6}$	$50.1_{\pm 1.2}$	$47.5_{\pm 0.8}$	$48.9_{\pm 0.7}$	$51.8_{\pm 0.6}$	$\underline{53.1_{\pm 0.9}}$	$\mathbf{55.2_{\pm 0.3}}$	$86.4_{\pm 0.2}$
	AP	$50.8_{\pm 0.6}$	$45.9_{\pm 1.5}$	$53.4_{\pm 1.4}$	$54.3_{\pm 0.9}$	$51.7_{\pm 0.8}$	$52.8_{\pm 1.8}$	$\underline{56.5_{\pm 0.5}}$	$\mathbf{59.1_{\pm 1.4}}$	$\mathbf{59.1_{\pm 0.1}}$	$87.5_{\pm 0.5}$
	RN	$52.1_{\pm 1.0}$	$44.6_{\pm 1.1}$	$49.7_{\pm 0.8}$	$51.5_{\pm 1.0}$	$51.9_{\pm 0.8}$	$51.1_{\pm 1.7}$	$52.7_{\pm 0.4}$	$\underline{59.1_{\pm 0.1}}$	$\mathbf{60.9_{\pm 0.4}}$	$89.3_{\pm 1.2}$
100 (7.7%)	CN	$52.2_{\pm 0.4}$	$45.1_{\pm 0.5}$	$54.4_{\pm 1.1}$	$53.4_{\pm 0.3}$	$55.0_{\pm 1.3}$	$53.2_{\pm 0.9}$	$55.5_{\pm 0.6}$	$\underline{58.9_{\pm 0.3}}$	$\mathbf{59.3_{\pm 0.3}}$	$86.4_{\pm 0.2}$
	AP	$59.4_{\pm 1.4}$	$54.8_{\pm 0.2}$	$61.7_{\pm 0.9}$	$58.3_{\pm 0.8}$	$56.4_{\pm 0.8}$	$56.1_{\pm 0.9}$	$59.9_{\pm 0.4}$	$\underline{64.3_{\pm 1.5}}$	$\mathbf{64.4_{\pm 0.1}}$	$87.5_{\pm 0.5}$
	RN	$61.5_{\pm 1.3}$	$50.4_{\pm 0.4}$	$59.3_{\pm 0.7}$	$58.9_{\pm 1.3}$	$60.2_{\pm 1.0}$	$58.3_{\pm 1.2}$	$57.6_{\pm 0.6}$	$\underline{65.7_{\pm 1.0}}$	$\mathbf{66.4_{\pm 0.1}}$	$89.3_{\pm 1.2}$

DiT (Peebles & Xie, 2023), and diffusion-based methods Minimax and MGD^3 (Gu et al., 2024; Chan Santiago et al., 2025). The ImageNet-1K table also includes SRe^{2}L and RDED (Yin et al., 2023; Sun et al., 2024). The recorded MaskDD synthesis setup uses 50-step DDIM sampling (Song et al., 2021). MGD^3 reports DDPM as its default sampler and also evaluates DDIM (Chan Santiago et al., 2025); its published results retain their source protocols. Sampler and evaluation settings should therefore not be assumed identical across all entries.

5.2 Downstream Classification

ImageWoof. Table 1 reports all methods across five image budgets and three classifiers. At 50 IPC, gains over MGD^3 are 2.3, 1.9, and 2.7 points for ConvNet-6, ResNetAP-10, and ResNet-18. The advantage is not universal: at 10 IPC, DiT scores 34.2% with ConvNet-6 versus 34.0% for MaskDD; at 20 IPC, MGD^3 scores 44.5% with ResNetAP-10 versus 44.2% for MaskDD. At 70 IPC, MGD^3 and MaskDD tie at 59.1% with ResNetAP-10. Appendix C.1 repeats the complete results in a transposed layout for easier reading.

Other ImageNet subsets. On ImageNette and the IDC subset, MaskDD has the highest reported mean in each of the six configurations in Table 12. At 50 IPC, its gains over Minimax are 4.1 and 6.0 percentage points, respectively. On ImageNet-100 (Table 2), it leads all three classifiers at 10 IPC. At 20 IPC, MGD^3 remains stronger for ConvNet-6 and ResNetAP-10, while MaskDD leads for ResNet-18. The benefit therefore depends on both the budget and the evaluation architecture.

ImageNet-1K. Table 3 reports 44.9% and 59.7% with ResNet-18 at 10 and 50 IPC, exceeding Minimax by 0.6 and 1.1 percentage points. These improvements are smaller than those on several ten-class subsets. Table 4 adds MGD^3, which is stronger with ResNet-18 at 50 IPC; the restricted ranking in Table 3 therefore does not imply superiority over all generative methods.

Transfer across evaluators. Reusing the same 50-IPC synthetic dataset yields $64.9 \pm 2.0\%$ with ResNet-50 and $67.8 \pm 1.6\%$ with ResNet-101 (Table 4). The gains over MGD^3 are only 0.3 and 0.1 points, smaller than the reported standard deviations. A separate ViT-B/16 evaluation reaches $63.5 \pm 0.4\%$, extending the assessment beyond the ResNet family. This shows that the synthetic set can train multiple architectures without regeneration; it does not establish architecture-independent superiority. The additional evaluation identifies the ResNet-18 backbone, but full training and labeling specifications remain necessary for replication.

Table 2: ImageNet-100 Top-1 accuracy (%). Bold and underline mark the highest and second-highest distinct means among compact-data methods in each column. CN: ConvNet-6; AP: ResNetAP-10; RN: ResNet-18.

	10 IPC			20 IPC		
Method	CN	AP	RN	CN	AP	RN
Random	17.0 ± 0.3	19.1 ± 0.4	17.5 ± 0.5	24.8 ± 0.2	26.7 ± 0.5	25.5 ± 0.3
Herding	17.2 ± 0.3	19.8 ± 0.3	16.1 ± 0.2	24.3 ± 0.4	27.6 ± 0.1	24.7 ± 0.1
IDC	$\underline{24.3 \pm 0.5}$	25.7 ± 0.1	$\underline{25.1 \pm 0.2}$	28.8 ± 0.3	29.9 ± 0.2	30.2 ± 0.2
Minimax	22.3 ± 0.5	24.8 ± 0.2	22.5 ± 0.3	29.3 ± 0.4	32.1 ± 0.3	31.2 ± 0.1
MGD^3	23.4 ± 0.9	$\underline{25.8 \pm 0.5}$	23.6 ± 0.4	$\mathbf{30.6 \pm 0.4}$	$\mathbf{33.9 \pm 1.1}$	$\underline{32.6 \pm 0.4}$
`MaskDD`	$\mathbf{24.4 \pm 0.7}$	$\mathbf{27.0 \pm 1.0}$	$\mathbf{26.9 \pm 0.3}$	$\underline{29.4 \pm 0.9}$	$\underline{32.9 \pm 0.5}$	$\mathbf{33.6 \pm 1.1}$
Full data	79.9 ± 0.4	80.3 ± 0.2	81.8 ± 0.7	79.9 ± 0.4	80.3 ± 0.2	81.8 ± 0.7

Table 3: ImageNet-1K Top-1 accuracy (%) with ResNet-18. Bold marks the highest mean among the listed methods. Complete labeling and training settings remain to be specified; Table 4 adds the MGD^3 comparison at 50 IPC.

IPC	SRe^{2}L	RDED	DiT	Minimax	`MaskDD`
10	21.3 ± 0.6	42.0 ± 0.1	39.6 ± 0.4	44.3 ± 0.5	$\mathbf{44.9 \pm 0.3}$
50	46.8 ± 0.2	56.5 ± 0.1	52.9 ± 0.6	58.6 ± 0.3	$\mathbf{59.7 \pm 0.4}$

Table 4: ImageNet-1K at 50 IPC across ResNet evaluators. The same `MaskDD` dataset is reused across architectures. MGD^3 values are quoted under its published cross-architecture protocol, rather than a shared DDIM evaluation. Differences in means do not establish significance.

Method	ResNet-18	ResNet-50	ResNet-101
MGD^3 (Chan Santiago et al., 2025)	60.2 ± 0.1	64.6 ± 0.4	67.7 ± 0.4
`MaskDD`	59.7 ± 0.4	64.9 ± 2.0	67.8 ± 1.6

5.3 Structural Controls and Schedule Sensitivity

Continuous scaling versus structural selection. The no-sparsity control in Table 13 learns per-class continuous channel scales without discrete pruning, support locking, or anchor rewind. It scores 32.5% and 40.1% at 10 and 50 IPC. This control tests multiplicative scaling, not a full FiLM adapter with additive shifts. The shared-mask control reaches 35.1% and 42.4%. For a consistent full-method reference, the table repeats the main ResNet-18 results, $40.2 \pm 0.4\%$ and $59.0 \pm 0.4\%$, with default $\lambda_{\text{sp}} = 5$. The control rows have separately recorded settings, including different sparsity coefficients and an unspecified evaluator. Their differences from this reference cannot be attributed to individual components without confirming matched generation, labeling, and evaluation protocols. These controls identify the comparisons needed to distinguish continuous scaling, class conditioning, and support-restricted refinement.

When to rewind. Table 5 varies the anchor and trigger while holding the remaining settings fixed. The default $(E_{\text{anc}}/E, E_{\text{trg}}/E) = (0.2, 0.6)$ reaches 59.3% in this sweep, versus 59.0% in the main evaluation. Neighboring settings $(0.2, 0.5)$ and $(0.3, 0.6)$ reach 58.6% and 58.1%. An anchor at $0.1E$ performs worse across all tested triggers, while moving the trigger to $0.7E$ reduces accuracy for the later anchors. Thus, the schedule has a useful intermediate region, but is not insensitive to timing. These single reported values do not quantify uncertainty or establish robustness across datasets.

Table 5: Rollback schedule on ImageWoof, 50 IPC, ResNet-18. Entries are Top-1 accuracy (%); error bars were not reported. The selected schedule is an intermediate setting, rather than the earliest or latest tested trigger.

	E_{trg}/E			
E_{anc}/E	0.4	0.5	0.6	0.7
0.1	49.2	52.6	53.0	51.2
0.2	57.2	58.6	**59.3**	55.3
0.3	56.4	57.3	58.1	55.6

Regularization strength. The extended affinity sweep in Appendix C.5 separates overlap reduction from useful routing. At matched 5.6% sparsity, removing affinity gives 57.3%, compared with 59.0% at $\lambda_{\text{aff}} = 0.05$. Increasing the coefficient to 0.2 further lowers raw mask IoU but reduces accuracy to 56.5%. The entropy sweep similarly favors moderate values: 0.01 and 0.02 yield 59.0% and 58.4%, whereas 0.1 yields 52.1%. Lower overlap or lower entropy alone is therefore an inadequate selection criterion. These are one-dimensional sweeps with other settings fixed, not a joint sensitivity analysis.

5.4 Mask Structure and Synthetic-Data Properties

Interpreting overlap at high retention. The masks retain approximately 94.4% of channels, so high raw overlap is expected even without semantic organization. Let r_i and r_j denote retention rates and J_{ij} the measured IoU. Independent channel selection gives the density baseline and adjusted overlap

$$J_{ij}^{\text{ind}} = \frac{r_i r_j}{r_i + r_j - r_i r_j}, \qquad J_{ij}^{\text{adj}} = \frac{J_{ij} - J_{ij}^{\text{ind}}}{1 - J_{ij}^{\text{ind}}}. \tag{14}$$

The first expression is a ratio of expected intersection and union sizes, approximating expected IoU when channel counts concentrate. For $r_i = r_j = 0.944$, it is 0.894. Consequently, the raw range 0.88–0.95 in Figure 2 mainly reflects a large shared core. The adjustment is defined when $J_{ij}^{\text{ind}} < 1$; positive values indicate above-baseline sharing and negative values indicate below-baseline overlap.

Across the 45 unordered ImageWoof class pairs, the reported adjusted overlaps have Spearman correlations of 0.404, 0.499, and 0.540 with WordNet taxonomy, DINOv2 real-image prototypes, and real-data ResNet-18 confusion, respectively. The analysis uses 10,000 joint class-label permutations of the external similarity matrices, preserving the dependence among pairwise entries; Holm-adjusted p-values are 0.0460, 0.0357, and 0.0460 (Appendix C.6). These associations support structured routing variations beyond density in this setting. They do not imply disjoint semantic factors or establish that routing differences cause improved classification.

Diversity and coverage. The supplementary ImageWoof analysis at 50 IPC reports intra-class LPIPS of 0.670 for `MaskDD`, 0.695 for DiT, and 0.625 without rollback (Table 6). Thus, diversity under this metric is lower than the dense baseline, while rollback recovers part of the reduction. Reported real-to-synthetic and hard-example nearest-neighbor distances also improve relative to DiT, but their feature-space and threshold specifications are incomplete (Appendix C.7). These results motivate a joint assessment of accuracy and coverage; they cannot establish distribution-wide preservation or rule out memorization. Matched and mismatched activation examples in Figure 3 provide complementary qualitative evidence of class-dependent routing, without proving causal foreground isolation.

Sampling cost. The recorded efficiency comparison gives 9.38 versus 9.43 seconds per 100 images and 6.67 versus 6.63 GB peak memory (Table 20). This supports approximately unchanged sampling cost, not a substantial speed or memory improvement. Channel masking does not by itself remove dense tensor operations or stored backbone parameters.

Table 6: Accuracy and diversity in the additional ImageWoof 50-IPC analysis. These runs are separate from Table 1; their baseline accuracies differ. The full diagnostic table and its reporting limitations appear in Appendix C.7.

Method	Top-1 accuracy (%)	Intra-class LPIPS
DiT	52.4	0.695
Minimax	56.1	0.675
`MaskDD` without rollback	54.3 ± 0.8	0.625
`MaskDD`	59.0 ± 0.4	0.670

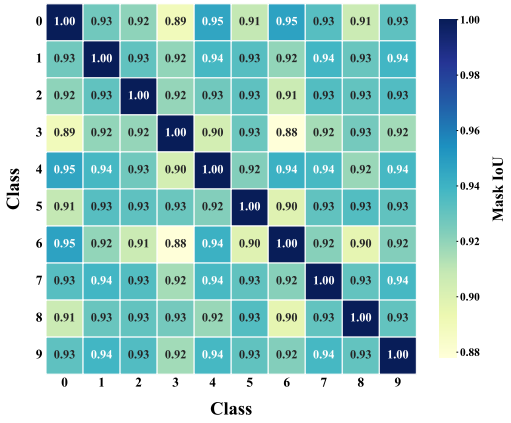

Figure 2: Pairwise overlap of active channel masks on ImageWoof. Raw overlap is high because masks retain most channels. Section 5.4 evaluates semantic associations after density correction.

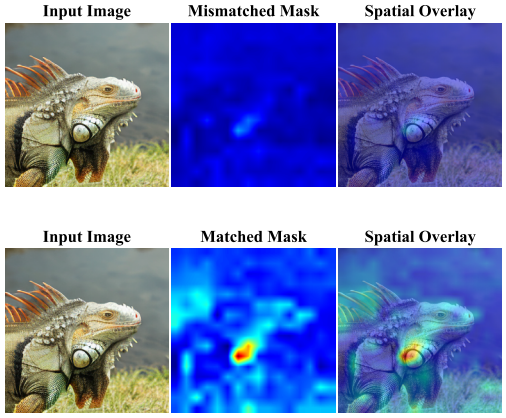

Figure 3: Feature activations on ImageNet-100 under mismatched Hummingbird routing (top) and matched Agama routing (bottom). Columns show the input image, mask response, and spatial overlay.

Budget-dependent utility. Table 1 distinguishes higher absolute accuracy from a growing advantage over competing methods. With ResNet-18, `MaskDD` improves from 40.2% to 66.4% as ImageWoof increases from 10 to 100 IPC, while its margin over MGD^3 narrows from 5.1 to 0.7 percentage points. More synthetic images therefore improve the reported classification accuracy without necessarily increasing the benefit of masking relative to another generator. The low-budget ConvNet-6 exception further shows that this relationship depends on the evaluator. Together, these observations motivate assessing generator adaptations across budgets and architectures rather than extrapolating their ranking from a single operating point. These differences compare reported means; establishing reliable advantages at any budget additionally requires matched protocols and uncertainty estimates for the comparison.

6 Conclusion

`MaskDD` adapts a frozen diffusion transformer through class-aware channel masks for dataset distillation. It combines denoising with structural regularization and separates support discovery from gate refinement through rollback. The reported results show improvements across several datasets and classifiers, particularly at larger ImageWoof budgets, while some settings favor existing methods. The additional analyses identify an intermediate rollback window, associations between adjusted mask overlap and external class structure, and transfer of one synthetic dataset across evaluator families. They also expose limits: large-scale margins are modest, higher accuracy does not imply universally higher diversity, and stronger polarization or separation can hurt utility. The mathematical analysis bounds gate-level hardening error rather than semantic isolation or downstream generalization. More complete protocol reporting, broader diversity tests, and physical compaction of supported channels remain important next steps.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex, assisted with restructuring and editing this manuscript, checking notation and mathematical arguments, interpreting the reported tables, redrawing method diagrams, and preparing the LaTeX format. This included reformulating the gate-gradient analysis and deriving a conditional bound on mask hardening error. Numerical experimental results were retained from the supplied research record; no new experimental results were generated during this revision. This statement describes the revision workflow and does not independently establish how the original experiments were conducted.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proc. CVPR*, pp. 10718–10727, 2022.

Jeffrey A. Chan Santiago, Praveen Tirupattur, Gaurav Kumar Nayak, Gaowen Liu, and Mubarak Shah. MGD^3: Mode-guided dataset distillation using diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 52861–52876, 2025.

Mingyang Chen, Jiawei Du, Bo Huang, Yi Wang, Xiaobo Zhang, and Wei Wang. Influence-guided diffusion for dataset distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proc. ICLR*, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proc. ICML*, pp. 3259–3269, 2020.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023.

Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: LoRA adaptors for precise control in diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *Proc. CVPR*, pp. 15793–15803, 2024.

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

Jeremy Howard. ImageNette: A smaller subset of 10 easily classified classes from ImageNet, 2019a. URL `https://github.com/fastai/imagenette`.

Jeremy Howard. ImageWoof: a subset of 10 classes from ImageNet that aren't so easy to classify, 2019b. URL `https://github.com/fastai/imagenette#imagewoof`.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning (ICML)*, 2022.

Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1100–1113, 2022.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations (ICLR)*, 2018.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. ICCV*, pp. 4195–4205, 2023.

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

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. D^4M: Dataset distillation via disentangled diffusion model. In *Proc. CVPR*, pp. 5809–5818, 2024.

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proc. CVPR*, pp. 9390–9399, 2024.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proc. ECCV*, pp. 776–794. Springer, 2020.

Haoxuan Wang, Zhenghao Zhao, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. CaO2: Rectifying inconsistencies in diffusion-based dataset distillation. In *Proc. ICCV*, pp. 4722–4731, 2025.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Max Welling. Herding dynamical weights to learn. In *Proc. ICML*, pp. 1121–1128, 2009.

Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. DiffFit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proc. ICCV*, pp. 4230–4239, 2023.

Linfeng Ye, Shayan Mohajer Hamidi, Guang Li, Takahiro Ogawa, Miki Haseyama, and Konstantinos N. Plataniotis. Information-guided diffusion sampling for dataset distillation. In *NeurIPS 2025 Workshop on Structured Probabilistic Inference and Generative Modeling*, 2025. URL https://openreview.net/forum?id=o2HVbnmazF.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at ImageNet scale from a new perspective. In *Advances in Neural Information Processing Systems*, volume 36, pp. 73582–73603, 2023.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6503–6512, 2023. doi: 10.1109/WACV56688.2023.00645.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations (ICLR)*, 2021.

A Analysis of the Mask Parameterization

The results below concern the stated gate parameterization. They do not prove that minimizing denoising loss identifies semantic foregrounds, maximizes classifier accuracy, or finds an optimal sparse architecture. The distinction matters because the generator is trained with denoising and structural losses, whereas utility is measured by a separate downstream classifier.

A.1 Deterministic Gates and Sources of Variance

At fixed A and $T > 0$, the map $P = \sigma(A/T)$ is deterministic, so $\mathrm{Cov}(P \mid A, T) = 0$. This says nothing about randomness in minibatches, noise samples, or diffusion timesteps. For a stochastic gate depending on additional randomness ξ, let G be its vector-valued gradient, with finite second moment. The law of total covariance gives

$$\mathrm{Cov}(G) = \mathbb{E}_\xi[\mathrm{Cov}(G \mid \xi)] + \mathrm{Cov}_\xi(\mathbb{E}[G \mid \xi]). \tag{15}$$

The second term is positive semidefinite. However, the first term need not equal the covariance of a deterministic-gate gradient. The decomposition therefore does not establish that deterministic masks always yield lower total gradient variance.

For a scalar perturbed sigmoid $P_\xi = \sigma((a + \xi)/T)$,

$$\left|\frac{\partial P_\xi}{\partial a}\right| = \frac{P_\xi(1 - P_\xi)}{T} \le \frac{1}{4T}. \tag{16}$$

This is an upper bound, not a lower bound asserting variance growth proportional to T^{-2}. A variance rate depends on the noise distribution, the logit, and the remaining loss derivatives. Training with stochastic masks also does not require resampling masks during inference: one can harden and hold a mask fixed before sampling. Thus, any claim about integration instability must specify the inference-time routing protocol.

A.2 Logit Gradients and Sparsity Pressure

Consider exploration at a fixed temperature, and index the $Q = CLD$ gates by q. Let $p_q = \sigma(a_q/T)$ and $F(P) = \mathcal{L}_{\text{diff}}(P) + \lambda_{\text{aff}}\mathcal{L}_{\text{affinity}}(P)$. Where F is differentiable, the chain rule gives

$$\frac{\partial \mathcal{L}}{\partial a_q} = \frac{p_q(1 - p_q)}{T}\left[\frac{\partial F}{\partial p_q} + \frac{\lambda_{\text{sp}}}{Q} + \frac{\lambda_{\text{ent}}}{Q}\log\frac{1 - p_q}{p_q}\right]. \tag{17}$$

The identity follows from $\partial p_q/\partial a_q = p_q(1 - p_q)/T$ and $h'(p) = \log((1 - p)/p)$. It holds for $0 < p_q < 1$; endpoint entropy values are handled by limits.

The sparsity contribution is nonnegative:

$$0 \le \frac{\partial(\lambda_{\text{sp}}\mathcal{L}_{\text{sparse}})}{\partial a_q} = \frac{\lambda_{\text{sp}}}{QT}p_q(1 - p_q) \le \frac{\lambda_{\text{sp}}}{4QT}. \tag{18}$$

A gradient-descent step therefore lowers the logit when only this penalty acts. For the full objective, an ordinary gradient-descent step raises the logit exactly when the bracket in Eq. (17) is negative. This is a local update condition, not a necessary and sufficient condition for eventual channel survival. AdamW moments and weight decay further modify the update.

Because $p_q > 0$ for finite a_q, the mask-mass penalty is smooth in the logits. Calling it a nonsmooth logit penalty or claiming that AdamW performs its proximal operator would be incorrect. The gate Jacobian also explains why saturation can slow recovery, but does not prove irreversible channel death in exact arithmetic.

A.3 A Bound on Hardening Error

Proposition A.1 (Entropy controls gate-level rounding error)**.** *Let $m \in [0, 1]^Q$ and $\widehat{m}_q = \mathbb{1}[m_q > 1/2]$. With natural logarithms and $0 \log 0 = 0$,*

$$\frac{1}{Q}\|m - \widehat{m}\|_1 \le \frac{1}{2\log 2}\,\frac{1}{Q}\sum_{q=1}^{Q} h(m_q). \tag{19}$$

Proof. For $p \in [0, 1/2]$, concavity of binary entropy and its values $h(0) = 0$ and $h(1/2) = \log 2$ imply $h(p) \geq 2p \log 2$. Symmetry, $h(p) = h(1-p)$, then gives $h(p) \geq 2 \log 2 \min\{p, 1-p\}$ on $[0, 1]$. The absolute rounding error is $\min\{p, 1-p\}$, including the tie at $p = 1/2$. Summing proves the claim. □

The result applies to support-restricted masks because inactive entries are exactly zero. If a denoiser output $g(m)$ is K_g-Lipschitz in the mask under the ℓ_1 norm on the relevant domain, then

$$\|g(m) - g(\widehat{m})\|_2 \leq K_g \|m - \widehat{m}\|_1 \leq \frac{K_g Q}{2 \log 2} \mathcal{L}_{\text{entropy}}. \tag{20}$$

Without a bound on K_g, low mask entropy alone does not establish negligible image distortion. Extending the result through a full sampler requires control of error propagation across denoising steps; downstream classification requires additional analysis.

A.4 What Channel Sparsity Does and Does Not Imply

Sparsity limits gate mass and, after hardening, the number of active channels. It is not in itself a variational bound on mutual information. For a fixed class c, if every soft gate is strictly positive, the map $H \mapsto \text{diag}(P_c)H$ is invertible channel-wise. Whenever mutual information is well defined, this invertible transformation preserves $I(X; H \mid Y = c)$. Thus, simply reducing positive gate magnitudes does not necessarily reduce information about X.

Hard masking is generally noninvertible. Conditional data processing can then give $I(X; \mathcal{M}_c^* \odot H \mid Y = c) \leq I(X; H \mid Y = c)$, but it does not identify whether the removed information concerns foreground, background, or the label. Nor does the denoising objective explicitly maximize a variational classifier likelihood. An information-bottleneck interpretation should therefore be treated as a possible motivation, not as an established equivalence of objectives.

For a token matrix $H \in \mathbb{R}^{N \times D}$ and a hard mask with r active channels, $\text{rank}(H \, \text{diag}(\mathcal{M}^*)) \leq r$. This elementary rank bound concerns the masked matrix itself. It does not automatically extend to the full DiT, which contains residual connections, conditioning, biases, and nonlinear operations. In particular, it does not prove that attention aligns with a foreground subspace.

A.5 Rollback and Local Curvature

At fixed temperature, the diagonal curvature contributed by the sparsity term is

$$\frac{\partial^2 (\lambda_{\text{sp}} \mathcal{L}_{\text{sparse}})}{\partial a_q^2} = \frac{\lambda_{\text{sp}}}{QT^2} p_q (1 - p_q)(1 - 2p_q). \tag{21}$$

Its sign changes at $p_q = 1/2$. Removing this term therefore does not uniformly reduce positive curvature, and neither removal nor rewinding guarantees a flat minimum. The defensible interpretation of rollback is operational: it restricts the admissible support, changes the initialization of active gates, resets optimizer state, and removes an explicit source of downward gate pressure. Evidence for its downstream benefit must come from controlled ablations.

Finally, support restriction is different from fixing every gate. With $\mathcal{M} = S \odot \sigma(A/T)$ and fixed S, gradients vanish on inactive coordinates and remain available on active ones. Final thresholding may remove additional channels, so $\text{supp}(\mathcal{M}^*) \subseteq \text{supp}(S)$. This is the precise support guarantee used in the revised algorithm.

B Implementation and Reporting Details

Available settings. The experimental description uses a pretrained DiT-XL/2 backbone at 256×256 resolution and optimizes mask logits with AdamW. The stated gate temperature decreases from 1.0 to 0.1. The implementation notes report $\lambda_{\text{sp}} = 5.0$, $\lambda_{\text{ent}} = 0.01$, and $\lambda_{\text{aff}} = 0.05$, and synthesis uses 50 DDIM steps (Song et al., 2021) with guidance scale $s = 4.0$. The stated hardware includes NVIDIA H800 and A100 GPUs, with RTX 4090 GPUs used for preliminary work. The negative logit sentinel is $\Omega = 100$.

The default sparsity coefficient is $\lambda_{\text{sp}} = 5.0$, consistent with the sparsity sweep and the full-method reference in Table 13; the 0.1 entries describe the recorded control configurations, not the default method. The record does not specify the mask initialization, learning rate, weight decay, batch size, training duration, annealing factor γ, insertion points within each DiT block, or the denominator floor δ. The supplementary schedule study specifies $E_{\text{anc}}/E = 0.2$ and $E_{\text{trg}}/E = 0.6$ for its default ImageWoof configuration. Entropy and affinity coefficients are shared across settings while sparsity is adjusted; the complete dataset-by-budget configuration map remains unspecified.

Differentiable refinement. Figure 4 summarizes the schedule; Algorithm 2 details the support-restricted refinement in Section 4. The default method uses deterministic sigmoid gates; Gumbel perturbations are confined to the robustness variants. Hard thresholds select the support and produce the final synthesis mask, while refinement uses differentiable gates restricted to the selected support. A hard indicator alone would require an additional gradient estimator to train the logits. The explicit support restriction and optimizer handling in the revised algorithm must be checked against the implementation associated with the reported experiments.

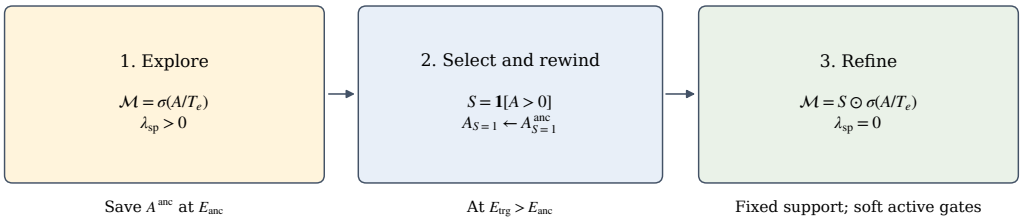

Figure 4: Rollback schedule. Exploration saves an anchor checkpoint; the trigger selects a support and rewinds surviving mask logits; refinement removes the sparsity penalty. The fixed object is the allowed support, while active gate values continue to change.

Algorithm 2 Details of a support-restricted refinement update

Require: Logits A, saved anchor A^{anc}, temperature T, optimizer.

At the rollback trigger

$S \leftarrow \mathbb{1}[A > 0]$; detach S from the computation graph.
$A \leftarrow S \odot A^{\text{anc}} - (1 - S)\Omega$.
Reset optimizer moments; set $\lambda_{\text{sp}} \leftarrow 0$.

For each subsequent minibatch

Form $P \leftarrow \sigma(A/T)$ and $\mathcal{M} \leftarrow S \odot P$.
Compute denoising and stabilized affinity with $\mathcal{M}$.
Compute entropy using $h(0) = h(1) = 0$ with numerically safe logarithms.
Backpropagate the objective and multiply the logit gradient by S.
Take an optimizer step; restore inactive entries of A to $-\Omega$.

After the final update

$\mathcal{M}^* \leftarrow S \odot \mathbb{1}[A > 0]$.
return $\mathcal{M}^*$.

Grouped affinity. For ImageNet-1K, the recorded strategy partitions $C = 1000$ classes into $G = 50$ disjoint groups of $m = 20$. Let these groups be $\mathcal{C}_g$. We write the average over the retained pairs as

$$\mathcal{L}^{\text{group}}_{\text{affinity}} = \frac{1}{L \sum_g \binom{|\mathcal{C}_g|}{2}} \sum_g \sum_{\substack{i<j \\ i,j \in \mathcal{C}_g}} \sum_{l=1}^{L} \cos_\delta(\mathcal{M}_{i,l}, \mathcal{M}_{j,l}), \tag{22}$$

where $\cos_\delta$ denotes the stabilized ratio in Eq. (8). There are $50\binom{20}{2} = 9{,}500$ pairs instead of $\binom{1000}{2} = 499{,}500$, a factor of approximately 52.6 fewer. For fixed group size, pairwise computation scales as

$O(CmLD)$ rather than $O(C^2LD)$. This is a polynomial reduction, not an exponential one. Cross-group overlaps are unpenalized; grouping does not guarantee global orthogonality. The class-to-group assignment remains to be specified.

Synthesis and evaluation protocol. The recorded DDIM step count and guidance scale do not specify the complete sampler configuration. In particular, Eq. (13) requires the routing $\mathcal{U}_y$ for the null-label branch: using dense routing, the target-class mask, or a separately learned mask are different choices. The supplied record does not resolve this choice or specify DDIM stochasticity, latent scaling, decoding and image preprocessing details.

The supplementary evaluation identifies ResNet-18 for the main ImageNet-1K results. The evaluation record refers to both hard-label and soft-label protocols, but does not map them unambiguously to all tables. It also omits the number of independent runs, some standard deviations, exact dataset class lists, downstream training schedules, and any teacher configuration for soft labels. These omissions limit reproducibility and interpretation of cross-method comparisons. Reported error bars are retained as standard deviations where supplied; no significance claim is inferred from them.

Baseline provenance and sampling. The Minimax and MGD^3 entries in the ImageWoof comparison (Table 1) are our reproductions, whereas the MGD^3 entries in Table 4 are quoted from its published evaluation. MGD^3 uses DDPM by default and separately studies DDIM (Chan Santiago et al., 2025). These alternatives do not establish which sampler was used for each reproduced result. The 50-step DDIM setting reported above describes the recorded `MaskDD` setup; a complete protocol comparison additionally requires the actual sampler, step count, label supervision, and evaluator settings for each baseline result.

C Additional Results and Analyses

C.1 Full Benchmark Tables

Tables 7–11 present the full ImageWoof comparison, transposed for legibility. DiT, DM, and IDC-1 results are quoted from Table 1 of Gu et al. (2024) and agree with Table 8 of Chan Santiago et al. (2025); Minimax and MGD^3 are our reproductions, and the remaining entries retain the supplied experiment record. Highest and second-highest means are marked independently of standard deviations; these markings are descriptive, not significance tests. Table 12 contains the ImageNette and IDC-subset comparison.

Table 7: Full ImageWoof results at 10 IPC (Top-1 accuracy, mean ± standard deviation, %).

Method	ConvNet-6	ResNetAP-10	ResNet-18
Random	24.3 ± 1.1	29.4 ± 0.8	27.7 ± 0.9
K-Center	19.4 ± 0.9	22.1 ± 0.1	21.1 ± 0.4
Herding	26.7 ± 0.5	32.0 ± 0.3	30.2 ± 1.2
DiT	$\mathbf{34.2 \pm 1.1}$	34.7 ± 0.5	34.7 ± 0.4
DM	26.9 ± 1.2	30.3 ± 1.2	33.4 ± 0.7
IDC-1	33.3 ± 1.1	$\underline{39.1 \pm 0.5}$	37.3 ± 0.2
Minimax	$\underline{34.1 \pm 0.4}$	35.7 ± 0.3	$\underline{37.6 \pm 0.9}$
MGD^3	33.5 ± 1.9	36.6 ± 0.9	35.1 ± 1.8
`MaskDD`	34.0 ± 0.2	$\mathbf{40.2 \pm 0.2}$	$\mathbf{40.2 \pm 0.4}$
Full dataset	86.4 ± 0.2	87.5 ± 0.5	89.3 ± 1.2

Table 8: Full ImageWoof results at 20 IPC (Top-1 accuracy, mean ± standard deviation, %).

Method	ConvNet-6	ResNetAP-10	ResNet-18
Random	29.1 ± 0.7	32.7 ± 0.4	29.7 ± 0.5
K-Center	21.5 ± 0.8	25.1 ± 0.7	23.6 ± 0.3
Herding	29.5 ± 0.3	34.9 ± 0.1	32.2 ± 0.7
DiT	36.1 ± 0.8	41.1 ± 0.8	40.5 ± 0.5
DM	29.9 ± 1.0	35.2 ± 0.6	29.8 ± 1.7
IDC-1	35.5 ± 0.8	43.4 ± 0.3	38.6 ± 0.2
Minimax	$\underline{36.8 \pm 0.2}$	43.3 ± 0.3	$\underline{40.9 \pm 0.6}$
MGD^3	36.2 ± 1.6	$\mathbf{44.5 \pm 2.8}$	40.3 ± 2.5
MaskDD	$\mathbf{36.9 \pm 0.2}$	$\underline{44.2 \pm 0.2}$	$\mathbf{44.5 \pm 0.5}$
Full dataset	86.4 ± 0.2	87.5 ± 0.5	89.3 ± 1.2

Table 9: Full ImageWoof results at 50 IPC (Top-1 accuracy, mean ± standard deviation, %).

Method	ConvNet-6	ResNetAP-10	ResNet-18
Random	41.3 ± 0.6	47.2 ± 1.3	47.9 ± 1.8
K-Center	36.5 ± 1.0	40.6 ± 1.0	39.6 ± 1.0
Herding	40.3 ± 0.7	49.1 ± 0.7	48.3 ± 1.2
DiT	46.5 ± 0.8	49.3 ± 0.2	50.1 ± 0.5
DM	44.4 ± 1.0	47.1 ± 1.1	46.2 ± 0.6
IDC-1	43.9 ± 1.2	48.3 ± 1.0	48.3 ± 0.8
Minimax	51.4 ± 0.4	54.4 ± 0.6	53.9 ± 0.6
MGD^3	$\underline{51.9 \pm 0.4}$	$\underline{55.6 \pm 1.0}$	$\underline{56.3 \pm 0.5}$
MaskDD	$\mathbf{54.2 \pm 0.5}$	$\mathbf{57.5 \pm 0.1}$	$\mathbf{59.0 \pm 0.4}$
Full dataset	86.4 ± 0.2	87.5 ± 0.5	89.3 ± 1.2

Table 10: Full ImageWoof results at 70 IPC (Top-1 accuracy, mean ± standard deviation, %).

Method	ConvNet-6	ResNetAP-10	ResNet-18
Random	46.3 ± 0.6	50.8 ± 0.6	52.1 ± 1.0
K-Center	38.6 ± 0.7	45.9 ± 1.5	44.6 ± 1.1
Herding	46.2 ± 0.6	53.4 ± 1.4	49.7 ± 0.8
DiT	50.1 ± 1.2	54.3 ± 0.9	51.5 ± 1.0
DM	47.5 ± 0.8	51.7 ± 0.8	51.9 ± 0.8
IDC-1	48.9 ± 0.7	52.8 ± 1.8	51.1 ± 1.7
Minimax	51.8 ± 0.6	$\underline{56.5 \pm 0.5}$	52.7 ± 0.4
MGD^3	$\underline{53.1 \pm 0.9}$	$\mathbf{59.1 \pm 1.4}$	$\underline{59.1 \pm 0.1}$
MaskDD	$\mathbf{55.2 \pm 0.3}$	$\mathbf{59.1 \pm 0.1}$	$\mathbf{60.9 \pm 0.4}$
Full dataset	86.4 ± 0.2	87.5 ± 0.5	89.3 ± 1.2

Table 11: Full ImageWoof results at 100 IPC (Top-1 accuracy, mean ± standard deviation, %).

Method	ConvNet-6	ResNetAP-10	ResNet-18
Random	52.2 ± 0.4	59.4 ± 1.4	61.5 ± 1.3
K-Center	45.1 ± 0.5	54.8 ± 0.2	50.4 ± 0.4
Herding	54.4 ± 1.1	61.7 ± 0.9	59.3 ± 0.7
DiT	53.4 ± 0.3	58.3 ± 0.8	58.9 ± 1.3
DM	55.0 ± 1.3	56.4 ± 0.8	60.2 ± 1.0
IDC-1	53.2 ± 0.9	56.1 ± 0.9	58.3 ± 1.2
Minimax	55.5 ± 0.6	59.9 ± 0.4	57.6 ± 0.6
MGD^3	$\underline{58.9 \pm 0.3}$	$\underline{64.3 \pm 1.5}$	$\underline{65.7 \pm 1.0}$
MaskDD	$\mathbf{59.3 \pm 0.3}$	$\mathbf{64.4 \pm 0.1}$	$\mathbf{66.4 \pm 0.1}$
Full dataset	86.4 ± 0.2	87.5 ± 0.5	89.3 ± 1.2

Table 12: ImageNette and the subset denoted IDC in the experiment record, evaluated with ResNetAP-10. Entries are Top-1 accuracy (%); bold and underline mark the highest and second-highest distinct means. The IDC dataset name is distinct from the IDC distillation method.

Dataset	IPC	Random	DiT	DM	Minimax	D^4M	MaskDD
Nette	10	54.2 ± 1.6	59.1 ± 1.2	60.8 ± 0.6	$\underline{62.0 \pm 0.2}$	60.4 ± 3.4	$\mathbf{64.8 \pm 0.7}$
	20	63.5 ± 0.5	64.8 ± 1.2	66.5 ± 1.1	$\underline{66.8 \pm 0.4}$	65.5 ± 1.2	$\mathbf{67.6 \pm 1.2}$
	50	76.1 ± 1.1	73.3 ± 0.1	76.2 ± 0.4	$\underline{76.6 \pm 0.2}$	73.8 ± 1.7	$\mathbf{80.7 \pm 0.5}$
IDC	10	48.1 ± 0.8	$\underline{54.1 \pm 0.4}$	52.8 ± 0.5	53.1 ± 0.2	51.1 ± 2.4	$\mathbf{54.8 \pm 1.4}$
	20	52.5 ± 0.9	58.9 ± 0.2	58.5 ± 0.4	$\underline{59.0 \pm 0.4}$	58.0 ± 1.4	$\mathbf{60.8 \pm 0.9}$
	50	68.1 ± 0.7	64.3 ± 0.6	69.1 ± 0.8	$\underline{69.6 \pm 0.2}$	64.1 ± 2.5	$\mathbf{75.6 \pm 0.6}$

C.2 Masking and Rollback Ablation

Table 13 uses the main-table ResNet-18 results as the full-method reference: $40.2 \pm 0.4\%$ at 10 IPC and $59.0 \pm 0.4\%$ at 50 IPC, with the default $\lambda_{\text{sp}} = 5$. The control rows retain their recorded settings and measurements. The no-sparsity row uses continuous per-class scaling without discrete pruning, support locking, or anchor rewind; it does not include the additive shift of a full FiLM adapter. The control evaluator and its correspondence to the main-table protocol remain unverified. Their differences from the reference therefore cannot isolate the effects of rollback, sparsity, or class conditioning; a matched comparison requires the same evaluator, labeling, generation, and training settings.

Table 13: ImageWoof controls and full-method reference. The reference repeats the ResNet-18 results in Table 1. Control configurations are not verified to match this protocol; differences across the two groups therefore do not isolate component effects.

Variant	λ_{sp}	Rollback	Sparsity (%)	10 IPC	50 IPC
Recorded controls					
Dense	0	No	0.0	32.0 ± 0.3	39.6 ± 0.4
No sparsity	0	No	0.0	32.5 ± 0.4	40.1 ± 0.3
No rollback	0.1	No	~ 85.4	21.4 ± 1.2	28.5 ± 1.0
Shared mask	0.1	Yes	~ 12.3	35.1 ± 0.6	42.4 ± 0.5
Main-table reference: ResNet-18					
Full method	5	Yes	~ 5.6	40.2 ± 0.4	59.0 ± 0.4

C.3 Stochastic Perturbation

Table 14 records a Gumbel-perturbed variant with LPIPS 0.681 and accuracy 39.0%, versus 0.673 and 40.2% for deterministic gates. Higher LPIPS is not sufficient evidence of better class coverage: it may also reflect irrelevant variation or artifacts. The comparison does not prove that stochastic routing is generally inferior. The conflicting IPC descriptions must be resolved before drawing a budget-specific conclusion. Moreover, adding a single Gumbel random variable inside a sigmoid should not automatically be identified with a standard binary Concrete relaxation, which uses a logistic perturbation, equivalently a difference of independent Gumbel variables.

Table 14: Recorded stochastic-mask comparison on ImageWoof. The original table specifies 10 IPC, whereas its accompanying prose specifies 50 IPC; the budget and classifier need verification. LPIPS and accuracy are retained without assuming a statistical difference.

Gate variant	Intra-class LPIPS	Top-1 accuracy (%)
Gumbel perturbation	0.681	39.0
Deterministic sigmoid	0.673	40.2

C.4 Hyperparameter Sensitivity

The recorded sparsity sweep in Table 15 uses ImageWoof at 50 IPC with ResNetAP-10. Accuracy peaks at $\lambda_{\text{sp}} = 5$ among the tested values. A larger coefficient coincides with more pruning and lower accuracy, while small coefficients produce little pruning. These are observations for this sweep, not proof of a universal optimal sparsity level. Table 5 adds the rollback-epoch sweep; a quantitative temperature-schedule sweep is still unavailable.

Table 15: Recorded sparsity sweep on ImageWoof, 50 IPC, ResNetAP-10. Error bars were not provided.

λ_{sp}	Channel sparsity at trigger (%)	Top-1 accuracy (%)
20	29.4	24.5
10	14.7	53.8
5	5.6	57.5
1	3.0	49.4
0.1	0.0	49.1

Table 16 records the entropy-weight sweep for ImageWoof at 50 IPC with ResNet-18. The highest mean among the tested settings occurs at $\lambda_{\text{ent}} = 0.01$. This supports the selected value within that experiment but does not quantify gate discretization error or isolate the refinement stage.

Table 16: Extended entropy sweep on ImageWoof, 50 IPC, ResNet-18. The pre-rollback entropy is reported separately from the schedule experiment. Error bars were not provided.

λ_{ent}	Top-1 accuracy (%)	Pre-rollback entropy
0	52.3	0.128
0.001	54.2	0.095
0.005	56.1	0.062
0.01	59.0	0.035
0.02	58.4	0.018
0.05	56.2	0.008
0.1	52.1	0.0025

C.5 Extended Affinity Analysis

Table 17 records the additional ImageWoof experiment at 50 IPC. The matched-sparsity control improves comparability of the channel budget, although the procedure used to achieve that budget remains to be specified. Its 57.3% accuracy is below the 59.0% default result. Stronger affinity reduces both soft cosine and hard-mask IoU but eventually lowers accuracy and LPIPS. Because mask density also changes, raw IoU should not be interpreted as a density-independent measure of class separation.

Table 17: Extended affinity sweep on ImageWoof, 50 IPC. Acc.: Top-1 accuracy; Sp.: channel sparsity; Cos.: soft-mask cosine; IoU: hard-mask overlap; Conf.: reported class confusion; R2S: relative real-to-synthetic nearest-neighbor distance. The confusion and distance implementations need further specification. Error bars were not supplied.

λ_{aff}	Acc. (%)	Sp. (%)	Cos.	IoU	Conf.	LPIPS	R2S
0	57.1	4.5	0.960	0.955	0.310	0.685	1.00
0, matched	57.3	5.6	0.950	0.945	0.300	0.680	0.98
0.01	57.8	5.2	0.940	0.930	0.265	0.678	0.95
0.025	58.4	5.4	0.925	0.910	0.235	0.674	0.91
0.05	59.0	5.6	0.895	0.890	0.200	0.670	0.86
0.1	58.1	8.5	0.845	0.820	0.220	0.658	0.94
0.2	56.5	18.5	0.750	0.680	0.275	0.635	1.08

Structural statistics at the trigger. For $E_{\mathrm{trg}}/E = 0.4, 0.5, 0.6, 0.7$, the schedule experiment reports sparsities of $1.5\%, 4.5\%, 5.4\%, 11.2\%$ and mean entropies of $0.20, 0.14, 0.08, 0.02$, respectively. Saving an anchor is passive and does not change the pre-trigger trajectory; these statistics therefore vary with trigger time rather than anchor time. Increasing sparsity and polarization accompany the late-trigger accuracy decline, but the measurements do not isolate its cause. The entropy of 0.08 at the default trigger differs from 0.035 in the extended entropy sweep (Table 16); we retain the measurements separately pending confirmation of their configurations and aggregation conventions.

C.6 Density-Adjusted Semantic Associations

The supplementary analysis evaluates all $\binom{10}{2} = 45$ unordered ImageWoof class pairs. External references comprise WordNet path-based taxonomy similarity, DINOv2 prototypes computed from real validation images, and symmetric confusion from a ResNet-18 trained on real ImageWoof data. Table 18 retains the reported rank correlations and permutation tests. Jointly permuting the rows and columns of an external matrix preserves the dependence induced by shared classes. Holm correction is applied across the three tests; it does not turn an association into a causal result.

Table 18: Reported semantic associations with density-adjusted mask IoU. Two-sided tests use 10,000 class-label permutations.

External reference	Spearman ρ	Raw p	Holm p
WordNet taxonomy	0.404	0.0230	0.0460
DINOv2 prototypes	0.499	0.0119	0.0357
Real-data confusion	0.540	0.0308	0.0460

The density reference in Eq. (14) is homogeneous across channels. With block-specific retention rates, a stratified approximation instead uses $\sum_l D_l r_{i,l} r_{j,l}$ divided by $\sum_l D_l (r_{i,l} + r_{j,l} - r_{i,l} r_{j,l})$. The analysis record mentions block stratification but does not provide per-block counts or establish whether adjustment preceded block aggregation. Releasing the masks, similarity matrices, embedding checkpoint, taxonomy transformation, and permutation code is necessary to reproduce the reported correlations. The homogeneous calculation explains the scale of raw overlap; it is not a reconstruction of the underlying permutation experiment.

C.7 Diversity, Coverage, and Redundancy Diagnostics

Table 19 reports an additional ImageWoof experiment at 50 IPC. The evaluation record describes a common generation and assessment protocol within this table. Its DiT and Minimax accuracies differ from the original main-table values, so these rows are kept as a separate analysis rather than substituted into Table 1.

Table 19: Additional ImageWoof diagnostics at 50 IPC. R2S, Hard R2S, and S2S are reported relative nearest-neighbor distances, with DiT normalized to 1. Smaller R2S/Hard R2S and larger S2S are treated as favorable in the evaluation record. Dup.: reported duplicate rate. Missing error bars are not inferred.

Method	Acc. (%)	Conf.	LPIPS	R2S	Hard R2S	S2S	Dup. (%)
DiT	52.4	0.320	0.695	1.00	1.00	1.00	10
Minimax	56.1	0.260	0.675	0.95	0.93	0.97	11
No rollback	54.3 ± 0.8	0.245	0.625	1.08	1.14	0.85	23
`MaskDD`	59.0 ± 0.4	0.200	0.670	0.86	0.78	1.02	8

R2S measures real-to-synthetic proximity, Hard R2S restricts that diagnostic to hard real examples, and S2S assesses within-synthetic nearest-neighbor separation. The supplied results do not specify the embedding network, distance normalization, hard-example rule, duplicate threshold, or exact confusion functional. These quantities should therefore be treated as reported diagnostics rather than fully reproducible estimators of distributional coverage. In particular, an 8% duplicate rate is meaningful only relative to the missing threshold definition, and within-synthetic redundancy does not measure memorization of training examples.

LPIPS is slightly lower than both DiT and Minimax, while the reported coverage distances are smaller. This is compatible with a tradeoff between unconstrained perceptual variability and task-relevant coverage in one setting. The results do not establish that diversity is preserved across datasets, budgets, or metrics. Synthetic-to-training nearest-neighbor inspection, a held-out comparison, and broader distributional tests would be needed to assess memorization and generalize the coverage claim.

C.8 Cross-Family Evaluation

The additional ImageNet-1K evaluation uses the same 50-IPC synthetic dataset for ResNet-18/50/101 and ViT-B/16. The ViT-B/16 result is $63.5 \pm 0.4\%$. The discussion also lists external baselines under a protocol denoted RD^3, but does not uniquely identify that protocol's source or complete settings. We therefore retain the `MaskDD` transfer measurement without treating those external values as a verified unified ranking. Training schedules, augmentation, label supervision, and independent-run counts remain to be supplied.

C.9 Sampling Efficiency

Table 20 reports nearly unchanged memory and sampling time. The numerical time difference is approximately 0.5%, without repeated timing measurements; it does not justify a robust acceleration claim. The reported active-count difference is 38.2 million parameters, but its relationship to masked activation channels, stored tensors, and executed operations is unspecified. Frozen weights reduce the number of optimized parameters, but backward computation through the frozen backbone is still required for mask gradients.

Table 20: Recorded ImageNet-100 sampling measurements at batch size 16. The parameter counts are reported as active counts, not verified allocated model sizes. Channel masking alone does not imply physical parameter removal. Hardware, warm-up, and timing variance are unspecified.

Method	Active params. (M)	Sparsity (%)	Memory (GB)	s/100 images
DiT	678.5	0.0	6.63	9.43
`MaskDD`	640.3	5.6	6.67	9.38

D Limitations and Broader Impacts

Objective and mechanism. The training loss preserves denoising ability and regularizes masks; it does not directly optimize the downstream classifier. Accuracy gains and qualitative heatmaps do not establish that background information is selectively removed. Controlled interventions, alternative regularizers, and quantitative feature analyses are needed to test that explanation.

Schedule sensitivity. Mask initialization, temperature decay, sparsity strength, and rollback timing can affect which channels remain available for refinement. A fixed early checkpoint may not provide a useful rewind state for every class or dataset. The additional rollback, affinity, and entropy sweeps cover one dataset and do not test joint hyperparameter interactions, and the available configurations contain inconsistencies described in Appendix B.

Scalability and efficiency. The class-specific logits require CLD parameters; all-pairs affinity scales quadratically in C. Grouping reduces pairwise work at the cost of omitting cross-group constraints. Activation masking also does not automatically reduce dense matrix-multiplication cost or allocated backbone memory. The current timing table is consistent with approximately unchanged sampling cost and does not establish a substantial speedup.

Evaluation coverage. Some low-budget settings favor baselines. The added coverage and semantic-association analyses are limited to ImageWoof, with metric definitions and raw artifacts still needed for full reproduction. LPIPS is lower than the dense baseline in the added comparison,

so improved accuracy should not be equated with universal diversity preservation. Recent generative baselines are not exhaustively evaluated under a common protocol. Missing protocol details and run counts prevent strong statistical conclusions. Moreover, the current record does not include a controlled domain-shift experiment, an independent replication, or an end-to-end distillation cost comparison including generator pretraining.

Broader impacts and future work. Compact synthetic datasets may reduce repeated downstream training cost, but using a pretrained generator can inherit dataset biases and memorization risks. Distillation does not provide a privacy guarantee. Future work could examine prompt-conditioned routing, overlap penalties that preserve useful shared features, and implementations that physically compact supported channel structures.

E Additional Related-Work Discussion

Structural adaptation versus sample construction. Distillation methods can act on the synthetic samples, the training signals used to construct them, or the generator. Gradient, distribution, and trajectory matching primarily define what information synthetic data should reproduce (Zhao et al., 2021; Zhao & Bilen, 2023; Cazenavette et al., 2022). Scalable non-diffusion methods include SRe^{2}L and RDED (Yin et al., 2023; Sun et al., 2024). Diffusion-based distillation adds a pretrained image prior, but its sampling and adaptation costs remain part of the pipeline (Gu et al., 2024; Su et al., 2024; Chan Santiago et al., 2025). In particular, iterative diffusion sampling should not be described as a one-pass generator unless the sampling procedure actually uses one step.

Conditional computation and concept control. Generative models also use specialization mechanisms for purposes other than distilled classifier training. eDiff-I uses denoisers specialized to different noise intervals (Balaji et al., 2022); Concept Sliders adjusts visual concepts through low-rank parameter directions (Gandikota et al., 2024); and concept erasure modifies a diffusion model to reduce specified content (Gandikota et al., 2023). These examples motivate adaptation within generative networks but do not establish the novelty or effectiveness of any particular mask schedule. The distinctive setting considered here is class-aware channel gating of a frozen DiT, with utility evaluated through classifiers trained on its synthetic images.

Scope of the lottery-ticket analogy. The original lottery ticket hypothesis concerns sparse networks trainable from particular initializations (Frankle & Carbin, 2019); weight rewinding extends the initialization used for retraining (Frankle et al., 2020). `MaskDD` instead reuses a fixed pretrained generator and rewinds its trainable gate logits. This is an algorithmic inspiration rather than a direct verification of the hypothesis in the generative setting.

F ADDITIONAL VISUALIZATIONS

Figures 5 and 6 compare generated examples across methods. Figures 7 and 8 show additional ImageNet-100 samples. These grids provide qualitative examples rather than quantitative measures of coverage, realism, or diversity.

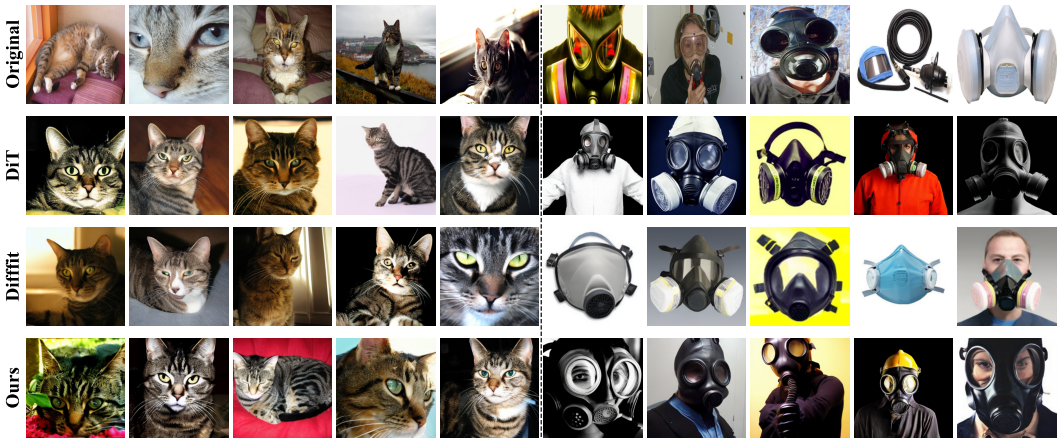

Figure 5: Real images and generated examples from DiT (Peebles & Xie, 2023), DiffFit (Xie et al., 2023), and `MaskDD`. The source description specifies matched generation seeds within each column; the dashed divider separates category groups.



Figure 6: ImageWoof examples from Minimax Diffusion (left) and `MaskDD` (right).



Figure 7: Additional `MaskDD` samples on ImageNet-100 (first grid).

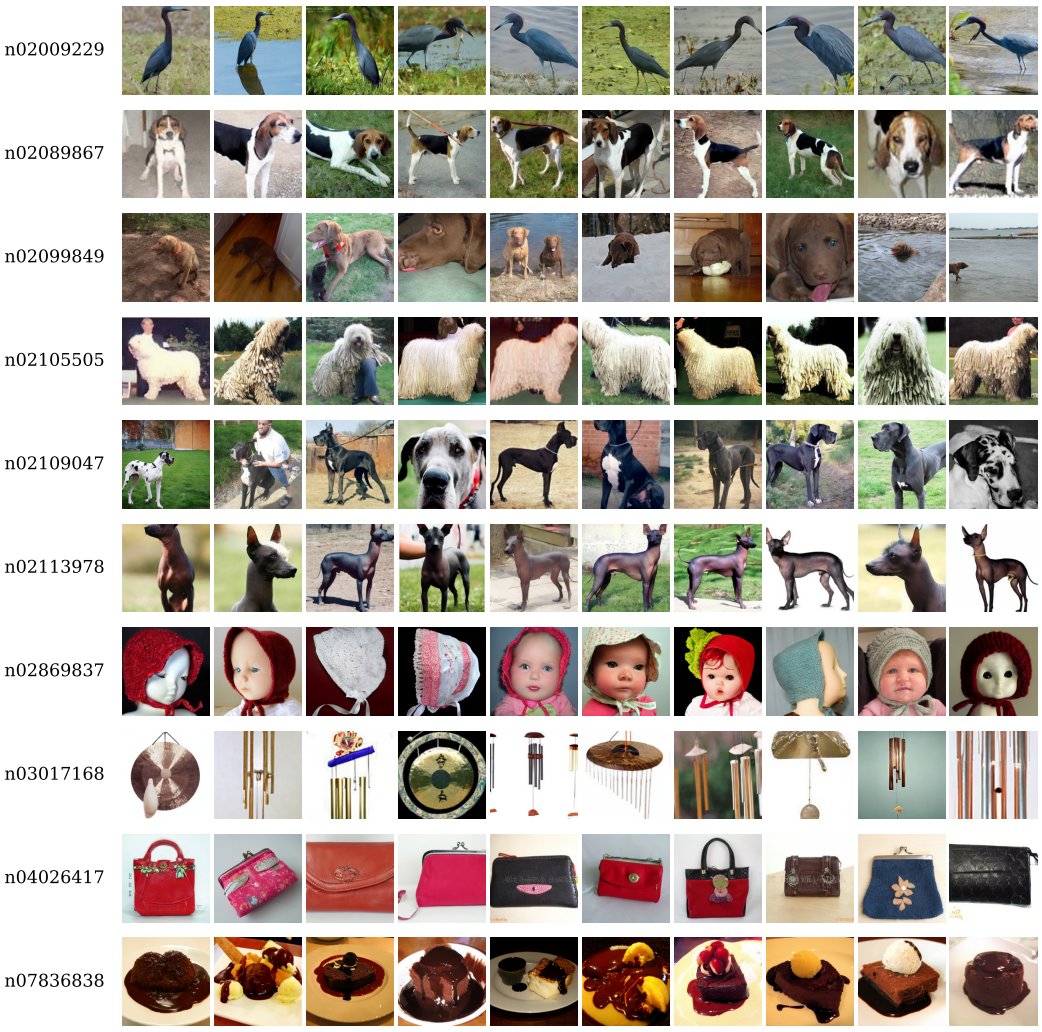

Figure 8: Additional `MaskDD` samples on ImageNet-100 (second grid).