CAG: CONTRACTION-AWARE GUIDANCE FOR DIFFUSION DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models can generate compact training sets without optimizing their parameters for each distillation task. Yet clear class identity alone does not make a small synthetic dataset informative: strong classifier-free guidance (CFG) can reduce variation among images of the same class. We introduce Contraction-Aware Guidance (CAG), an inference-time correction that complements CFG during late denoising. CAG reconstructs the predicted clean latent, decodes it with a frozen autoencoder, and increases the dispersion of its feature responses using a frozen visual encoder. A timestep gate restricts this correction to the later sampling steps, and a noise reparameterization incorporates it into the existing solver. The objective acts on individual images; its effect on dataset diversity is evaluated separately. On ImageWoof at 10 images per class, CAG improves ResNetAP-10 accuracy from 34.7% to 42.6%, while increasing intra-class LPIPS from 0.55 to 0.63 with a 0.7-point reduction in class consistency. Experiments on five ImageNet-based benchmarks examine transfer across classifier architectures and integration with existing distillation pipelines. CAG requires no additional model training, with a measured generation cost of 1.83 seconds per image on an RTX 4090.

1 INTRODUCTION

Dataset distillation seeks a compact set of training examples that retains useful information from a much larger dataset (Wang et al., 2018; Liu et al., 2022). Its success is determined by the performance of a model trained on that compact set, rather than by how convincing the images appear. With only a few images per class (IPC), each example must convey recognizable class information, and the examples together must provide complementary supervision. Repeated versions of a class prototype can satisfy the first requirement while falling short of the second.

Conventional approaches optimize synthetic images through gradients, features, or training trajectories (Zhao et al., 2021; Zhao & Bilen, 2023; Cazenavette et al., 2022). Pretrained diffusion models offer a different route: they provide a generative prior from which high-resolution, class-conditional examples can be sampled (Ho et al., 2020; Peebles & Xie, 2023). Recent distillation methods adapt this route through generator fine-tuning, mode guidance, or geometric constraints (Gu et al., 2024; Chan-Santiago et al., 2025; Roy et al., 2026). These approaches make the sampling procedure an important part of dataset construction. A generator's ability to produce realistic images does not, by itself, establish that a small sample from it is an effective training set (Ravuri & Vinyals, 2019; He et al., 2023).

Classifier-free guidance (CFG) exposes this distinction. It strengthens a class-conditional prediction relative to an unconditional prediction and generally improves conditional fidelity (Ho & Salimans, 2022). In a low-IPC regime, stronger guidance can help because an ambiguous image occupies a substantial fraction of its class's budget. However, increasing guidance can also reduce variation. Prior analyses distinguish early changes in mode selection from later concentration within modes (Jin et al., 2026). Our ImageWoof diagnostics show the same practical tension: increasing CFG from 1 to 4 raises target-class confidence from 0.3409 to 0.9310, but reduces cross-sample feature variance from 0.1491 to 0.1075. Confidence and variation therefore respond differently to stronger guidance.

This observation motivates a targeted intervention: retain the class-conditioned denoising prediction, but modify the fine-grained structure represented by its clean estimate during the later sampling

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.

steps. We propose *Contraction-Aware Guidance* (CAG), which adds a feature-dispersion correction to a frozen diffusion sampler. At an active step, CAG reconstructs the clean latent implied by CFG, decodes that estimate, and measures the variance of responses in a pretrained visual encoder. It then takes a gradient-ascent step in the clean latent and converts the corrected estimate back into the solver's noise parameterization. The generator, decoder, and encoder remain frozen.

Two distinctions are essential to this design. First, CAG increases response dispersion *within an individual image*; it does not directly maximize distances between generated images. Whether that local objective improves dataset diversity must be assessed empirically. Second, the timestep gate is predefined. It reflects a stage-specific motivation but does not detect contraction online. We therefore evaluate the method using downstream accuracy together with independent perceptual-distance evaluation and semantic diagnostics. Here, independent means that the perceptual metric is not optimized by CAG; an increase in the optimized statistic alone is insufficient evidence of success.

The results support CAG as a useful sampling intervention, with benefits that depend on the benchmark and evaluator. On ImageWoof, its means exceed the quoted DiT results in all 15 architecture–budget settings, with 11 outright leads and one tie for best among compact-data methods. Comparisons with published baselines are descriptive because protocol matching is unverified. On ImageNet-1K, Minimax+CAG exceeds Minimax by 0.9 percentage points at both budgets. These comparisons establish neither uniform superiority nor recovery of missing modes; Section 4 distinguishes their evidential scope.

Our contributions are threefold: (i) a clean-estimate feature-dispersion correction that can be inserted into an existing noise-prediction sampler without training its components; (ii) a joint evaluation of diversity, class fidelity, and training utility that distinguishes useful variation from unconstrained dispersion; and (iii) experiments across five benchmarks, together with representation ablations and an explicit accounting of inference-time cost.

2 Related work

Dataset distillation. Gradient matching (Zhao et al., 2021), distribution matching (Zhao & Bilen, 2023), and trajectory matching (Cazenavette et al., 2022) optimize small synthetic datasets to approximate aspects of learning on real data. Efficient parameterizations help scale this optimization (Kim et al., 2022). At ImageNet scale, SRe^{2}L separates squeezing, recovery, and relabeling (Yin et al., 2023), while RDED emphasizes diversity and realism (Sun et al., 2024). Their use of pretrained models and label supervision makes evaluation protocol part of the comparison: hard-label and teacher-assisted soft-label results should be interpreted separately. We retain this distinction throughout our experiments.

Diffusion-based dataset construction. Minimax Diffusion fine-tunes a generator to balance representativeness and diversity (Gu et al., 2024). D^{4}M uses a disentangled diffusion framework (Su et al., 2024), while MGD^3 guides generation toward modes extracted from real data (Chan-Santiago et al., 2025). ManifoldGD constructs hierarchical representatives and projects mode-alignment directions onto estimated local tangent spaces (Roy et al., 2026). CAG uses a different signal: the responses of a frozen encoder to the current predicted image. Its correction neither requires real-data prototypes nor explicitly estimates a tangent space. The MGD^3 integration comparison is exploratory because matched sampling protocols are unverified. When attached to a fine-tuned Minimax model, CAG adds no further training; it does not remove the training already required by that baseline.

Guidance and sampling dynamics. CFG balances conditional fidelity against diversity (Ho & Salimans, 2022). CADS perturbs conditioning to improve diversity (Sadat et al., 2024); CFG++ changes the guidance formulation to address sampling behavior (Chung et al., 2025); interval guidance restricts when guidance is applied (Kynkäänniemi et al., 2024). Stage-wise analyses explain why the effect of CFG changes during denoising (Jin et al., 2026). CAG retains CFG and gates an additional feature-based direction. Its distinguishing question is whether this late correction produces a more useful compact training set. Accordingly, scheduling and noise-injection alternatives are compared using both semantic diagnostics and downstream classifiers.

3 CONTRACTION-AWARE GUIDANCE

3.1 PROBLEM FORMULATION AND NOTATION

We generate K labeled images for each of C classes, giving a synthetic training set of CK images at IPC $= K$. Downstream classifiers are trained on this set and evaluated on real held-out images. Figure 1 summarizes CAG: it changes sampling from a pretrained diffusion prior, rather than learning that prior from the compact set.

Let z_t be the noisy latent at a sampled noise level, and write $a_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t = \sqrt{1-\bar{\alpha}_t}$. The model predicts noise $\epsilon_\theta(z_t, t, y)$ for class y, with $\varnothing$ denoting the unconditional input. CFG and its predicted clean latent are

$$\begin{aligned} \epsilon_t^{\text{cfg}} &= \epsilon_\theta(z_t, t, \varnothing) + w\big[\epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, \varnothing)\big], \\ \hat{z}_{0|t} &= \frac{z_t - \sigma_t \epsilon_t^{\text{cfg}}}{a_t}. \end{aligned} \tag{1}$$

Here w is the CFG scale. We call $\hat{z}_{0|t}$ a *clean estimate*, rather than an exact projection onto a data manifold: it is obtained algebraically from the guided model prediction. In particular, a guided prediction need not be the exact conditional posterior mean. The frozen decoder D maps this latent to an image estimate $\hat{x}_{0|t} = D(\hat{z}_{0|t})$.

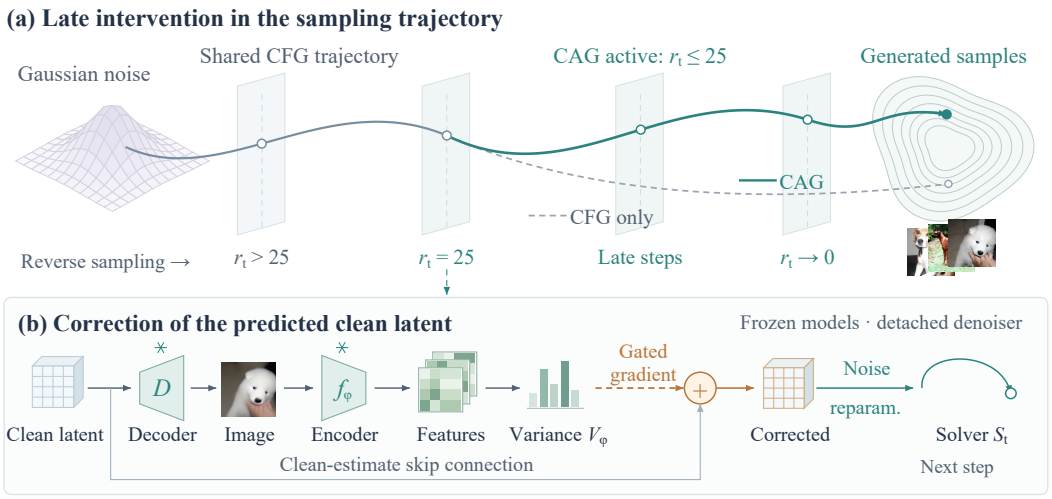

Figure 1: CAG overview. (a) CAG shares the early CFG trajectory and intervenes only after the late gate opens. (b) A feature-variance gradient corrects the clean latent; noise reparameterization feeds the original solver. Surfaces, contours, trajectories, images, and feature responses are illustrative, not measured diagnostics.

3.2 PER-IMAGE FEATURE DISPERSION OBJECTIVE

The visual encoder is trained on images, whereas z_t contains timestep-dependent noise. We therefore compute its responses on $\hat{x}_{0|t}$ rather than directly decoding the noisy state. This provides an image-space estimate of the structure currently implied by the sampler. It does not ensure that every intermediate reconstruction is natural, but avoids deliberately treating the noisy latent as a clean decoder input.

For a selected encoder layer ℓ, let $F^\ell = f_\phi^\ell(\hat{x}_{0|t})$ and flatten its channel and spatial axes into M_ℓ entries. With μ_ℓ their arithmetic mean, we define

$$V_\phi(\hat{x}_{0|t}) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{1}{M_\ell} \sum_{j=1}^{M_\ell} (F_j^\ell - \mu_\ell)^2. \tag{2}$$

The averaging normalizes for the number of responses at each layer. It does not normalize activation scales across layers, so the choice of encoder, layers, and image preprocessing affects the gradient magnitude. Our default uses VGG-16 features at layers 8, 15, and 22.

Equation (2) measures response dispersion across coordinates of a *single* image. It can increase through spatial variation, differences in channel means, or both. It is neither a direct estimate of dataset coverage nor an exact measure of high-frequency image energy. Two identical images can each have a large value of V_ϕ, so maximizing this quantity cannot by itself prevent duplicates. The motivation is more limited: a pretrained representation supplies a structured, differentiable direction for changing local appearance. Whether repeated application produces useful differences across samples is tested in Section 4.3.

3.3 Late-stage correction and sampler integration

Let r_t denote the reverse-step index in the respaced sampling schedule, descending from 49 to 0 for 50 steps. We activate the correction when $g_t = \mathbb{I}[r_t \leq r_s]$. The default $r_s = 25$ therefore selects 26 steps. This step index is distinct from the original diffusion training timestep; the coefficients a_t and σ_t are evaluated at the actual sampled noise level. At an active step, CAG performs

$$\begin{aligned} \tilde{z}_{0|t} &= \hat{z}_{0|t} + \eta g_t \nabla_{\hat{z}_{0|t}} V_\phi\big(D(\hat{z}_{0|t})\big)\,, \\ \tilde{\epsilon}_t &= \frac{z_t - a_t \tilde{z}_{0|t}}{\sigma_t}. \end{aligned} \tag{3}$$

The existing solver then computes $z_{t^-} = \mathcal{S}_t(z_t, \tilde{\epsilon}_t)$, where t^- is the next, lower noise level. The reparameterization is applied at sampled states with $\sigma_t > 0$; the final noise-free output is decoded without dividing by zero. For $g_t = 0$ or $\eta = 0$, the update reduces exactly to the baseline prediction.

The sign in Eq. (3) is gradient *ascent* on dispersion. Equivalently, it is descent on the energy $-V_\phi$. Substituting the first expression into the second gives $\tilde{\epsilon}_t = \epsilon_t^{\text{cfg}} - (a_t/\sigma_t)\eta g_t \nabla V_\phi$. Thus, CAG changes the denoising prediction through a clean-latent correction; it is not additive random noise applied after a solver step. A constant η also does not mean a constant perturbation in noise space, because a_t/σ_t varies with time.

The gate preserves the early baseline trajectory and delays intervention until the current clean estimates contain more developed structure. It also limits the number of backward passes. This is a fixed schedule, not a learned controller or a test of the local contraction rate. We use "contraction-aware" to refer to the stage-specific motivation for this choice. The method does not calculate a trajectory Jacobian or require a pair of samples during generation.

Gradient path and cost. At each active step, the clean estimate is detached from the denoiser computation and treated as the variable to differentiate. Gradients pass through D and f_ϕ to that variable; the denoiser is not differentiated, model parameters are frozen, and the graph is discarded after the update. Consequently, there is no backpropagation through the complete sampling trajectory and no optimization of a persistent synthetic dataset. There is nevertheless an inference cost: each active step adds decoding, feature extraction, and a backward pass. Algorithm 1 in Appendix A specifies this computation.

3.4 Local interpretation and scope of the correction

Locally, a sufficiently small ascent step increases the chosen objective to first order: the change in $V_\phi \circ D$ is $\eta g_t \|\nabla(V_\phi \circ D)\|^2 + O(\eta^2)$ under smoothness. That statement concerns one clean estimate. It does not imply that different trajectories separate, that class identity is preserved, or that downstream accuracy increases. The decoder, encoder, and sampler all affect the resulting image, and their nonlinearities can alter the direction of the observed change. Appendix C gives a concise conditional analysis of this distinction.

This scope guides the evaluation. We check three outcomes jointly: whether samples differ more in perceptual or feature space, whether they retain their class semantics, and whether models trained on them improve on real validation images. A useful intervention must balance these outcomes. In particular, larger perceptual distance without preserved class identity is not evidence of better distillation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks and supervision. We evaluate 256×256 synthetic images on ImageNette, ImageWoof (Howard, 2019a;b), ImageIDC (Kim et al., 2022), ImageNet-100 (Tian et al., 2020), and ImageNet-1K (Russakovsky et al., 2015). Budgets range from 10 to 100 IPC, depending on the benchmark. The first four use hard labels: classifiers are trained from scratch on the compact dataset and evaluated on real validation images. ImageNet-1K follows the Minimax soft-label evaluation configuration (Gu et al., 2024; Chan-Santiago et al., 2025), using ResNet-18 as both teacher and student; the pretrained teacher supplies targets for augmented views. Its results therefore measure the combination of synthetic images and teacher supervision, and are reported separately from the hard-label experiments. Unless stated otherwise, accuracy values are means and standard deviations over three runs; we do not interpret small mean differences as statistical significance.

Generation and comparisons. The default backbone is DiT-XL/2-256 (Peebles & Xie, 2023), with `sd-vae-ft-mse` (Stability AI, 2022), 50 sampling steps, and CFG scale $w = 4.0$. CAG uses pretrained VGG-16 (Simonyan & Zisserman, 2015), layers $8, 15, 22$, strength $\eta = 0.09$, and gate $r_s = 25$. MGD^3 uses DDPM by default and separately evaluates DDIM (Chan-Santiago et al., 2025; Song et al., 2021); sampler matching in our comparison is unverified. Published baselines in Tables 1–3 are identified by source. CAG denotes DiT+CAG unless another pipeline is named. Minimax+CAG retains its fine-tuned baseline checkpoint.

The measurements below answer three questions: does the correction improve training utility across datasets and classifiers; does it improve variation while retaining semantics; and is the effect attributable to its structured correction rather than simply weakening guidance or adding noise? All numerical results are restricted to the recorded benchmark and ablation settings. Cross-table differences should not be treated as matched-seed treatment effects. Appendix A.1 records the available protocol details and the limits of comparisons between separately reported experiments.

4.2 DOWNSTREAM CLASSIFICATION PERFORMANCE

IPC (Ratio)	Evaluator	Random	K-Center	Herding	DiT	DM	IDC-1	Minimax	MGD^3	ManifoldGD	CAG (Ours)	Full
10 (0.8%)	C-6	$24.3_{\pm 1.1}$	$19.4_{\pm 0.9}$	$26.7_{\pm 0.5}$	$34.2_{\pm 1.1}$	$26.9_{\pm 1.2}$	$33.3_{\pm 1.1}$	$\mathbf{37.0_{\pm 1.0}}$	$34.73_{\pm 1.1}$	$\underline{36.9_{\pm 0.6}}$	$35.8_{\pm 0.7}$	$86.4_{\pm 0.2}$
	RAP-10	$29.4_{\pm 0.8}$	$22.1_{\pm 0.1}$	$32.0_{\pm 0.3}$	$34.7_{\pm 0.5}$	$30.3_{\pm 1.2}$	$39.1_{\pm 0.5}$	$39.2_{\pm 1.3}$	$\underline{40.4_{\pm 1.9}}$	$38.3_{\pm 0.4}$	$\mathbf{42.6_{\pm 1.3}}$	$87.5_{\pm 0.5}$
	R-18	$27.7_{\pm 0.9}$	$21.1_{\pm 0.4}$	$30.2_{\pm 1.2}$	$34.7_{\pm 0.4}$	$33.4_{\pm 0.7}$	$37.3_{\pm 0.2}$	$37.6_{\pm 0.9}$	$38.5_{\pm 2.5}$	$\mathbf{39.2_{\pm 0.7}}$	$\underline{38.9_{\pm 1.7}}$	$89.3_{\pm 1.2}$
20 (1.6%)	C-6	$29.1_{\pm 0.7}$	$21.5_{\pm 0.8}$	$29.5_{\pm 0.3}$	$36.1_{\pm 0.8}$	$29.9_{\pm 1.0}$	$35.5_{\pm 0.8}$	$37.6_{\pm 0.2}$	$\underline{39.0_{\pm 3.46}}$	$37.7_{\pm 0.6}$	$\mathbf{39.1_{\pm 1.1}}$	$86.4_{\pm 0.2}$
	RAP-10	$32.7_{\pm 0.4}$	$25.1_{\pm 0.7}$	$34.9_{\pm 0.1}$	$41.1_{\pm 0.8}$	$35.2_{\pm 0.6}$	$43.4_{\pm 0.3}$	$\underline{45.8_{\pm 0.5}}$	$43.6_{\pm 1.6}$	$45.6_{\pm 0.5}$	$\mathbf{46.1_{\pm 1.9}}$	$87.5_{\pm 0.5}$
	R-18	$29.7_{\pm 0.5}$	$23.6_{\pm 0.3}$	$32.2_{\pm 0.6}$	$40.5_{\pm 0.5}$	$29.8_{\pm 1.7}$	$38.6_{\pm 0.2}$	$\underline{42.5_{\pm 0.6}}$	$41.9_{\pm 2.1}$	$42.4_{\pm 0.4}$	$\mathbf{42.6_{\pm 1.1}}$	$89.3_{\pm 1.2}$
50 (3.8%)	C-6	$41.3_{\pm 0.6}$	$36.5_{\pm 1.0}$	$40.3_{\pm 0.7}$	$46.5_{\pm 0.8}$	$44.4_{\pm 1.0}$	$43.9_{\pm 1.2}$	$53.9_{\pm 0.6}$	$\underline{54.5_{\pm 1.6}}$	$51.3_{\pm 0.2}$	$\mathbf{56.7_{\pm 1.3}}$	$86.4_{\pm 0.2}$
	RAP-10	$47.2_{\pm 1.3}$	$40.6_{\pm 1.0}$	$49.1_{\pm 0.7}$	$49.3_{\pm 0.2}$	$47.1_{\pm 1.1}$	$48.3_{\pm 1.0}$	$56.3_{\pm 1.0}$	$56.5_{\pm 1.9}$	$\underline{57.9_{\pm 0.6}}$	$\mathbf{58.0_{\pm 2.0}}$	$87.5_{\pm 0.5}$
	R-18	$47.9_{\pm 1.8}$	$39.6_{\pm 1.0}$	$48.3_{\pm 1.2}$	$50.1_{\pm 0.5}$	$46.2_{\pm 0.6}$	$48.3_{\pm 0.8}$	$57.1_{\pm 0.6}$	$\mathbf{58.3_{\pm 1.4}}$	$\underline{58.2_{\pm 0.6}}$	$\mathbf{58.3_{\pm 0.8}}$	$89.3_{\pm 1.2}$
70 (5.4%)	C-6	$46.3_{\pm 0.6}$	$38.6_{\pm 0.7}$	$46.2_{\pm 0.6}$	$50.1_{\pm 1.2}$	$47.5_{\pm 0.8}$	$48.9_{\pm 0.7}$	$\underline{55.7_{\pm 0.9}}$	$55.1_{\pm 2.5}$	$55.3_{\pm 0.6}$	$\mathbf{56.1_{\pm 1.4}}$	$86.4_{\pm 0.2}$
	RAP-10	$50.8_{\pm 0.6}$	$45.9_{\pm 1.5}$	$53.4_{\pm 1.4}$	$54.3_{\pm 0.9}$	$51.7_{\pm 0.8}$	$52.8_{\pm 1.8}$	$58.3_{\pm 0.2}$	$60.2_{\pm 2.4}$	$\underline{60.6_{\pm 0.1}}$	$\mathbf{61.5_{\pm 1.9}}$	$87.5_{\pm 0.5}$
	R-18	$52.1_{\pm 1.0}$	$44.6_{\pm 1.1}$	$49.7_{\pm 0.8}$	$51.5_{\pm 1.0}$	$51.9_{\pm 0.8}$	$51.1_{\pm 1.7}$	$58.8_{\pm 0.7}$	$59.7_{\pm 2.7}$	$\mathbf{61.0_{\pm 0.6}}$	$\underline{60.9_{\pm 2.3}}$	$89.3_{\pm 1.2}$
100 (7.7%)	C-6	$52.2_{\pm 0.4}$	$45.1_{\pm 0.5}$	$54.4_{\pm 1.1}$	$53.4_{\pm 0.3}$	$55.0_{\pm 1.3}$	$53.2_{\pm 0.9}$	$\underline{61.1_{\pm 0.7}}$	$60.1_{\pm 1.2}$	$60.6_{\pm 0.2}$	$\mathbf{63.2_{\pm 1.5}}$	$86.4_{\pm 0.2}$
	RAP-10	$59.4_{\pm 1.0}$	$54.8_{\pm 0.2}$	$61.7_{\pm 0.9}$	$58.3_{\pm 0.8}$	$56.4_{\pm 0.8}$	$56.1_{\pm 0.9}$	$64.5_{\pm 0.2}$	$\underline{66.5_{\pm 1.0}}$	$65.2_{\pm 0.3}$	$\mathbf{66.7_{\pm 0.8}}$	$87.5_{\pm 0.5}$
	R-18	$61.5_{\pm 1.3}$	$50.4_{\pm 0.4}$	$59.3_{\pm 0.7}$	$58.9_{\pm 1.3}$	$60.2_{\pm 1.0}$	$58.3_{\pm 1.2}$	$65.7_{\pm 0.4}$	$\underline{68.8_{\pm 0.7}}$	$66.3_{\pm 0.6}$	$\mathbf{69.8_{\pm 0.6}}$	$89.3_{\pm 1.2}$

Table 1: ImageWoof top-1 accuracy (%) at 256×256; C-6/RAP-10/R-18 denote ConvNet-6/ResNetAP-10/ResNet-18. Random, Herding, DiT, DM, IDC-1, Minimax, MGD^3, and Full are quoted from Chan-Santiago et al. (2025, Table 8). Bold/underlining mark the best/second distinct compact-data means, including ties. Protocol differences limit causal interpretation.

ImageWoof and classifier transfer. Table 1 reports five IPC budgets and three evaluators. At IPC $= 10$, CAG exceeds the published DiT means: 35.8 versus 34.2 on ConvNet-6, 42.6 versus 34.7 on ResNetAP-10, and 38.9 versus 34.7 on ResNet-18. Minimax leads on ConvNet-6 and ManifoldGD on ResNet-18. At IPC $= 50$, CAG leads on the first two evaluators and ties MGD^3 on ResNet-18; small margins do not establish significance.

CAG exceeds the listed DiT means in all 15 settings, with 11 outright leads and one tie for best. Besides the two IPC-10 exceptions, ManifoldGD leads on ResNet-18 at IPC $= 70$ (61.0% versus 60.9%). These are descriptive comparisons across recorded protocols. The full-data gap remains substantial: ResNetAP-10 at IPC $= 50$ reaches 58.0% versus 87.5% with all training data.

	IPC = 10 (0.8%)			**IPC = 20** (1.6%)		
Method	C-6	RAP-10	R-18	C-6	RAP-10	R-18
Random	$17.0_{\pm 0.3}$	$19.1_{\pm 0.4}$	$17.5_{\pm 0.5}$	$24.8_{\pm 0.2}$	$26.7_{\pm 0.5}$	$25.5_{\pm 0.3}$
Herding (Welling, 2009)	$17.2_{\pm 0.3}$	$19.8_{\pm 0.3}$	$16.1_{\pm 0.2}$	$24.3_{\pm 0.4}$	$27.6_{\pm 0.1}$	$24.7_{\pm 0.1}$
IDC-1 (Kim et al., 2022)	$\underline{24.3_{\pm 0.5}}$	$25.7_{\pm 0.1}$	$25.1_{\pm 0.2}$	$28.8_{\pm 0.3}$	$29.9_{\pm 0.2}$	$30.2_{\pm 0.2}$
MGD^3 (Chan-Santiago et al., 2025)	$23.4_{\pm 0.9}$	$25.8_{\pm 0.5}$	$23.6_{\pm 0.4}$	$\underline{30.6_{\pm 0.4}}$	$\underline{33.9_{\pm 1.1}}$	$\underline{32.6_{\pm 0.4}}$
MinimaxDiffusion (Gu et al., 2024)	$22.3_{\pm 0.5}$	$24.8_{\pm 0.2}$	$22.5_{\pm 0.3}$	$29.3_{\pm 0.4}$	$32.3_{\pm 0.1}$	$31.2_{\pm 0.1}$
CAG+ DiT	$23.1_{\pm 1.1}$	$\underline{25.9_{\pm 0.7}}$	$\underline{25.2_{\pm 0.9}}$	$28.3_{\pm 0.3}$	$31.2_{\pm 0.2}$	$32.5_{\pm 0.5}$
CAG+ MGD^3	$\mathbf{24.5_{\pm 1.7}}$	$\mathbf{26.4_{\pm 0.8}}$	$\mathbf{25.7_{\pm 2.1}}$	$\mathbf{31.0_{\pm 1.2}}$	$\mathbf{34.7_{\pm 0.6}}$	$\mathbf{33.5_{\pm 2.4}}$
Full Dataset	$79.9_{\pm 0.4}$	$80.3_{\pm 0.2}$	$81.8_{\pm 0.7}$	$79.9_{\pm 0.4}$	$80.3_{\pm 0.2}$	$81.8_{\pm 0.7}$

Table 2: ImageNet-100 accuracy (%); abbreviations follow Table 1. Non-CAG rows are quoted from Chan-Santiago et al. (2025, Table 2). Bold/underlining mark the top two means. Comparisons with CAG do not establish effects under matched sampling protocols.

Scaling to more classes. On ImageNet-100 (Table 2), CAG+MGD^3 has higher means than the quoted MGD^3 results in all six settings (0.4–2.1 points; approximately 1.0 on average). Unverified sampler and evaluation matching prevents attributing these descriptive differences solely to CAG. DiT+CAG is less competitive, especially on ConvNet-6 at IPC $= 20$.

Dataset	IPC	Random	DiT	DM	Minimax	D^{4}M	Minimax+CAG
Nette	10	$54.2_{\pm 1.6}$	$59.1_{\pm 0.7}$	$60.8_{\pm 0.6}$	$\underline{62.0_{\pm 0.2}}$	$60.4_{\pm 3.4}$	$\mathbf{62.1_{\pm 0.7}}$
	20	$63.5_{\pm 0.5}$	$64.8_{\pm 1.2}$	$66.5_{\pm 1.1}$	$\underline{66.8_{\pm 0.4}}$	$65.5_{\pm 1.2}$	$\mathbf{67.2_{\pm 1.2}}$
	50	$76.1_{\pm 1.1}$	$73.3_{\pm 0.9}$	$76.2_{\pm 0.4}$	$\underline{76.6_{\pm 0.2}}$	$73.8_{\pm 1.7}$	$\mathbf{78.8_{\pm 0.4}}$
IDC	10	$48.1_{\pm 0.8}$	$\underline{54.1_{\pm 0.4}}$	$52.8_{\pm 0.5}$	$53.1_{\pm 0.2}$	$51.1_{\pm 2.4}$	$\mathbf{54.4_{\pm 0.7}}$
	20	$52.5_{\pm 0.9}$	$58.9_{\pm 0.2}$	$58.5_{\pm 0.4}$	$\underline{59.0_{\pm 0.4}}$	$58.0_{\pm 1.4}$	$\mathbf{59.8_{\pm 1.1}}$
	50	$68.1_{\pm 0.7}$	$64.3_{\pm 0.6}$	$69.1_{\pm 0.8}$	$\underline{69.6_{\pm 0.2}}$	$64.1_{\pm 2.5}$	$\mathbf{70.5_{\pm 0.6}}$

Table 3: ImageNette/ImageIDC accuracy (%) using ResNetAP-10. Non-CAG values are quoted from Chan-Santiago et al. (2025, Table 1), using its DiT-based D^{4}M row. Bold/underlining indicate the best/second-best means.

Integration with Minimax. Minimax+CAG improves Minimax in all six settings (Table 3), by 0.1–2.2 points on ImageNette and 0.8–1.3 on ImageIDC. The smaller ImageNette changes are modest relative to run variation, supporting compatibility rather than a decisive advantage throughout. Appendix D shows matched-seed examples.

IPC	SRe^{2}L	RDED	DiT	Minimax	Minimax+CAG
10	$21.3_{\pm 0.6}$	$42.0_{\pm 0.1}$	$39.6_{\pm 0.4}$	$44.3_{\pm 0.5}$	$\mathbf{45.2_{\pm 0.8}}$
50	$46.8_{\pm 0.2}$	$56.5_{\pm 0.1}$	$52.9_{\pm 0.6}$	$58.6_{\pm 0.3}$	$\mathbf{59.5_{\pm 1.0}}$

Table 4: ImageNet-1K under the Minimax soft-label evaluation configuration (Gu et al., 2024; Chan-Santiago et al., 2025), with ResNet-18 as both teacher and student. SRe^{2}L (Yin et al., 2023) and RDED (Sun et al., 2024) are included as large-scale baselines. The highest listed mean is bold.

On ImageNet-1K (Table 4), Minimax+CAG exceeds Minimax by 0.9 points at both budgets. The generator and soft-label teacher supply information beyond the stored images; these 1,000-class results do not establish that CAG alone closes the scale gap.

4.3 GUIDANCE STRENGTH, DIVERSITY, AND SEMANTIC FIDELITY

CFG scale	Class confidence	Cross-sample variance
1	0.3409	0.1491
2	0.7994	0.1280
3	0.9177	0.1185
4	0.9310	0.1075
5	0.9454	0.1042
6	0.9315	0.1020
7	0.8871	0.1016

Table 5: Generation diagnostics for the CFG-scale sweep on ImageWoof at IPC $= 10$. A frozen ImageNet-pretrained VGG-16 supplies confidence and cross-sample feature variance. These diagnostics are reported separately from the joint utility comparison in Table 6.

Stronger guidance is not uniformly better. Table 5 isolates the generation trade-off within the CFG-scale sweep. From $w = 1$ to $w = 4$, confidence increases from 0.3409 to 0.9310 while feature variance falls by approximately 27.9%. Confidence peaks at $w = 5$ and then declines, whereas feature variance decreases throughout the sweep. Thus stronger guidance does not monotonically improve both diagnostics. This table motivates retaining a strong semantic signal while adding a separate correction; downstream utility is assessed jointly with generation quality in Table 6.

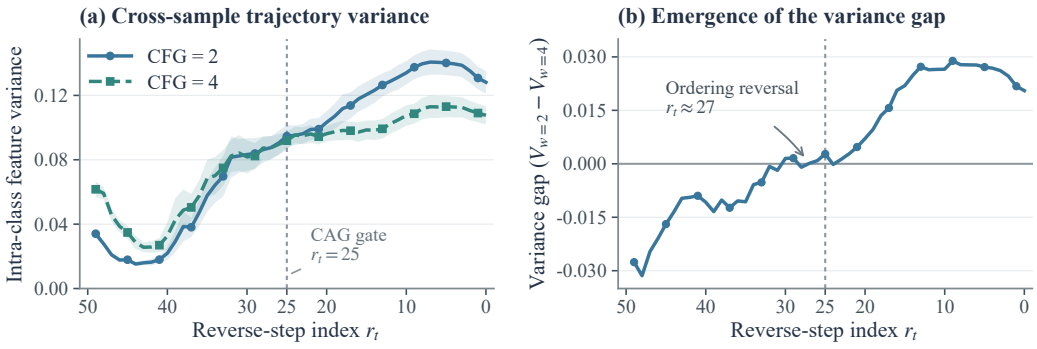

Figure 2: ImageWoof clean-estimate diagnostics recovered from the earlier vector figure. (a) Cross-sample intra-class embedding variance; shaded envelopes are retained, but their statistical interpretation is unverified. (b) Variance at CFG $= 2$ minus variance at CFG $= 4$. The ordering reverses near step 27; CAG starts at step 25. These feature-space summaries do not estimate the full sampler Jacobian.

When the variance gap emerges. Figure 2 follows the clean-estimate embeddings through sampling. Strong CFG does not begin with lower variance than moderate CFG; the ordering reverses around respaced step 27 and the gap grows toward the clean endpoint. The marked gate places CAG within this later interval. This comparison motivates the intervention window; it does not measure the effect of CAG itself. Nor does it prove that all latent-space directions contract or distinguish loss of entire modes from changes within a mode. The following controlled comparison evaluates the final dataset.

Method	LPIPS	Feature var.	Confidence	Consistency	Acc. (%)
Original CFG	0.55	0.108	0.935	97.2	$34.7_{\pm 0.5}$
Lower CFG	0.65	0.129	0.815	89.5	$35.1_{\pm 1.2}$
CFG decay	0.59	0.116	0.905	93.5	$36.3_{\pm 0.4}$
Early-stop CFG	0.61	0.120	0.885	92.0	$37.5_{\pm 0.9}$
Random noise	0.66	0.128	0.915	94.5	$37.3_{\pm 1.1}$
CAG	0.63	0.121	0.928	96.5	$42.6_{\pm 1.3}$

Table 6: Joint diversity, semantics, and utility on ImageWoof at IPC $= 10$. LPIPS and feature variance compare samples within a class; consistency is the percentage assigned to the intended class. Accuracy uses ResNetAP-10. The recorded comparison uses the same checkpoint, sampling steps, generation seeds, and evaluator protocol.

Useful variation versus raw dispersion. Table 6 compares CAG with weaker guidance, guidance scheduling, and random latent perturbations. Relative to original CFG, CAG raises intra-class LPIPS (Zhang et al., 2018) from 0.55 to 0.63 and cross-sample feature variance from 0.108 to 0.121, corresponding to relative increases of 14.5% and 12.0%. Confidence decreases by 0.007 and consistency by 0.7 percentage points. Accuracy increases by 7.9 points and exceeds the strongest alternative, early-stop CFG, by 5.1 points.

Lower CFG and random noise produce larger diversity scores than CAG but lower accuracy. Random noise has LPIPS 0.66 versus 0.63, yet gives 37.3% rather than 42.6% accuracy; CAG also retains higher class consistency. This supports useful variation within the recorded comparison, not verified distributional coverage or elimination of semantic drift. VGG-based LPIPS is an independent evaluation metric in the sense that CAG does not optimize it. It shares the VGG feature family with the correction, so it does not provide encoder-independent validation. Downstream accuracy supplies a separate utility measure; missing comparator and metric settings limit reproduction.

4.4 Feature encoder analysis and computational cost

Frozen encoder	IPC $= 10$	IPC $= 50$
VGG-16 (default) (Simonyan & Zisserman, 2015)	$42.6_{\pm 1.3}$	$58.0_{\pm 2.0}{}^{\dagger}$
VGG-19 (Simonyan & Zisserman, 2015)	$39.6_{\pm 1.1}$	$56.7_{\pm 1.3}$
ResNet-50 (He et al., 2016)	$41.7_{\pm 1.0}$	$57.1_{\pm 1.5}$
ViT-B/16 (Dosovitskiy et al., 2021)	$43.0_{\pm 1.1}$	$55.3_{\pm 1.5}$
DINOv2-B/14 (Oquab et al., 2023)	$41.5_{\pm 1.0}$	$56.5_{\pm 1.3}$

Table 7: Frozen-encoder comparison on ImageWoof, evaluated with ResNetAP-10. †The IPC-50 VGG-16 value is reused from Table 1 as a benchmark reference, not a new ablation run. Its differences from the other IPC-50 entries are descriptive, not verified paired effects.

The representation matters. In the objective-space comparison on ImageWoof at IPC $= 50$, the reported ResNet-18 accuracies are $54.4 \pm 0.6\%$ with random VGG-16 responses and $54.7 \pm 0.7\%$ with pixel variance. The pretrained VGG-16 reference from Table 1 is $58.3 \pm 0.8\%$. Its descriptive margins are 3.9 and 3.6 points. Because this reference is reused rather than a verified paired rerun, these margins do not isolate the causal effect of the representation. Pretrained features can also respond to nuisance patterns.

Table 7 includes convolutional and Transformer-based representations. At IPC $= 10$, ViT-B/16 has the highest listed mean, with accuracies ranging from 39.6 to 43.0%. At IPC $= 50$, the alternative encoders range from 55.3 to 57.1%; the reused VGG-16 benchmark reference is 58.0%. These records indicate compatibility with several representations but do not establish a controlled IPC-50 encoder ranking. Encoder-specific activation scales, preprocessing, and feature selection remain important implementation considerations.

Method	Seconds/image	Peak GiB	Backward steps	Est. 500 images (min)
CFG	0.99	3.10	0	8.25
CAG	1.83	4.40	26	15.25

Table 8: Generation cost on one RTX 4090, averaged over 50 images after five warm-up images. The final column extrapolates $500 \times$ seconds/image$/60$ from rounded timings; it is not a separate end-to-end measurement.

Inference overhead. Table 8 reports 0.99 seconds per image and 3.10 GiB peak memory for CFG, versus 1.83 seconds and 4.40 GiB for CAG: approximately $1.85\times$ the runtime and 41.9% more memory. Timing covers reverse sampling and final decoding, excluding model loading and file writing. CAG activates backward computation at 26 of 50 steps (52%). Generating 500 images is estimated to take 15.25 minutes. This overhead is confined to dataset construction; downstream training uses stored images without CAG-specific computation.

Scope of cost evaluation. The comparison fixes denoising steps and excludes generator/encoder pretraining, Minimax fine-tuning, and classifier training. Matching IPC likewise fixes dataset size, not total compute or access to pretrained information. CAG adds generation work to produce a reusable dataset. Whether this is preferable to more baseline sampling, sample selection, or another solver requires a separate comparison at equal end-to-end compute.

4.5 Implications for compact dataset generation

The diagnostic and utility results suggest a distinction between increasing variation and making that variation useful for learning. In Table 6, lower CFG and random perturbations produce greater perceptual dispersion than CAG, whereas CAG yields higher classification accuracy and class consistency. This pattern motivates assessing variation together with semantic fidelity and downstream utility. The trajectory analysis addresses a different design question: when to intervene. Its late-emerging variance gap motivates preserving the early guided trajectory and applying the feature correction after the clean estimate has developed more structure.

The benchmark and integration results address complementary aspects of this design. Evaluation with multiple classifiers examines whether generated images support different downstream learners, while the Minimax integration examines the correction within a fine-tuned generation pipeline. These comparisons place CAG at the sampling stage, where it can modify a fixed generator's output without changing the classifier training procedure. Their protocol qualifications are stated alongside the corresponding tables and in Appendix A.1.

This separation also clarifies the practical role of generation cost. Once images are stored, they can be reused across downstream training runs without repeating the correction. The added work is therefore attached to constructing the dataset rather than to each classifier update. For applications that reuse a compact dataset, both generation cost and downstream utility are relevant evaluation criteria.

5 Conclusion

We presented CAG, an inference-time correction for diffusion-based dataset distillation. CAG combines a per-image feature dispersion objective with a late sampling gate, then incorporates the corrected clean estimate into the existing sampler through noise reparameterization. It keeps pretrained components frozen and requires no pairwise sample interactions. The reported experiments support its utility for compact classification datasets, with the joint evaluation showing a favorable balance between sample variation, semantic fidelity, and downstream accuracy.

The present evaluation focuses on ImageNet-derived classification tasks and a fixed correction schedule. Extending the approach to other domains and adaptive schedules is a natural direction for future work. More broadly, the results motivate studying inference-time feature corrections as a means of adapting pretrained generators to the training needs of compact datasets.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and revision, including language editing, organization, LaTeX formatting, and checking the consistency of the written presentation of methods and reported results. The authors have manually reviewed and checked the AI-assisted writing. The authors take full responsibility for the final manuscript, including the accuracy of its text, claims, citations, and reported results.

REFERENCES

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proc. CVPR*, 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*, volume 267, pp. 52861–52876, 2025. URL https://proceedings.mlr.press/v267/santiago25a.html.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *Proc. ICLR*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nUmCcZ5RKF.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 2020.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, 2019a. URL https://github.com/fastai/imagenette#imagenette. Official dataset repository; accessed 2026-09-26.

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. Official dataset repository; accessed 2026-09-26.

Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. In *Proc. ICLR*, 2026.

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.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Proc. NeurIPS*, 2024.

Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. *Advances in neural information processing systems*, 35:1100–1113, 2022.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023. URL https://github.com/facebookresearch/dinov2. Author-provided reference and official model repository.

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

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fcf55a303b71b84d326fb1d06e332a26-Abstract.html.

Ayush Roy, Wei-Yang Alex Lee, Rudrasis Chakraborty, and Vishnu Suresh Lokhande. Manifoldgd: Training-free hierarchical manifold guidance for diffusion-based dataset distillation. *arXiv preprint arXiv:2602.23295*, 2026.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *Proc. ICLR*, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. URL https://www.robots.ox.ac.uk/~vgg/research/very_deep/.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Stability AI. sd-vae-ft-mse: Fine-tuned autoencoder, 2022. URL https://huggingface.co/stabilityai/sd-vae-ft-mse. Official model card; accessed 2026-09-26.

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*, 2024.

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

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.

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

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*, 2021. URL https://openreview.net/forum?id=mSAKhLYLSsl.

A IMPLEMENTATION AND REPORTING DETAILS

Algorithm 1 CAG for a noise-prediction diffusion sampler

Require: Frozen denoiser ϵ_θ, decoder D, encoder f_ϕ; class y; guidance w; correction η; gate r_s; solver $\mathcal{S}$

Sample initial latent z from the baseline initialization
for respaced step $r = 49, 48, \ldots, 0$ **do**
  Obtain the actual noise level t and schedule coefficients a_t, σ_t
  Compute ϵ^{cfg} and $\hat{z}_0$ using Eq. (1), without a denoiser gradient graph
  if $r \leq r_s$ **then**
    $u \leftarrow \text{stopgrad}(\hat{z}_0)$; enable gradients for u
    $v \leftarrow V_\phi(D(u))$ using Eq. (2)
    $d \leftarrow \nabla_u v$; $\tilde{z}_0 \leftarrow \text{stopgrad}(u + \eta d)$
    $\tilde{\epsilon} \leftarrow (z - a_t \tilde{z}_0)/\sigma_t$
  else
    $\tilde{\epsilon} \leftarrow \epsilon^{\text{cfg}}$
  end if
  $z \leftarrow \mathcal{S}_t(z, \tilde{\epsilon})$; discard the current gradient graph
end for
return $D(z)$ at the clean endpoint

The reverse-step index controls the gate; it is not substituted into the original diffusion noise schedule. The update uses the schedule coefficients associated with the sampler's actual timestep. For a stochastic solver, its random draws remain part of the baseline solver. Paired comparisons hold those draws and initial latents fixed. Differentiating through the decoder and encoder requires retaining their activation graph for the current step even though their parameters are frozen. The denoiser prediction and graphs from earlier steps are not retained.

A.1 REPORTED CONFIGURATIONS AND COMPARISON SCOPE

The reported default is DiT-XL/2-256 with `sd-vae-ft-mse`, 50 sampling steps, $w = 4$, VGG-16 layers $8, 15, 22$, $\eta = 0.09$, and $r_s = 25$. The original experimental description mentions some dataset–IPC-specific parameter adjustments but does not enumerate them. Exact per-setting overrides, encoder normalization and layer-index conventions, generation seeds, downstream optimizer and augmentation settings, training length, and the distinction between generation and classifier seeds are not fully specified in the available record. For ImageNet-1K, both teacher and student are ResNet-18, matching the Minimax evaluation configuration. The remaining omissions limit exact reproduction and should not be inferred from the defaults.

Benchmark results and reused references have distinct roles. The ImageWoof IPC-50 CAG references are $58.0 \pm 2.0\%$ for ResNetAP-10 and $58.3 \pm 0.8\%$ for ResNet-18, both taken from Table 1. Reuse in an auxiliary comparison is marked with a dagger. These entries are not additional experimental replicates and are not asserted to share seeds or every setting with the comparator entries. Only the explicitly paired IPC-10 comparison in Table 6 is used for the corresponding joint diversity, semantics, and utility analysis. The CFG-scale table reports generation diagnostics without attaching a conflicting classifier-accuracy series.

MGD^3 comparison provenance. The MGD^3 ImageNet-100 baseline is quoted from Chan-Santiago et al. (2025, Table 2), whose default sampler is DDPM. The available CAG integration description specifies DDIM, but its effective configuration is unverified. The source also evaluates DDIM separately on ImageNette; this does not establish a matched ImageNet-100 control. Moreover, its ImageWoof experiment trains DiT on ImageWoof. Quoted baselines are literature comparisons, not verified paired reruns; their differences cannot isolate CAG's effect.

The cost analysis uses 0.99 and 1.83 seconds per image: a $1.85\times$ ratio and 0.84-second increase. Linear extrapolation to 500 images gives 8.25 and 15.25 minutes, excluding loading and file writing. These are throughput-based estimates, not full-dataset measurements or scaling guarantees for other batch sizes or devices.

B ADDITIONAL ANALYSES AND ABLATIONS

The complete ImageWoof comparison appears in Table 1. Averaging cell-wise differences over its 15 architecture–budget settings gives 6.84 points over DiT, 1.68 over Minimax, and 1.23 over MGD^3. The latter two use the quoted values from Chan-Santiago et al. (2025, Table 8), including its original numerical precision. These equally weighted summaries are neither statistical tests nor verified paired effects.

B.1 ALTERNATIVE SAMPLING STRATEGIES AT HIGHER IPC

Method	IPC=10	IPC=50
Original CFG	$34.7_{\pm 0.5}$	$49.3_{\pm 0.2}$
Lower CFG	$35.1_{\pm 1.2}$	$50.4_{\pm 0.7}$
CFG Decay	$36.3_{\pm 0.4}$	$\underline{55.5_{\pm 1.9}}$
Early-stop CFG	$\underline{37.5_{\pm 0.9}}$	$54.6_{\pm 1.5}$
Random Noise Injection	$37.3_{\pm 1.1}$	$55.3_{\pm 2.1}$
Full CAG	$\mathbf{42.6_{\pm 1.3}}$	$58.0^{\dagger}_{\pm 2.0}$

Table 9: Comparison with simple alternatives on ImageWoof using the same DiT-XL/2 checkpoint and 50-step sampler. ResNetAP-10 Top-1 accuracy (%) on the real validation set is reported as mean $\pm$ standard deviation over three seeds after training on each generated dataset. †The IPC-50 CAG entry is reused from Table 1 as a benchmark reference, not a new paired run.

Table 9 repeats the IPC-10 accuracies from Table 6 and adds IPC $= 50$. At the larger budget, CFG decay reaches 55.5%, compared with the reused CAG reference of 58.0%. This 2.5-point gap is descriptive, not a verified paired effect. The exact guidance scales, schedules, stopping point, and noise magnitude are unspecified, so better tuned alternatives are not ruled out.

B.2 COMPARISON OF OPTIMIZATION SPACES

Objective space	Accuracy (%)
Random VGG-16 variance	54.4 ± 0.6
Pixel variance	$\underline{54.7 \pm 0.7}$
Pretrained VGG-16 reference	$58.3 \pm 0.8^{\dagger}$

Table 10: Objective-space comparison on ImageWoof at IPC $= 50$, evaluated with ResNet-18. †The pretrained VGG-16 reference is reused from Table 1; it is not a new paired ablation run. Means and standard deviations are reported, and cross-row differences are descriptive.

In Table 10, the reused pretrained VGG-16 reference exceeds the random-feature and pixel-space entries by 3.9 and 3.6 points. These descriptive differences do not isolate encoder pretraining, activation scale, semantic structure, or unreported configuration changes.

B.3 Exploratory analysis of correction timing and strength

Timestep Threshold		Correction Strength	
r_s	Mean accuracy (%)	η	Mean accuracy (%)
5	32.37	0.00	34.37
10	35.10	0.02	35.97
15	35.55	0.04	36.57
20	37.10	0.06	37.97
25	38.47	0.08	38.03
30	37.58	0.10	38.47
49	30.82	0.12	38.00

Table 11: Exploratory mean-accuracy sweeps on ImageWoof at IPC $= 10$. The gate r_s is varied with $\eta = 0.10$, and η with $r_s = 25$. Standard deviations are omitted because the available summaries disagree at the shared setting. These results are not used to claim statistical significance.

Table 11 favors an intermediate gate and moderate strength: $r_s = 25$ exceeds 5 and 49, and $\eta = 0.10$ exceeds 0 and 0.12. The shared mean is 38.47%, but its standard deviation is inconsistent and the evaluator and full configuration are unspecified. These exploratory results neither establish significance nor override the default $\eta = 0.09$. Incompletely specified projection comparisons are excluded from accuracy claims.

C Local interpretation and its limits

These identities describe the objective and its one-step effect, without establishing global contraction reversal or recovery of the real-data distribution.

What feature variance contains. For a tensor with C channels and N spatial positions, let μ_c and μ denote channel and overall means. Expanding $F_{c,p} - \mu = (F_{c,p} - \mu_c) + (\mu_c - \mu)$ and summing the zero cross terms gives

$$\mathrm{Var}_{c,p}(F_{c,p}) = \frac{1}{C}\sum_c \mathrm{Var}_p(F_{c,p}) + \mathrm{Var}_c(\mu_c). \tag{4}$$

Either component can increase independently. Parseval's identity relates the first to non-DC feature energy, not exclusively high frequencies. With a nonlinear decoder and encoder, Eq. (2) measures feature dispersion rather than a fixed output-image Fourier band.

Transport through a deterministic DDIM step. Consider deterministic DDIM without additional clipping. Its next step is $z_{t^-} = a_{t^-}\hat{z}_{0|t} + \sigma_{t^-}\epsilon_t$. At a fixed input z_t, define $U = V_\phi \circ D$ and $d_t = \nabla U(\hat{z}_{0|t})$. Substitution of Eq. (3) gives the exact difference

$$z_{t^-}^{\mathrm{CAG}} - z_{t^-}^{\mathrm{CFG}} = \eta g_t \left(a_{t^-} - \frac{a_t \sigma_{t^-}}{\sigma_t}\right) d_t. \tag{5}$$

The coefficient is positive when $a_{t^-}/\sigma_{t^-} > a_t/\sigma_t$, with the clean endpoint included by continuity. The identity applies at a shared input; later trajectories may diverge. Clipping and nonlinear solver modifications require separate analysis.

Pairwise separation requires an additional condition. For two states at the same timestep, let Δ be their baseline next-state difference, $d_i - d_j$ their gradient difference, and $b_t = a_{t^-} - a_t\sigma_{t^-}/\sigma_t$. Under the same DDIM assumptions, the squared-distance change is

$$\begin{aligned} &\|\Delta + \eta g_t b_t (d_i - d_j)\|^2 - \|\Delta\|^2 \\ &\quad = 2\eta g_t b_t \langle \Delta, d_i - d_j\rangle + \eta^2 g_t^2 b_t^2 \|d_i - d_j\|^2. \end{aligned} \tag{6}$$

For small η and positive b_t, the first-order sign depends on alignment with Δ, which per-image variance ascent does not guarantee. Greater one-step separation alone establishes neither semantic fidelity nor training utility.

D MATCHED-SEED QUALITATIVE COMPARISONS

Figures 3 and 4 show ImageNette comparisons at IPC $= 10$. Within each baseline–CAG pair, the class condition, initial latent, and reverse-solver random seed are held fixed. This pairing helps attribute visible changes to the correction; selected examples cannot quantify distributional coverage or artifact frequency. Quantitative results appear in the main text.



Figure 3: Matched-seed ImageNette examples from DiT (left) and DiT+CAG (right). Rows identify classes and columns identify samples. Corresponding positions show the same sampling seed. These examples illustrate changes in appearance, background, and structure; they do not establish semantic preservation across the generated distribution.

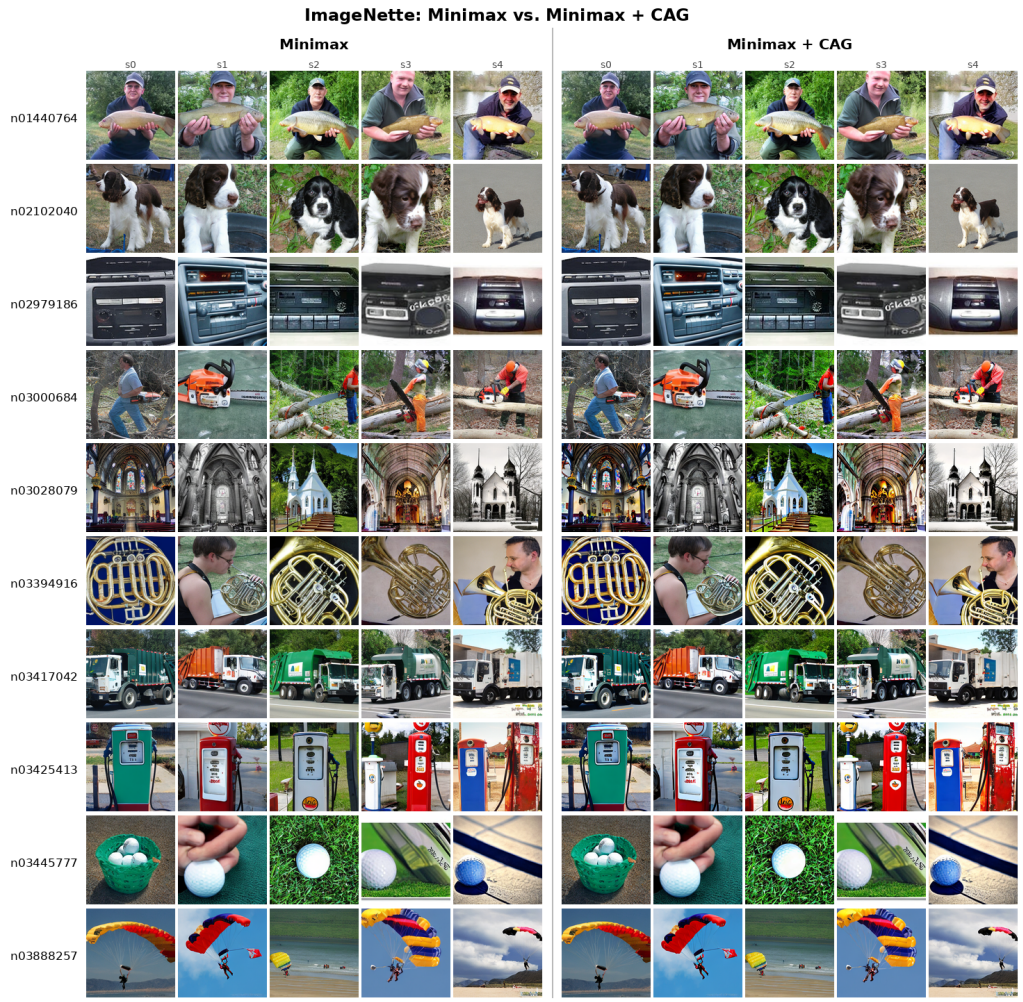

Figure 4: Matched-seed ImageNette examples from Minimax (left) and Minimax+CAG (right), arranged as in Figure 3. Each pair uses the same baseline checkpoint. CAG applies an inference-time correction without additional model training.

E Resource and metric provenance

References identify the official ImageNette/ImageWoof repository, the VGG, ResNet, ViT, and DINOv2 methods, and the `sd-vae-ft-mse` model card. Resource links do not identify exact releases or checkpoint revisions used in each run. Compression ratios follow the cited tables; their denominators have not been independently verified. No public default is substituted for missing experimental settings.

Figure 2 was recovered from an earlier vector figure without smoothing. Its shaded envelopes preserve that figure's geometry; no per-class measurements are available to verify their interpretation as standard errors. LPIPS uses a VGG backbone and is evaluated separately from the CAG objective. Its version, pair sampling, feature normalization, classifier-logit mapping, sample counts, and aggregation order are not fully documented. The diagnostic values are reported summaries, not independently recomputed metrics. Equation (2) defines the single-image optimization objective, not the cross-sample diagnostic. Pairing claims describe the recorded protocol and have not been checked against generation manifests.