DENSITY IS NOT TRAINING UTILITY: A LEARNING-RESPONSE ANALYSIS OF SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Visually realistic synthetic images do not necessarily improve downstream performance, and evaluating each synthetic dataset through retraining is costly. We study this problem through the *learning response*: how training on synthetic samples changes predictions on real target samples. Our analysis relates utility to whether these changes correct existing target errors and to the curvature of the target loss. We show that utility depends on current data coverage, and that image distributions alone cannot determine utility rankings across tasks. We propose Interventional Godambe Risk Tomography (IGRT) to predict changes in target negative log-likelihood without retraining for each candidate set. Using a reference model fitted on real data, IGRT combines local approximations of parameter shifts with uncertainty in fitted parameters due to data sampling. Controlled experiments show that its predictions capture the effects of candidate labels, current data coverage, and augmentation size. Across multiple downstream tasks, IGRT predicts the direction and magnitude of augmentation effects and selects subsets that improve performance over distribution-based selection strategies and model-based data valuation methods under matched data and training budgets.

1 INTRODUCTION

Recent image generators produce increasingly realistic images and offer greater control over their content (Rombach et al., 2022; Xie et al., 2025a;b; Z-Image Team, 2025; Zhang et al., 2023). This makes synthetic data attractive for expanding training sets when real examples are scarce or unevenly distributed across classes and visual conditions (Zhao et al., 2024; Dunlap et al., 2023). Yet improvements in visual realism do not consistently translate into better performance on real recognition tasks (Adamkiewicz et al., 2026; Fan et al., 2024). Although synthetic augmentation can help (He et al., 2023; Azizi et al., 2023), targeted retrieval of real images can still yield larger gains (Geng et al., 2024; Burg et al., 2023). Understanding when generated images become useful training data is therefore important for deciding whether and how to use them. The benefit of synthetic data depends on what the current learning problem requires (Li et al., 2025; Nguyen et al., 2026). Realistic, correctly labeled images may repeat information already covered by the training set (Li et al., 2025), while label errors can encourage predictions that conflict with the target task (Dunlap et al., 2023; Geng et al., 2024). Visual quality alone therefore does not establish whether additional data will help a particular learner. The relevant question is whether learning from those data improves predictions on real target samples, given the existing training data and the task to be solved.

Common evaluation methods provide only part of the evidence needed for this decision. FID, feature density, and precision–recall characterize fidelity and coverage (Heusel et al., 2017; Naeem et al., 2020); covariance matching uses distributional structure to guide selection (Rezaei et al., 2026). These measurements do not directly quantify the change in target performance caused by adding synthetic data under a specified training rule.

We assume access to a labeled current set D, a labeled synthetic candidate set S, and real target data T representing the intended application. The training rule $\mathcal{A}$ specifies the learner, data weights, regularization, optimizer, and training budget. We define the utility $U_{\mathcal{A}}(S \mid D)$ (Equation 1) of S as the decrease in expected target loss when the same rule is applied to the augmented data instead of D

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

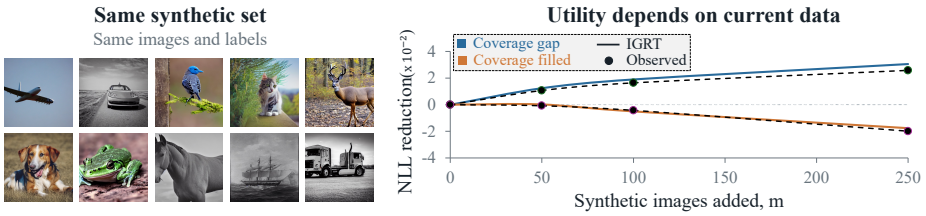

Figure 1: **The same synthetic data can help or hurt depending on current data coverage.** Coverage gap underrepresents a region within each class; Coverage filled increases coverage of that region by replacing real samples while keeping class counts fixed. Both current sets contain 1,000 real images and receive identical synthetic additions. Solid lines show IGRT predictions; black points show measured test NLL reductions, connected by dashed lines. Positive values indicate benefit.

alone. We use negative log-likelihood (NLL) as the target loss. Positive (negative) utility indicates decreased (increased) target loss. Our central question is:

Can we predict the sign and magnitude of $U_{\mathcal{A}}(S \mid D)$ *without retraining for each candidate set?* (Q)

To answer (Q), we study the learning response: how training on synthetic samples changes predictions on real target samples. The training signal on the synthetic samples is observable, but the resulting change in target predictions, ΔF_T, is not known before training. A local expansion of target NLL relates utility to the alignment of ΔF_T with existing target errors and to a quadratic term determined by target-loss curvature. This distinguishes updates that correct target errors from those that are misdirected or too large. We analyze this response in a probabilistic reference model with fixed features. Augmentation is represented by $P_\alpha = (1-\alpha)P_D + \alpha P_S$, where α is the synthetic-data weight, allowing differences in both image statistics and label relationships. Under smoothness and local regularity conditions, we derive an expansion of expected target loss that separates systematic parameter shifts from variation due to data sampling. The analysis explains how the same synthetic data can supply a missing learning signal for one current set but little additional benefit for another.

We therefore develop **Interventional Godambe Risk Tomography** (IGRT) to estimate these effects without retraining for each candidate set. At a reference model fitted on real data, mean gradients provide update directions, loss Hessians determine the local response, and gradient covariances describe sampling variability (Godambe, 1960; White, 1982; Müller, 2013). IGRT combines these statistics at different augmentation weights to approximate the current-only and augmented fits, then evaluates their difference through the gradient and curvature of target NLL. Figure 1 compares these predictions with measured gains from the same synthetic sets under different current-data coverage. The additions improve test NLL in the gap condition but worsen it in the filled condition. Influence functions similarly connect changes in training-example weights to changes in test loss (Koh & Liang, 2017). Compared with the first-order influence approximation, IGRT uses candidate-mixture curvature and retains quadratic target-loss and sampling-variance terms. This yields estimates for comparing synthetic datasets and choosing addition counts. For interpretation, a reference-intervention analysis relates changes in the measured statistics to known label and coverage interventions.

We evaluate IGRT across natural-image, testing both utility prediction and data selection under multiple downstream learners. We first compare predicted changes in target NLL with those measured after actual downstream training. We then use the predictions to select synthetic subsets and assess their downstream performance under matched data and training budgets. These experiments test whether IGRT can predict augmentation effects and use those predictions to identify useful training data across tasks and domains.

Our contributions are threefold:

- **A learning-response analysis of synthetic data utility.** We connect augmentation utility to changes in real target predictions and derive a local expansion separating systematic parameter shifts from data-sampling variability (Theorem 1). The analysis shows how current data coverage affects utility (Proposition 1) and why image marginals alone cannot determine utility rankings for every target task (Proposition 2).

- **Utility estimation without retraining for each candidate set.** We develop IGRT to predicts target NLL changes across synthetic datasets and augmentation sizes using a reference model fitted on real data. It combines mixture gradient statistics with training and target-loss curvature to capture parameter shifts and sampling variability (Section 4.3).
- **Predictive and practical validation across downstream tasks.** Across classification and segmentation tasks with multiple learners and generators, IGRT predicts the direction and magnitude of augmentation effects. Under matched data and training budgets, its selected subsets improve downstream performance over distribution-based selection and model-based data valuation methods. Controlled comparisons test how candidate labels, current data coverage, and augmentation size affect these predictions (Section 5).

2 Related Work

Synthetic data quality and utility. FID, KID, CMMD, and MAUVE measure distributional discrepancies (Heusel et al., 2017; Bińkowski et al., 2018; Jayasumana et al., 2024; Pillutla et al., 2023), while precision–recall and density–coverage distinguish fidelity from coverage (Kynkäänniemi et al., 2019; Naeem et al., 2020). Sample-level auditing additionally examines copying of training examples (Alaa et al., 2022). For training-set construction, prompting, uncertainty-guided sampling, and consistency filtering improve relevance and label reliability (Lei et al., 2025; Luo et al., 2024; He et al., 2023; Zhou et al., 2026), while adaptive generation targets the current learner's needs (Li et al., 2025; Nguyen et al., 2026). However, scaling studies and comparisons with real-image retrieval show that improvements in synthesis do not consistently produce larger training gains (Fan et al., 2024; Geng et al., 2024; Burg et al., 2023). Analyses of joint real–synthetic training relate these outcomes to distribution mismatch. Dohmatob et al. (2025) characterize the additional test error caused by differences between real and synthetic labeling functions. Under high-dimensional regression assumptions, Rezaei et al. (2026) analyze distribution shifts and derive covariance matching as a selection principle. These results link selection to learning outcomes, not image quality alone.

Data valuation and risk estimation. Data Shapley and Beta Shapley aggregate marginal predictive gains over training subsets (Ghorbani & Zou, 2019; Kwon & Zou, 2022). Distributional Shapley extends this view to underlying data distributions (Ghorbani et al., 2020), with regression analyses relating value to feature geometry, label residuals, and sample size (Kwon et al., 2021). Influence functions approximate test-loss changes under sample reweighting (Koh & Liang, 2017), with related methods addressing group perturbations, second-order effects, and resampling (Koh et al., 2019; Basu et al., 2020; Giordano et al., 2019). TracIn, Datamodels, and TRAK use gradients or subset-training outcomes for data attribution (Pruthi et al., 2020; Ilyas et al., 2022; Park et al., 2023). Related selection methods optimize validation objectives or match gradients (Killamsetty et al., 2021b;a). For expected-risk analysis, misspecified-model theory describes parameter variability through sandwich covariance, retaining the distinction between gradient covariance and loss curvature (Godambe, 1960; Huber, 1967; White, 1982). Shimodaira (2000) derives expected target-loss expansions under covariate shift and model misspecification, while active learning uses training and target information matrices to guide data acquisition (Chaudhuri et al., 2015).

3 Preliminaries

We formalize (Q) and introduce the reference learning problem used to estimate augmentation effects.

Data and utility. Let $D = \{z_i\}_{i=1}^n$ be the current set and $S = \{z_j^s\}_{j=1}^m$ a synthetic candidate set to be added. Each observation $z = (x, y)$ contains an image and its supervision. The distributions P_D, P_S, P_T describe the current, candidate, and real target data, respectively. For loss $\ell(z;\theta) = -\log p_\theta(y \mid x)$, write $L_P(\theta) = \mathbb{E}_P \ell(z;\theta)$ and $R_T = L_{P_T}$. For training rule $\mathcal{A}$, define

$$U_{\mathcal{A}}(S \mid D) = \mathbb{E}_\omega[R_T\{\mathcal{A}(D;\omega)\} - R_T\{\mathcal{A}(D \cup S;\omega)\}], \tag{1}$$

where ω denotes training randomness and positive utility indicates lower target loss. Given a labeled target set T for the intended task, we estimate utility without retraining $\mathcal{A}$ for each candidate set.

Reference learner and augmentation. We analyze unpenalized empirical risk minimization for the reference model $p_\theta(y \mid \phi(x))$, with fixed features ϕ, identifiable parameters $\theta \in \mathbb{R}^p$, and

possible model misspecification. For nonempty B, define $\widehat{L}_B(\theta) = |B|^{-1}\sum_{z\in B}\ell(z;\theta)$. Let $\theta_P \in \arg\min_\theta L_P(\theta)$ and $\widehat{\theta}_B \in \arg\min_\theta \widehat{L}_B(\theta)$, with $\theta_T = \theta_{P_T}$. Write $\widehat{\theta}_{P,n_P}$ for a fit to n_P observations from P. For $m > 0$, equal-weight addition gives $\alpha = m/(n+m)$, $n_\alpha = n+m$, and

$$\begin{aligned} P_\alpha &= (1-\alpha)P_D + \alpha P_S, \\ \widehat{\theta}_{D\cup S} &\in \arg\min_\theta \big\{(1-\alpha)\widehat{L}_D(\theta) + \alpha\widehat{L}_S(\theta)\big\}. \end{aligned} \tag{2}$$

For $m = 0$, the fit is $\widehat{\theta}_D$. The reference utility averages over new training-data draws:

$$\mathcal{V}_\alpha(P_S; P_D, P_T) = \mathbb{E}_{D',S'|\pi}\left[R_T(\widehat{\theta}_{D'}) - R_T(\widehat{\theta}_{D'\cup S'})\right], \tag{3}$$

where π fixes the source counts n, m and samples observations independently within and across sources. Thus, $\mathcal{V}_\alpha$ concerns data-sampling variability under reference ERM, whereas $U_{\mathcal{A}}$ concerns training randomness for fixed datasets. We derive the score from the former and evaluate its ability to predict the latter in the augmentation experiments.

Local quantities and assumptions. At a common reference point θ_0, define

$$\begin{aligned} s(z) &= \nabla_\theta \ell(z;\theta_0), \quad & h(z) &= \nabla^2_\theta \ell(z;\theta_0), \\ b_P &= \mathbb{E}_P s, & H_P &= \mathbb{E}_P h, \\ M_P &= \mathbb{E}_P[ss^\top], & J_P &= M_P - b_P b_P^\top. \end{aligned} \tag{4}$$

The mean gradient b_P describes the training signal at the reference point, and H_P relates this signal to a local parameter change. The score covariance J_P measures variation across observations. The target derivatives $g_T = \nabla R_T(\theta_0)$ and $G_T = \nabla^2 R_T(\theta_0)$ determine how parameter changes affect target loss. The effective score covariance $\mathcal{J}_{P,\pi}$ additionally accounts for the sampling design; it equals J_P for a single i.i.d. source. The population analysis uses $\theta_0 = \theta_T$, where $g_T = 0$. Section 4.3 replaces this point and its moments with empirical estimates.

We assume bounded third derivatives in a neighborhood $\mathcal{N}$ of θ_T, uniformly bounded $\|H_P^{-1}\|$, and moment-controlled asymptotic linearity of the reference estimator. Along the local sequence, $\theta_P \in \mathcal{N}$ and $\|b_P\| + n_P^{-1/2} \to 0$. The estimator and remainder moment conditions are stated in Appendix B.3.

4 Theoretical Results

We connect utility to target-prediction changes and derive a local target-risk expansion. We examine its dependence on current data and task and use it to construct IGRT.

4.1 Learning response and change in target risk

From prediction changes to utility. For classification, fix the current set, candidate set, and training randomness. Let $F_0(T)$ denote the current-only logits on q target examples, and let $P_0(T)$ be their softmax probabilities. Augmentation changes these logits by $\Delta F_T = F_{D\cup S}(T) - F_0(T)$. With one-hot target labels Y_T, define the current residuals $r_T = Y_T - P_0(T)$. A Taylor expansion gives the empirical target-NLL reduction

$$\Delta R_T \approx \frac{1}{q} r_T^\top \Delta F_T - \frac{1}{2q}\Delta F_T^\top W_T \Delta F_T, \tag{5}$$

where ΔR_T is NLL before minus NLL after augmentation. The matrix W_T has blocks $\mathrm{diag}(p_i) - p_i p_i^\top$, with p_i the current probabilities for target example i. The first term measures correction of existing target errors. The second accounts for the size of the response under target-loss curvature. Utility therefore depends on where the candidate training signal changes target predictions, not only on the strength of that signal. Appendix C gives the integral identity and Taylor remainder.

Response of the reference learner. We linearize the population estimating equation at the same target optimum θ_T for both current-only and augmented fits:

$$\delta_P := \theta_P - \theta_T = -H_P^{-1} b_P + O(\|b_P\|^2). \tag{6}$$

The difference $\delta_{P_\alpha} - \delta_{P_D}$ approximates the parameter change caused by augmentation. For affine logits, the target-logit Jacobian maps this change to ΔF_T.

Finite training samples also produce variation around each population fit. Under model misspecification, sandwich covariance accounts for the difference between training curvature and score variability (Godambe, 1960; Huber, 1967; White, 1982). Combining this covariance with a local expansion of expected target loss (Shimodaira, 2000) yields the following result.

Theorem 1 (Local expansion of expected target risk)**.** *Under the local regularity and moment conditions stated above and in Appendix B.3, evaluate all population quantities at the target optimum θ_T. For a training distribution P and sampling policy π,*

$$\mathbb{E}R_T(\widehat{\theta}_{P,n_P}) - R_T(\theta_T) = \underbrace{\tfrac{1}{2} b_P^\top H_P^{-\top} G_T H_P^{-1} b_P}_{B_T(P)} + \underbrace{\frac{1}{2n_P} \operatorname{tr}\big(G_T H_P^{-1} \mathcal{J}_{P,\pi} H_P^{-\top}\big)}_{V_T(P,n_P)} + o(\|b_P\|^2 + n_P^{-1}). \tag{7}$$

Proof sketch. The population estimating equation gives Equation 6. The asymptotically linear representation of the fitted estimator gives parameter covariance $n_P^{-1} H_P^{-1} \mathcal{J}_{P,\pi} H_P^{-\top}$ to leading order. Expanding target loss at the stationary point θ_T separates squared mean displacement from sampling covariance. Moment bounds control the expected remainder and finite-sample bias.

The terms B_T and V_T capture the contributions of systematic displacement and sampling variability, respectively. Comparing current-only and augmented training gives

$$\begin{aligned} \mathcal{V}_\alpha = {} & B_T(P_D) - B_T(P_\alpha) \\ & + V_T(P_D, n) - V_T(P_\alpha, n_\alpha) + o(\epsilon_\alpha), \end{aligned} \tag{8}$$

where $\epsilon_\alpha = \|b_D\|^2 + \|b_\alpha\|^2 + n^{-1} + n_\alpha^{-1}$. Adding data changes both terms: a reduction in sampling variability can be offset by an unfavorable displacement, while a candidate that corrects the current fit can improve target risk even without closely matching the image distribution. Adding more synthetic samples changes both the mixture weight and sample size, so utility need not increase.

For a correctly specified, unpenalized likelihood at $P = P_T$, $b_T = 0$ and $J_T = H_T = F_T$, recovering the usual Fisher covariance expression. Appendix A.2 gives this special case and an example in which a large empirical score outer product accompanies harmful augmentation. These cases distinguish information about a fitted model from the direction of its change under a candidate distribution.

Numerical validation. We evaluate Theorem 1 in Gaussian linear regression with $p = 8$, zero-mean inputs, and equal current and candidate sample counts. Target and candidate input covariances are I_8; the current covariance is $\operatorname{diag}(\gamma I_4, I_4)$. Current and target labels share coefficient θ_T and unit noise variance. Candidate labels have a shift δ in the first coefficient and noise variance σ_S^2. The predictions in Figure 2 capture lower risk with more data, higher risk under label shift, and greater noise sensitivity under weak coverage.

4.2 Dependence on current data and target task

Equation 8 conditions utility on both the current training distribution and the target loss. The following two constructions make these dependencies explicit.

Proposition 1 (Current-set dependence)**.** *The expected utility in Equation 3 is not generally a function of P_S alone. The same clean candidate distribution can substantially reduce target NLL along a poorly represented direction in the current data, yet offer arbitrarily little benefit once that direction is well estimated. A conditionally biased candidate can help in the former case and harm in the latter.*

For example, consider a Gaussian linear model in which the current design has information L in a target-relevant direction and the candidate adds m units of information in that direction. With unit noise variance and unit target curvature, the clean utility is

$$\frac{1}{2}\left(\frac{1}{L} - \frac{1}{L+m}\right) = \frac{m}{2L(L+m)}. \tag{9}$$

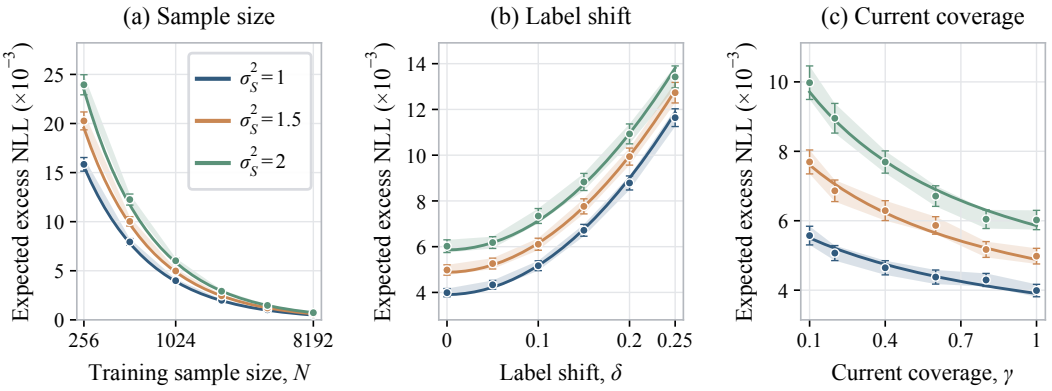

Figure 2: **Numerical validation of the risk expansion.** Colors indicate candidate noise variance σ_S^2. Curves show population-moment predictions from Equation 7; points show mean excess NLL from 500 independent data draws and least-squares refits. Shading and error bars give 95% Monte Carlo intervals for the mean. Unless varied on the horizontal axis, $N = 1024$, $\delta = 0$, and $\gamma = 1$.

The candidate distribution is unchanged, but its benefit decreases as the current estimate becomes more precise. If candidate labels have an offset c in this direction, the additional risk is $\frac{1}{2}\{mc/(L+m)\}^2$. This gives an explicit trade-off between useful information and systematic error, rather than identifying repetition or image similarity with zero utility by definition.

Proposition 2 (No universal image-only utility ranking)**.** *There exist tasks with label mechanisms depending on different image coordinates and two candidate distributions whose utility ordering reverses between tasks while all observed image marginals remain unchanged. No scalar based solely on these image marginals can rank the candidates correctly for both tasks.*

Proof sketch. Let the target distribution assign equal probability to $X = e_1$ and $X = e_2$, with L current samples at each location. Candidate S_i adds m samples at e_i. For task $t \in \{1, 2\}$, let $Y = X_t + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 1)$, and use the reference model $p_\theta(Y \mid X, t) = \mathcal{N}(\theta X_t, 1)$. Only samples at e_t inform θ, so the expected utilities satisfy

$$\begin{aligned} \mathcal{V}^{(1)}(S_1) = \mathcal{V}^{(2)}(S_2) = \frac{1}{4}\left(\frac{1}{L} - \frac{1}{L+m}\right) > 0, \\ \mathcal{V}^{(1)}(S_2) = \mathcal{V}^{(2)}(S_1) = 0. \end{aligned} \tag{10}$$

The utility ranking therefore reverses, although all image marginals remain unchanged across tasks. An image-only score cannot reproduce both rankings. Appendix B.3 gives the risk calculation.

4.3 IGRT ESTIMATION AND SELECTION

Using one real-data reference fit, we aggregate candidate statistics and estimate parameter changes to compare target risks before and after augmentation.

A shared reference for all candidates. We fit the reference model on one part of the labeled target data to obtain $\widehat{\theta}_T$. A disjoint part supplies the target-risk derivatives g_T, G_T; held-out evaluation data are reserved for measuring downstream outcomes. Using the same reference point for all candidates makes their current-only and augmented risk estimates directly comparable.

At $\widehat{\theta}_T$, we estimate current and candidate moments by sample averages, using the same symbols for their estimates. For augmentation weight α,

$$b_\alpha = (1-\alpha)b_D + \alpha b_S, \qquad H_\alpha = (1-\alpha)H_D + \alpha H_S, \tag{11}$$

$$M_\alpha = (1-\alpha)M_D + \alpha M_S. \tag{12}$$

With fixed source counts, score variability is centered within each source before combination:

$$\mathcal{J}_{P_\alpha,\pi} = (1-\alpha)J_D + \alpha J_S, \qquad \mathrm{Cov}(\bar{s}) = \mathcal{J}_{P_\alpha,\pi}/n_\alpha. \tag{13}$$

The alternative covariance for random source membership is derived in Appendix B. The reference fit and current-data statistics are reused across candidate sets and addition counts.

From parameter changes to target risk. For each training distribution P, let b_P^{tr} be the gradient of the reference training objective at $\widehat{\theta}_T$, including any deterministic training penalty. Let A_P be its Hessian with the numerical damping used in the linear solve. The corresponding displacement estimate satisfies $A_P\widehat{\delta}_P = -b_P^{\mathrm{tr}}$. The penalty and damping conventions are given in Appendix A.

The separate risk-scoring data may yield $g_T \neq 0$, so we retain the linear term:

$$\widehat{\mathcal{E}}_T(P, n_P) = g_T^\top \widehat{\delta}_P + \tfrac{1}{2}\widehat{\delta}_P^\top G_T \widehat{\delta}_P + \frac{1}{2n_P}\operatorname{tr}\big(G_T A_P^{-1} \mathcal{J}_{P,\pi} A_P^{-\top}\big)\,. \tag{14}$$

The first two terms evaluate the mean parameter displacement under the target loss; the trace term accounts for data-sampling variability. This is the empirical score motivated by Theorem 1, whose remainder is stated at the population optimum.

The estimated augmentation benefit is

$$\widehat{\mathcal{U}}_\alpha = \widehat{\mathcal{E}}_T(P_D, n) - \widehat{\mathcal{E}}_T(P_\alpha, n_\alpha). \tag{15}$$

Changing the addition count changes both the mixture moments and n_α, so the score accounts for the response and sample-size effects of augmentation.

For affine reference logits, this comparison also gives the estimated target prediction change

$$\widehat{\Delta F_T} = Q_T\big(\widehat{\delta}_{P_\alpha} - \widehat{\delta}_{P_D}\big), \tag{16}$$

where $Q_T = \partial F_\theta(T)/\partial\theta^\top|_{\widehat{\theta}_T}$. This connects the parameter calculation to the learning response in Section 4.1. Low-rank and matrix-free implementations are described in Appendix A.

Comparing sets and addition counts. Let $\mathcal{C}$ contain the feasible candidate-set and budget pairs evaluated by the selection procedure. We return the pair with the largest predicted benefit:

$$(\widehat{S}, \widehat{m}) \in \arg\max_{(S,m)\in\mathcal{C}} \widehat{\mathcal{U}}(S, m \mid D). \tag{17}$$

We score predefined candidate sets at each evaluated addition count. For pool selection, we evaluate additions and exchanges conditional on the selected subset, updating candidate statistics without refitting the reference model. We use constrained beam construction or same-class exchanges, with search budgets and response approximations specified by each experiment. Fixed-budget comparisons use the same m; including $m = 0$ with score zero allows the decision to use no synthetic data.

Explaining candidate responses. Let d_S denote the candidate's standardized measurement changes. The reference-intervention matrix B collects corresponding changes from matched label, coverage, and appearance interventions. We fit $d_S \approx B\widehat{a}$. The coefficients describe the response in terms of these interventions, while the residual measures reconstruction error. Risk prediction uses the candidate's measured moments directly. Appendix A.3 defines the constrained fit and reconstruction residual; risk-based abstention requires separate calibration.

5 Experimental Results

Following the risk analysis in Theorem 1, we evaluate IGRT through utility prediction and downstream data selection. Controlled experiments vary the current data and target task to examine the dependencies in Propositions 1 and 2.

Experimental setup. We evaluate classification on CIFAR-10 and ImageNet-100, semantic segmentation on ADE20K, and object detection on PASCAL VOC, using CNN and Transformer learners. Classification data are generated with EDM, Stable Diffusion, and Z-Image-Turbo; segmentation and detection data use ControlNet-based synthesis and foreground compositing, respectively. Appendix B.4 details data generation, splits, reference models, downstream training, and random seeds. Appendix B.7 shows synthetic images and their training annotations.

Baselines. We compare IGRT with existing methods in two selection settings. For *candidate-set selection*, we construct equally sized candidate sets by varying generation sources, configurations,

Table 1: Downstream performance (%; higher is better). Subscripts give available standard deviations. For selection methods, † marks candidate-set selection (A); unmarked values indicate selection from a shared pool (B). Paired values follow A† / B. Dashes denote unreported results.

Type	Method	CIFAR-10 Accuracy ↑	ImageNet-100 Accuracy ↑	ADE20K mIoU ↑	VOC AP50 ↑
Control	Real only	$49.38_{\pm 1.69}$	$58.60_{\pm 1.91}$	$36.38_{\pm 2.52}$	$61.00_{\pm 1.52}$
	Additional real	$61.51_{\pm 0.29}$	$62.90_{\pm 1.40}$	$36.98_{\pm 0.61}$	$64.94_{\pm 1.17}$
Random		$56.31_{\pm 0.84}$	$59.00_{\pm 1.89}$	$36.22_{\pm 2.61}$	$60.97_{\pm 1.95}$
Distribution	FID	$50.81_{\pm 1.69}^{\dagger}$	$59.10_{\pm 2.09}^{\dagger}$	$35.69_{\pm 1.19}^{\dagger}$	$60.11_{\pm 1.78}^{\dagger}$
	CMMD	$51.46_{\pm 1.03}^{\dagger}$	$57.00_{\pm 1.69}^{\dagger}$	$35.69_{\pm 1.19}^{\dagger}$	$60.11_{\pm 1.78}^{\dagger}$
	FD-DINOv2	$51.46_{\pm 1.03}^{\dagger}$	$57.00_{\pm 1.69}^{\dagger}$	$35.69_{\pm 1.19}^{\dagger}$	$60.11_{\pm 1.78}^{\dagger}$
	PRDC Recall	$51.46_{\pm 1.03}^{\dagger}$	$59.10_{\pm 2.09}^{\dagger}$	$36.16_{\pm 2.48}^{\dagger}$	$59.87_{\pm 1.25}^{\dagger}$
Feature	Text matching	$56.44_{\pm 1.07}$	$58.50_{\pm 2.81}$	$36.54_{\pm 1.46}$	$61.07_{\pm 1.86}$
	Center matching	$\underline{58.64}_{\pm 1.19}$	$60.10_{\pm 1.82}$	$35.35_{\pm 1.64}$	$61.33_{\pm 0.99}$
	K-means	$57.31_{\pm 0.54}$	$57.00_{\pm 2.44}$	$36.23_{\pm 1.88}$	$60.54_{\pm 2.37}$
	Cov. matching	$56.59_{\pm 1.56}$	$57.50_{\pm 2.13}$	$36.39_{\pm 2.25}$	$61.13_{\pm 0.66}$
Information	BAIT	$58.99_{\pm 2.47}$	$\underline{61.00}_{\pm 2.48}$	$\underline{36.48}_{\pm 2.25}$	$61.29_{\pm 0.54}$
Response	First-order influence	$50.81_{\pm 1.69}^{\dagger}$ / $55.81_{\pm 2.07}$	$59.10_{\pm 2.09}^{\dagger}$ / $58.60_{\pm 2.18}$	$36.23_{\pm 1.20}^{\dagger}$ / $36.39_{\pm 0.81}$	$60.11_{\pm 1.78}^{\dagger}$ / $\underline{61.73}_{\pm 1.64}$
	SOURCE	$51.46_{\pm 1.03}^{\dagger}$ / $56.35_{\pm 1.89}$	$59.10_{\pm 0.73}^{\dagger}$ / $60.04_{\pm 2.82}$	$36.23_{\pm 1.20}^{\dagger}$ / $36.30_{\pm 0.46}$	$61.50_{\pm 2.09}^{\dagger}$ / $59.17_{\pm 2.69}$
	LWP	$51.46_{\pm 1.03}^{\dagger}$ / $54.98_{\pm 1.78}$	$57.00_{\pm 1.69}^{\dagger}$ / $58.57_{\pm 1.72}$	$36.16_{\pm 0.60}^{\dagger}$ / $34.06_{\pm 2.36}$	$59.10_{\pm 0.73}^{\dagger}$ / $59.55_{\pm 0.54}$
	IGRT **(Ours)**	$52.50_{\pm 0.46}^{\dagger}$ / $\mathbf{59.08}_{\pm 2.71}$	$59.10_{\pm 2.09}^{\dagger}$ / $\mathbf{61.27}_{\pm 2.22}$	$36.23_{\pm 1.20}^{\dagger}$ / $\mathbf{36.58}_{\pm 2.14}$	$61.50_{\pm 2.09}^{\dagger}$ / $\mathbf{62.08}_{\pm 2.53}$

and mixture proportions. Baselines include random set selection, FID (Heusel et al., 2017), CMMD (Jayasumana et al., 2024), FD-DINOv2 (Stein et al., 2023), and PRDC Recall (Kynkäänniemi et al., 2019). FD-DINOv2 computes Fréchet distance in DINOv2 feature space. For *selection from a candidate pool*, each method selects 1,000 images from the same pool. Baselines include random sampling, text matching and K-means (Lin et al., 2023), center matching (He et al., 2023), covariance matching (Rezaei et al., 2026), and BAIT (Ash et al., 2021). Model-based attribution baselines in both settings include first-order influence (Koh & Liang, 2017), SOURCE (Bae et al., 2024), and LWP (Li et al., 2026). SOURCE uses approximate unrolling of training dynamics, while LWP learns parameter importance weights for gradient-based attribution. For first-order influence, sample scores are aggregated for candidate-set selection.

5.1 Can IGRT select useful synthetic data?

Table 1 compares both selection settings under fixed current data, augmentation sizes, and training budgets. Selection precedes downstream training. Comparisons with current-only and random-selection baselines measure net augmentation and additional selection gains, respectively.

Across classification, segmentation, and detection, IGRT achieves the best or tied-best performance among selection methods in both settings (Table 1). Its selected data also improve on current-only training. Additional candidate-set rankings and pool-selection visualizations are provided in Appendices B.8 and B.9.

5.2 Can IGRT predict augmentation effects and choose the augmentation size?

We compare predicted and measured risk reductions after retraining across synthetic datasets and addition counts (Figure 3). Target risk is classification NLL, mean pixel NLL with equal image weights for segmentation, or a fixed-weight composite detection loss. We report MAE across all dataset–size pairs. Distributional metrics are used only for ranking and selection, not for predicting risk-reduction magnitudes.

5.3 How do labels, current data, and target risk affect utility?

Figure 4 compares predicted and measured utility under three controlled changes: (a) labels or segmentation annotations with synthetic images fixed; (b) coverage across equally sized current sets receiving identical synthetic additions; and (c) target-risk class weights with ADE20K images, annotations, and model outputs fixed.

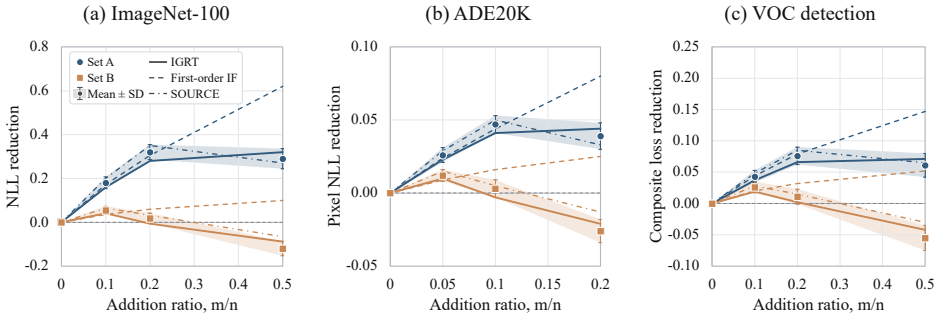

Figure 3: **Predicted and measured augmentation gains.** Colors denote candidate sets. Solid, dashed, and dash-dotted lines show IGRT, first-order IF, and SOURCE predictions. Points show mean paired-retraining gains, with bars and shading indicating ± 1 SD across training seeds. Positive values indicate lower target risk. Straight segments connect evaluated addition ratios.

Table 2: **Prediction error and selection regret (all $\downarrow$; $\times 10^{-3}$).** MAE measures error in predicted risk reduction. Set and size regret measure the shortfall from the best evaluated candidate set and augmentation size, respectively. Subscripts denote standard deviations.

	ImageNet-100			ADE20K		
Method	MAE	Set regret	Size regret	MAE	Set regret	Size regret
SOURCE	$21.6_{\pm 4.7}$	$6.8_{\pm 2.9}$	$9.4_{\pm 3.1}$	$8.2_{\pm 2.0}$	$2.9_{\pm 1.1}$	$4.1_{\pm 1.6}$
IGRT, $J = H$	$14.8_{\pm 3.1}$	$4.6_{\pm 1.8}$	$6.2_{\pm 2.4}$	$6.0_{\pm 1.4}$	$2.1_{\pm 0.9}$	$3.0_{\pm 1.3}$
IGRT, without covariance	$17.3_{\pm 3.8}$	$5.1_{\pm 2.2}$	$7.5_{\pm 2.8}$	$6.7_{\pm 1.8}$	$1.9_{\pm 0.8}$	$3.5_{\pm 1.5}$
IGRT	$\mathbf{8.9}_{\pm 2.2}$	$\mathbf{2.4}_{\pm 1.1}$	$\mathbf{3.3}_{\pm 1.4}$	$\mathbf{3.8}_{\pm 0.9}$	$\mathbf{1.1}_{\pm 0.5}$	$\mathbf{1.7}_{\pm 0.7}$

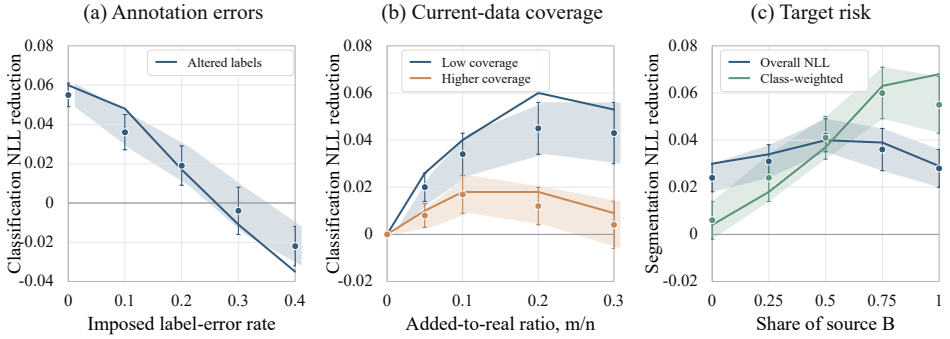

Figure 4: **Factors affecting augmentation utility.** (a) Label corruption with fixed synthetic images. (b) Identical synthetic additions to equally sized current datasets with different coverage. (c) Overall versus prespecified class-weighted segmentation NLL on identical outputs, varying the source-B fraction. Lines: IGRT predictions; points: paired retraining means; bands/bars: ± 1 SD across training seeds. Positive values indicate lower target risk.

6 Discussion and Limitations

IGRT uses a local risk approximation that accounts for candidate labels, current-data coverage, and the target objective to guide candidate-set and augmentation-size selection. This complements generative augmentation (He et al., 2023; Azizi et al., 2023) by matching synthetic data to the intended learner.

Attainable gains depend on generator quality and controllability. Beyond classification, segmentation and detection require precise image–mask or image–box alignment. Spatial conditioning improves control (Zhang et al., 2023) but does not guarantee consistent boundaries, locations, or annotations, making reliable generation a practical constraint. Combining better task-conditioned generation with utility-guided selection could extend augmentation benefits across tasks.

Reproducibility statement

Section 5 describes the evaluation protocol. Appendices A and B provide implementation details, experimental configurations, and theoretical assumptions and proofs.

Ethics statement

We evaluate IGRT on public vision benchmarks. Its utility estimates depend on the target task and reference learner and do not certify fairness, privacy, or safety. Synthetic data may reproduce biases or memorized content; these risks require separate assessment.

AI use statement

Generative AI tools assisted writing, theoretical development, experimental design, coding, and result interpretation. Generative models produced synthetic training images as described in Section 5. The authors take responsibility for the final content and results.

References

Krzysztof Adamkiewicz, Brian Bernhard Moser, Stanislav Frolov, Tobias Christian Nauen, Federico Raue, and Andreas Dengel. When pretty isn't useful: Investigating why modern text-to-image models fail as reliable training data generators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36660–36669, June 2026.

Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 290–306. PMLR, 2022.

Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham M. Kakade. Gone fishing: Neural active learning with fisher embeddings. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8927–8939. Curran Associates, Inc., 2021.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves ImageNet classification. *Transactions on Machine Learning Research*, 2023.

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger Grosse. Training data attribution via approximate unrolling. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 715–724, 2020.

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

Max F. Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. Image retrieval outperforms diffusion models on data augmentation. *Transactions on Machine Learning Research*, 2023.

Kamalika Chaudhuri, Sham M. Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://papers.neurips.cc/paper_files/paper/2015/hash/ca9c267dad0305d1a6308d2a0cf1c39c-Abstract.html.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/284afdc2309f9667d2d4fb9290235b0c-Abstract-Conference.html.

Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79024–79034. Curran Associates, Inc., 2023.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training ... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7382–7392, 2024.

Scott Geng, Cheng-Yu Hsieh, Vivek Ramanujan, Matthew Wallingford, Chun-Liang Li, Pang Wei Koh, and Ranjay Krishna. The unmet promise of synthetic training images: Using retrieved real images performs better. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019.

Amirata Ghorbani, Michael Kim, and James Zou. A distributional framework for data valuation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3535–3544. PMLR, 2020. URL `https://proceedings.mlr.press/v119/ghorbani20a.html`.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A swiss army infinitesimal jackknife. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1139–1147, 2019.

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 221–233. University of California Press, 1967.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9525–9587, 2022.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, June 2024.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474, 2021a.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8110–8118, 2021b.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017.

Pang Wei Koh, Kai-Siang Ang, Hubert Teo, and Percy Liang. On the accuracy of influence functions for measuring group effects. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8780–8802, 2022.

Yongchan Kwon, Manuel A. Rivas, and James Zou. Efficient computation and analysis of distributional Shapley values. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 793–801. PMLR, 2021. URL https://proceedings.mlr.press/v130/kwon21a.html.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Shiye Lei, Hao Chen, Sen Zhang, Bo Zhao, and Dacheng Tao. Image captions are natural prompts for training data synthesis. *International Journal of Computer Vision*, 133(8):5435–5454, 2025.

Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. Learning to weight parameters for training data attribution. In *International Conference on Learning Representations*, 2026.

Zhiteng Li, Lele Chen, Jerone T. A. Andrews, Yunhao Ba, Yulun Zhang, and Alice Xiang. GenDataAgent: On-the-fly dataset augmentation with synthetic data. In *International Conference on Learning Representations*, 2025.

Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 638–647, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/GCV/html/Lin_Explore_the_Power_of_Synthetic_Data_on_Few-Shot_Object_Detection_CVPRW_2023_paper.html.

Yimin Luo, Qinyu Yang, Yuheng Fan, Haikun Qi, and Menghan Xia. Measurement guidance in diffusion models: Insight from medical image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7983–7997, 2024.

Ulrich K. Müller. Risk of bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849, 2013.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 2020.

Dang Nguyen, Jiping Li, Jinghao Zheng, and Baharan Mirzasoleiman. Do we need all the synthetic data? targeted image augmentation via diffusion models. In *International Conference on Learning Representations*, 2026.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113, 2023.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. MAUVE scores for generative models: Theory and practice. *Journal of Machine Learning Research*, 24(356):1–92, 2023.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Parham Rezaei, Filip Kovačević, Francesco Locatello, and Marco Mondelli. High-dimensional analysis of synthetic data selection. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 94117–94168, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, June 2022.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4. URL https://www.sciencedirect.com/science/article/pii/S0378375800001154.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025a.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68578–68598. PMLR, 2025b.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 2021.

Z-Image Team. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. URL https://arxiv.org/abs/2511.22699.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, October 2023.

Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. LTGC: Long-tail recognition via leveraging LLMs-driven generated content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19510–19520, June 2024.

Xinrui Zhou, Yuhao Huang, Haoran Dou, Shijing Chen, Ao Chang, Jia Liu, Weiran Long, Jian Zheng, Erjiao Xu, Jie Ren, Alejandro F. Frangi, Ruobing Huang, Jun Cheng, Xiaomeng Li, Wufeng Xue, and Dong Ni. Ctrl-GenAug: Controllable generative augmentation for medical sequence classification. *International Journal of Computer Vision*, 134(5), 2026. Article 216.

A Implementation Details and Supplementary Properties

This appendix records the numerical conventions and interpretive extensions of the score derived in Section 4.

We write $D \oplus_{\alpha,\pi} S$, or $D \oplus S$ when the policy is clear, for augmentation at weight α under sampling policy π. Ordinary addition is $D \cup S$ with $n_\alpha = n + m$; replacement keeps $n_\alpha = n$.

A.1 Regularization, displacement, and matrix-free computation

All empirical moments are evaluated at the shared real-data reference fit $\widehat{\theta}_T$. A deterministic quadratic training penalty gives

$$b_P^{\mathrm{tr}} = b_P + \lambda_{\mathrm{tr}} R\widehat{\theta}_T, \qquad H_P^{\mathrm{tr}} = H_P + \lambda_{\mathrm{tr}} R. \tag{18}$$

Numerical damping is separate from the training objective:

$$A_P = H_P^{\mathrm{tr}} + \lambda_{\mathrm{num}} D_T, \qquad D_T \succeq 0, \tag{19}$$

and the computed displacement is

$$\widehat{\delta}_P = -A_P^{-1} b_P^{\mathrm{tr}}. \tag{20}$$

Deterministic ridge changes the mean gradient and Hessian but does not add score noise. For randomized source membership, the mixture covariance has the additional term

$$\alpha(1-\alpha)(b_D - b_S)(b_D - b_S)^\top, \tag{21}$$

which is absent under fixed source counts. No density-ratio weighting is introduced unless it is part of the stated training rule.

The parameter-error second moment is approximated by

$$\mathsf{C}_P = \widehat{\delta}_P \widehat{\delta}_P^\top + n_P^{-1} A_P^{-1} \mathcal{J}_{P,\pi} A_P^{-\top}, \qquad \Delta\mathsf{C} = \mathsf{C}_D - \mathsf{C}_{D\oplus S}. \tag{22}$$

For a local quadratic target loss with Hessian $G \succeq 0$, the quadratic contribution to utility is $\frac{1}{2}\langle G, \Delta\mathsf{C}\rangle$; a linear term is included when $g_T \neq 0$. Hessian–vector products and randomized trace estimates avoid storing dense parameter-sized matrices. Using common random trace vectors for current-only and augmented fits can reduce noise in their difference. Generator likelihood or denoising loss may be included as an auxiliary attribution measurement, but does not replace the predictive model gradients or the target-loss geometry.

A.2 Correct specification and a harmful high-score candidate

Proposition 3 (Clean Fisher reduction)**.** *If the unpenalized likelihood is correctly specified and $P = P_T$, then $b_T = 0$ and $\mathcal{J}_{T,\pi} = J_T = H_T = F_T$. Consequently,*

$$\mathcal{E}_T(P_T, n) = \frac{1}{2n}\,\mathrm{tr}(G_T F_T^{-1}) + o(n^{-1}). \tag{23}$$

For the same target NLL, $G_T = F_T$, giving $p/(2n)$ to leading order in identifiable p-dimensional coordinates. This identity concerns zero or vanishing training ridge. Fixed ridge instead uses Equation 18 and leaves the score covariance unchanged.

Proposition 4 (High empirical Fisher can be harmful)**.** *There exist candidate samples whose empirical score outer product is arbitrarily large while adding them strictly increases target NLL.*

The Gaussian construction in Appendix B.3 makes this score grow as M^2, while augmentation introduces a squared-bias contribution $\frac{1}{2}\alpha^2 M^2$ that can exceed the reduction in sampling variance.

A.3 Reference-intervention attribution

Each reference intervention is compared with a matched control of the same size. Interventions include image blur, texture or condition changes, label swaps, and removal of selected within-class regions. Let d_k be the standardized measurement change for intervention k and define

$$\tau(P) = \big[L_b b_P, \mathrm{vech}(U^\top H_P U), \mathrm{vech}(U^\top M_P U)\big], \qquad B = [d_k]_{k=1}^K, \quad C = [\Delta\tau_k]_{k=1}^K. \tag{24}$$

For candidate measurements d_S, fit

$$\widehat{a} = \arg\min_{a_Q \geq 0, a_C} \|W_r(d_S - Ba)\|_2^2 + \lambda_Q \|a_Q\|_1 + \lambda_C \|a_C\|_2^2. \tag{25}$$

Coefficients associated with conditional-validity violations are nonnegative; coverage coefficients may be signed. The reconstruction $C\widehat{a}$ attributes moment changes and constructs counterfactual moment vectors. It does not replace the candidate's empirical moments in the score.

The reconstruction residual is

$$\rho_S = \|W_r(d_S - B\widehat{a})\|_2. \tag{26}$$

The coefficients describe the candidate response relative to the chosen reference interventions, without guaranteeing unique mechanism recovery. The residual measures dictionary reconstruction error and can reflect measurement noise, regularization, or responses outside the reference span. It is not a calibrated measure of utility-prediction error; risk-based abstention requires separate calibration.

B Additional Derivations and Protocol Details

This appendix gives the assumptions, derivations, evaluation protocol, and complete numerical definitions used in the main text.

B.1 Notation and regularity assumptions

Let $Z = (X, Y, C)$ denote an image, its supervision, and optional conditioning context. For a training distribution P, define $L_P(\theta) = \mathbb{E}_P[\ell(Z; \theta)]$ and let $\theta_P^\star$ be its local minimizer. We assume that the loss is three times continuously differentiable in a neighborhood of the target optimum θ_T, that the relevant score moments are finite, and that the population expected Hessian is nonsingular. Empirical inverses use the separately damped matrix defined in Equation 19. The local expansion is intended for mixtures whose pseudo-true optimum remains in this neighborhood. This locality assumption is separate from intervention coverage: the reconstruction residual measures whether a candidate-set response is represented by the reference-intervention matrix.

B.2 Mixture moments under alternative sampling designs

For $Q_\alpha = (1-\alpha)P_D + \alpha P_S$, let $b_P = \mathbb{E}_P[s]$, $M_P = \mathbb{E}_P[ss^\top]$, and $J_P = M_P - b_P b_P^\top$. Linearity of expectation gives

$$b_\alpha = (1-\alpha)b_D + \alpha b_S, \tag{27}$$

$$M_\alpha = (1-\alpha)M_D + \alpha M_S. \tag{28}$$

Expanding $J_\alpha^{\mathrm{iid}} = M_\alpha - b_\alpha b_\alpha^\top$ yields

$$J_\alpha^{\mathrm{iid}} = (1-\alpha)J_D + \alpha J_S \tag{29}$$

$$+ \alpha(1-\alpha)(b_D - b_S)(b_D - b_S)^\top. \tag{30}$$

This is the covariance of one observation from an i.i.d. mixture, including the randomness of source membership. It is not the correct score covariance for a stratified augmentation experiment with fixed n_D and n_S. In that design, writing $N = n_D + n_S$ and $\alpha = n_S/N$,

$$\mathrm{Cov}\left(\frac{1}{N}\sum_{i=1}^{n_D} s_{D,i} + \frac{1}{N}\sum_{j=1}^{n_S} s_{S,j}\right) = \frac{n_D J_D + n_S J_S}{N^2} \tag{31}$$

$$= \frac{1}{N}\{(1-\alpha)J_D + \alpha J_S\}. \tag{32}$$

Hence the between-source term is required for randomized source membership and must be excluded when the counts are conditioned upon. The implementation stores raw b, H, M for each source and selects the score covariance only after the sampling protocol and realized integer counts are known.

If the reference model is fitted with a deterministic quadratic penalty, its score shift is $c = \lambda_{\mathrm{tr}} R\widehat{\theta}_T$. Applying the shift after mixing gives $b^{\mathrm{tr}} = b + c$ and $H^{\mathrm{tr}} = H + \lambda_{\mathrm{tr}} R$. Equivalently, the shifted second moment is $M^{\mathrm{tr}} = M + bc^\top + cb^\top + cc^\top$, so $M^{\mathrm{tr}} - b^{\mathrm{tr}}(b^{\mathrm{tr}})^\top = J$: deterministic ridge does not add score noise. Numerical damping is not included in any of these moments.

B.3 PROOF DETAILS FOR THE MAIN RESULTS

For the expected-risk expansion, the usual distributional asymptotic linearization is supplemented by moment control. Writing $e_{P,n} = \widehat{\theta}_{P,n} - \theta_P$, we require a centered-score linear representation whose remainder has second moment $o(n^{-1})$ and mean $O(n^{-1})$, and $\mathbb{E}\|e_{P,n}\|^3 = O(n^{-3/2})$, uniformly along the local sequence. The expected contribution of events outside the smooth neighborhood is $o(\|b_P\|^2 + n^{-1})$. These conditions justify taking expectations of the local expansion; a finite third score moment alone is not used as a substitute for control of the estimator remainder.

Theorem 1. Write the population estimating equation as $\Psi_P(\theta) = \mathbb{E}_P[s(Z;\theta)]$. A Taylor expansion at θ_T gives

$$0 = \Psi_P(\theta_P) = b_P + H_P(\theta_P - \theta_T) + r_P, \qquad \|r_P\| \le C\|\theta_P - \theta_T\|^2. \tag{33}$$

Uniform invertibility and the inverse-function theorem therefore imply

$$\delta_P := \theta_P - \theta_T = -H_P^{-1} b_P + O(\|b_P\|^2). \tag{34}$$

For i.i.d. observations, the asymptotic linear representation for the misspecified M-estimator is

$$\widehat{\theta}_{P,n} - \theta_P = -H_P^{-1} \frac{1}{n} \sum_{i=1}^{n} \{s(Z_i;\theta_P) - \Psi_P(\theta_P)\} + o_p(n^{-1/2}), \tag{35}$$

where evaluating H_P and J_P at θ_T rather than θ_P changes the final expression by $O(\|b_P\|/n)$. Hence

$$\mathrm{Cov}(\widehat{\theta}_{P,n}) = \frac{1}{n} H_P^{-1} \mathcal{J}_{P,\pi} H_P^{-\top} + o(n^{-1}), \tag{36}$$

where $\mathcal{J}_{P,\pi} = J_P$ in the single-source i.i.d. case. For the fixed-count two-source design, apply the same expansion separately to the two independent stratum means; Equation 32 then gives $\mathcal{J}_{P,\pi} = (1-\alpha)J_D + \alpha J_S$. For randomized source membership, Equation 30 gives $\mathcal{J}_{P,\pi} = J_\alpha^{\mathrm{iid}}$. Finally, with $g_T = 0$,

$$R_T(\theta_T + u) - R_T(\theta_T) = \tfrac{1}{2} u^\top G_T u + O(\|u\|^3). \tag{37}$$

Substitute $u = \delta_P + e_{P,n}$ and take expectations using the stated moment conditions. The displacement remainder contributes $O(\|b_P\|^3)$, the mean-bias cross term is $O(\|b_P\|/n)$, and the third-order sampling remainder is $O(n^{-3/2})$. Covariance and evaluation-point remainders are also $o(\|b_P\|^2 + n^{-1})$, yielding Equation 7. When $g_T \neq 0$, the exact linear contribution is $g_T^\top \mathbb{E}[\widehat{\theta}_{P,n} - \theta_T]$. Replacing the expected displacement by $-H_P^{-1} b_P$ introduces an additional mean-displacement error, including finite-sample mean bias. These terms need not be negligible at order $\|b_P\|^2 + n^{-1}$, so the preceding small-o remainder applies only to the stated $g_T = 0$ result.

Proposition 1. Consider a two-coordinate homoscedastic linear model with target curvature $G_T = I$. Let the current design information be $A_D = \mathrm{diag}(n, L)$ and let the candidate contribute $A_S = m e_2 e_2^\top$. Ignoring constants, its clean variance reduction in the second direction is

$$\frac{1}{L} - \frac{1}{L+m} = \frac{m}{L(L+m)}. \tag{38}$$

It is large when the second direction is missing (L small) and converges to zero as current coverage $L \to \infty$, although P_S is unchanged. If the candidate has a fixed label offset c along e_2 and the same unit noise variance, its squared displacement adds $\frac{1}{2}\{mc/(L+m)\}^2$ to target risk. The exact utility in this direction is therefore

$$\frac{m}{2L(L+m)} - \frac{m^2 c^2}{2(L+m)^2}. \tag{39}$$

When $mc^2 > 1$, this is positive for $L < m/(mc^2 - 1)$ and negative for larger L. Both the candidate distribution and its count are fixed in this comparison. Thus the change in target NLL after augmentation cannot be a function of P_S alone.

Table 3: Proposed real-data allocation, in images. All listed subsets are mutually disjoint. The additional-real control uses 1,000 images in every task.

Dataset	$\lvert D\rvert$	$\lvert T_{\mathrm{fit}}\rvert$	$\lvert T_{\mathrm{score}}\rvert$	$\lvert T_{\mathrm{dev}}\rvert$	$\lvert R_{\mathrm{add}}\rvert$	Evaluation split
CIFAR-10	1000	1000	1000	1000	1000	Official test
ImageNet-100	2000	1000	1000	1000	1000	Selected-class validation
ADE20K	5000	1000	1000	500	1000	Official validation
PASCAL VOC	2000	1000	1000	500	1000	VOC 2007 test

Proposition 2. Let the target image coordinates take values $X = e_1$ and $X = e_2$ with equal probability. Task $t \in \{1, 2\}$ has labels $Y = X_t + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$, and reference likelihood $p_\theta(Y \mid X, t) = \mathcal{N}(\theta X_t, 1)$. The current design contains L examples at each coordinate. Candidate S_1 adds m examples at e_1, and S_2 adds m examples at e_2; labels follow the corresponding task. Under task 1, only observations at e_1 provide information about θ. Consequently, the expected target-NLL benefit of S_1 is

$$\frac{1}{4}\left(\frac{1}{L} - \frac{1}{L+m}\right) > 0, \tag{40}$$

whereas that of S_2 is zero. Task 2 reverses these benefits. The target, current, and candidate image marginals are unchanged across tasks; only the label mechanism and corresponding prediction problem change. An image-only scalar must return the same ordering in both tasks and therefore cannot match both strict utility orderings.

Proposition 3. For an unpenalized correctly specified likelihood at P_T, differentiation under the integral gives $\mathbb{E}_T[s] = 0$ and Bartlett's identity gives $\mathbb{E}_T[ss^\top] = \mathbb{E}_T[h] = F_T$. Thus $b_T = 0$ and $J_T = H_T = F_T$. Substitution into Theorem 1 eliminates the displacement term and gives $\mathrm{tr}(G_T F_T^{-1})/(2n)$. Damping is a numerical regularizer and is taken to zero for this population identity.

Proposition 4. Let the target set and current-set observations be $Y \sim \mathcal{N}(0, 1)$ and use a Gaussian reference model with loss $\ell(Y; \theta) = \frac{1}{2}(Y - \theta)^2$. A synthetic candidate set contains m deterministic observations $Y = M$. At $\theta_T = 0$, its score outer product is M^2, which can be made arbitrarily large. With n current-set observations, the augmented estimator is

$$\widehat{\theta}_{n,m} = \frac{\sum_{i=1}^{n} Y_i + mM}{n + m}. \tag{41}$$

The expected target excess risk is exactly

$$\frac{1}{2}\mathbb{E}[\widehat{\theta}^2_{n,m}] = \frac{1}{2}\left\{\frac{n}{(n+m)^2} + \left(\frac{mM}{n+m}\right)^2\right\}, \tag{42}$$

whereas the current-set-only target NLL is $1/(2n)$. The augmented target NLL exceeds the baseline for sufficiently large $|M|$, despite its arbitrarily large empirical Fisher score. Its nonzero mean score $b_S = -M$ produces exactly the omitted bias direction.

B.4 Main experimental protocol

B.4.1 Data splits and random seeds

For each dataset, we will construct mutually disjoint real-data subsets: the current training set D, reference-fitting set T_{fit}, risk-scoring set T_{score}, development set T_{dev}, and additional-real-data control R_{add}. The remaining training images will form a generation-support set G. Only G may supply source images, masks, foreground objects, or backgrounds for image-conditioned generation. The official evaluation split E will remain inaccessible to generation, selection, hyperparameter tuning, and reference fitting.

Classification splits will be class balanced. For ImageNet-100, we will choose 100 classes without replacement from the sorted ImageNet-1K class identifiers using seed 2027, and fix this class list

across all experiments. ADE20K and VOC splits will use iterative multilabel stratification to preserve class coverage. VOC training images will come from the union of the 2007 and 2012 trainval splits, with duplicate image identifiers removed.

All methods will receive the same D, candidate data, and labeled reference subsets. The additional labeled data used for scoring will be reported separately from the downstream training set size. Neither T_{fit} nor T_{score} will be appended to downstream training.

We will fix the split seed to 2027, the reference-training seed to 2718, the selection seed to 31415, and the numerical-probe seed to 1618. For task index t, generation-configuration index c, and image index i, the generation seed will be

$$s_{\mathrm{gen}} = 10^6 t + 10^4 c + i, \qquad 0 \le i < 2000.$$

Indices will be assigned before generation and recorded with each image.

The main results will use five downstream training seeds, $\{0, 1, 2, 3, 4\}$. Data splits, generated images, selected subsets, and scoring randomness will remain fixed across these five runs. Consequently, the reported standard deviations will describe downstream training variability, not variability across generated pools or dataset splits.

B.4.2 Synthetic-data generation

Generators will remain frozen during candidate construction. For each generator, the final configuration will identify the exact checkpoint, repository revision, sampler, resolution, inference precision, and software version. No generator or filtering parameter will be selected using E. Use of public pretrained generators will be distinguished from task-specific generation or fine-tuning; disjoint experimental splits will not be presented as a guarantee about public-model pretraining exposure.

CIFAR-10 classification. We will use a class-conditional CIFAR-10 EDM checkpoint, Stable Diffusion 1.4, and Z-Image-Turbo. EDM will generate 32×32 images using an 18-step deterministic Heun schedule with $\sigma_{\min} = 0.002$, $\sigma_{\max} = 80$, $\rho = 7$, and zero stochastic churn. Stable Diffusion will use 512×512 generation, 50 DDIM steps, guidance scale 7.5, and $\eta = 0$. Z-Image-Turbo will use 1024×1024 generation with its native sampler, nine inference steps, and classifier-free guidance disabled. Text-generated images will be resized to 32×32 with antialiased bicubic interpolation for downstream CIFAR-10 training.

Two prompt templates will be used:

> Basic: "A photograph of a {class}."
> Detailed: "A realistic photograph of a {class}, with the main subject clearly visible, natural lighting, and a plausible background."

The intended class will supply the synthetic training label.

The ten candidate configurations will comprise five primitive sources: EDM, SD-basic, SD-detailed, Z-Image-basic, and Z-Image-detailed; four equal mixtures: EDM/SD-basic, EDM/Z-Image-basic, SD-basic/Z-Image-basic, and SD-detailed/Z-Image-detailed; and one approximately equal three-way mixture of EDM, SD-detailed, and Z-Image-detailed. Rounding of mixture counts will be deterministic and class balanced.

ImageNet-100 classification. We will use Stable Diffusion 1.4 and Z-Image-Turbo with the same sampling settings as above. Prompts will include the class name and its fixed disambiguating description. The eight configurations will cross two generators, two prompt templates, and two context instructions: ordinary lighting and background, or varied lighting and background. All prompt strings and class descriptions will be fixed before selection. Generated images will retain their original resolution until the downstream training transform is applied.

ADE20K semantic segmentation. We will use Stable Diffusion 1.5 with a segmentation-conditioned ControlNet checkpoint. Conditioning masks will come exclusively from G and will be converted using the checkpoint's prescribed semantic color encoding. Images and masks will share the same geometric transform; mask resizing will use nearest-neighbor interpolation.

Generation will use 512×512 resolution, 50 DDIM steps, guidance scale 7.5, and $\eta = 0$. Ten configurations will cross ControlNet strengths $\{0.6, 0.8, 1.0, 1.2, 1.4\}$ with basic and detailed prompts. Basic prompts will list the scene and classes present in the mask; detailed prompts will additionally specify natural appearance, lighting, and material descriptions. Scene descriptions will be derived only from the source annotations and a fixed template bank.

The conditioning mask will be retained as the intended supervision. It will not be assumed to be a verified pixel-perfect annotation of the generated image. This distinction allows the experiment to assess candidate data with imperfect image–annotation consistency. All selection methods will receive the same generated images and the same intended masks.

PASCAL VOC object detection. We will construct synthetic detection images by pasting one foreground object into an annotated background from G. Foreground cutouts will use instance masks from the portion of the VOC segmentation training data assigned to G. Object labels will come from the corresponding source annotations.

Eight configurations will cross: (i) original versus diffusion-edited foreground appearance; (ii) pasted-object area fractions of $[0.02, 0.06]$ versus $[0.06, 0.15]$ of the destination image; and (iii) uniform placement versus placement sampled from the current-set object-center distribution. Foreground appearance editing will use a Stable Diffusion 1.5 inpainting checkpoint with 50 sampling steps, guidance scale 7.5, and denoising strength 0.5. The original alpha mask will be reapplied after editing.

Pasted objects will remain within image boundaries and will not overlap existing annotated bounding boxes. An unsuccessful placement will be retried at most 100 times before the composition is rejected. The new bounding box will be computed from the transformed foreground mask; existing background boxes and class labels will be retained. This construction avoids silently treating occluded or removed background objects as unchanged annotations.

Common generation rules. Each configuration will provide 2,000 accepted images. Candidate configurations will use separate random streams. Only unreadable files, empty images, invalid annotations, and violations of the stated compositing constraints will be rejected before selection. No method-specific or downstream-performance-based filtering will be applied to the shared pool.

Classification configurations will be class balanced. For segmentation, each generated image will have a designated conditioning class present in its source mask. For detection, its designated class will be the pasted-object class. These tags will define common selection quotas without replacing the full segmentation or detection annotations.

B.4.3 Candidate-set selection, pool selection, and addition size

Setting A: candidate-set selection. We will form one 1,000-image candidate set from each generation configuration using a fixed, stratified ordering. The candidate-menu sizes will therefore be $K = 10, 8, 10, 8$ for CIFAR-10, ImageNet-100, ADE20K, and VOC, respectively. Every method will score the same menu and select one set before downstream training. Ties will be resolved by the fixed configuration index.

Setting B: selection from a shared pool. The shared pool will contain all 2,000 images from each configuration, giving pool sizes of 20,000, 16,000, 20,000, and 16,000. Every method will select exactly 1,000 images without replacement. Classification selections will use equal class quotas. Segmentation and detection selections will use the same quotas over the designated conditioning or pasted-object classes. A fixed largest-remainder rule will allocate non-integer quotas.

For subset objectives such as covariance matching, BAIT, and IGRT, we will start from the same stratified random subset and consider 3,000 same-tag exchange proposals. A proposal will be accepted only if it improves the method's objective. IGRT will update the selected-set moments after each accepted exchange without refitting its reference model. Methods that assign independent sample scores will select the highest-scoring feasible samples. The search budget and scoring time will be reported separately from downstream training cost.

Controls. Real-only training will use D. The additional-real control will use $D \cup R_{\mathrm{add}}$, with 1,000 added real images and the same downstream update budget. Uniform selection will sample 1,000 synthetic images under the same quota constraints. The additional-real control will be described as a reference comparison, not as a mathematical upper bound.

Addition-size experiments. For each generation configuration, we will construct nested candidate subsets using a fixed ordering. The evaluated addition counts will be

$$\begin{aligned} \text{CIFAR-10}: &\quad m \in \{0, 50, 100, 250, 500, 1000\}, \\ \text{ImageNet-100 and VOC}: &\quad m \in \{0, 200, 400, 600, 800, 1000\}, \\ \text{ADE20K}: &\quad m \in \{0, 250, 500, 750, 1000\}. \end{aligned}$$

The plotted addition ratio will be $m/|D|$. The same image identifiers and nested prefixes will be used by all estimators. Size selection will maximize predicted utility over the evaluated counts, including $m = 0$ with utility zero.

B.4.4 Reference models and IGRT computation

The reference model used for scoring will be distinguished from the downstream learner used to measure utility. Reference fitting and all candidate scores will be completed before the corresponding augmented downstream runs.

Classification reference. We will use frozen, normalized 384-dimensional DINOv2 ViT-S/14 features and a multinomial logistic head fitted on T_{fit}. The average-loss objective will use weight ridge 0.01 and an unpenalized intercept. Optimization will use deterministic L-BFGS with at most 500 iterations and gradient tolerance 10^{-6}. The same reference fit will score all candidate sets and addition counts within a dataset.

Segmentation reference. We will freeze the feature extractor of a real-only segmentation model trained on D with the fixed reference seed. A pixel-classification head will be fitted on T_{fit}. Reference fitting and moment estimation will use 1,024 uniformly sampled valid pixels per image, with pixel indices fixed across candidate comparisons. Each image will receive equal weight regardless of its number of valid pixels. Final downstream evaluation will use all valid pixels.

Detection reference. We will initialize from a real-only detector trained on D with the fixed reference seed. Backbone features will remain frozen while prediction heads are fitted on T_{fit}. Anchor assignments, proposal sampling, and RoI matching will be fixed during local moment estimation. The scoring parameter block will include the prediction layers for proposal objectness and regression and RoI classification and regression. This is a local head-based approximation, not a simulation of full detector retraining.

Scoring and numerical settings. Target derivatives will be estimated on T_{score}. Current and candidate moments will use their supplied annotations. For segmentation and detection, one image will constitute one observation: its pixel-, anchor-, and instance-level losses will first be aggregated before computing the image gradient. The score covariance will therefore not treat pixels or proposals from the same image as independent training observations.

We will use the fixed-source-count covariance and mixture weight $\alpha = m/(|D| + m)$. Training ridge and numerical damping will remain separate. A fixed damping value

$$\lambda_{\mathrm{num}} = 10^{-3} \max\left\{ \frac{\mathrm{tr}(H_D)}{p}, 10^{-8} \right\}$$

will be computed once per task and reused across candidate mixtures. Inverse-Hessian products will use float64 conjugate-gradient solves with relative residual tolerance 10^{-5} and at most 200 iterations. Trace estimates will use 128 common Rademacher probes for current-only and augmented calculations. Solver failures will be recorded rather than replaced using observed downstream outcomes.

All methods will share the same available real-data splits. Distributional metrics will be evaluated only for ranking and selection, with their score direction fixed in advance. They will not be interpreted

Table 4: Proposed downstream training configurations. Budgets are optimizer updates, not epochs. Learning rates are the initial rates after warm-up.

Dataset	Learner	Initialization	Optimizer	LR	Batch	Updates
CIFAR-10	ResNet-18	Random	SGD	0.1	128	4000
ImageNet-100	ResNet-50	ImageNet-1K	SGD	0.01	128	8000
ADE20K	SegFormer-B0	ImageNet-1K backbone	AdamW	6×10^{-5}	16	40000
VOC	Faster R-CNN, R50-FPN	ImageNet-1K backbone	SGD	0.02	16	24000

as calibrated NLL predictions. LWP will likewise be evaluated as a ranking method unless a separate calibration procedure is explicitly specified.

B.4.5 Downstream learners and training budgets

Every augmented run will start from the same initialization as its paired real-only run. We will not warm-start an augmented model from a completed real-only model. For pretrained architectures, all comparisons will use the same pretrained weights and identically initialized task-specific heads.

CIFAR-10 will use a 3×3, stride-one ResNet stem without the initial max-pooling layer. ImageNet-100 will replace the pretrained classifier with a 100-class head. ADE20K will use a 150-class segmentation head. VOC will use 20 foreground classes plus background, with randomly initialized detection-specific layers. No COCO-pretrained detector weights will be used in this design.

SGD will use momentum 0.9. Weight decay will be 5×10^{-4} for CIFAR-10 and 10^{-4} for ImageNet-100 and VOC. AdamW will use $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay 0.01, excluding biases and normalization parameters from decay.

Classification learning rates will follow cosine decay after a warm-up covering 5% of the update budget. ADE20K will use 1,500 warm-up updates followed by polynomial decay with power 1. VOC will use 500 warm-up updates and learning-rate reductions by a factor of 10 at updates 16,000 and 22,000. The final scheduled checkpoint will be evaluated; no evaluation-set-based checkpoint selection will be allowed.

Training transforms. CIFAR-10 will use four-pixel padding, random 32×32 cropping, and horizontal flips with probability 0.5. ImageNet-100 will use random resized 224×224 crops with scale range $[0.08, 1]$ and horizontal flips. ADE20K will use random scaling in $[0.5, 2]$, 512×512 crops, and horizontal flips, applying identical geometric transforms to images and masks. VOC will resize the shorter image side to 800 pixels subject to a maximum longer side of 1,333 pixels and use horizontal flips, with matching box transformations. Mixup, CutMix, and label smoothing will be disabled in this design to keep annotation interventions and NLL comparisons directly interpretable.

Fixed computation across addition counts. Mini-batches will be sampled uniformly from the concatenated dataset $D \cup S$. The number of optimizer updates, batch size, learning-rate schedule, and augmentation policy will remain fixed as m changes. Thus, equal image weights correspond to $\alpha = m/(|D| + m)$.

Epoch counts will be reported as derived quantities:

$$E_{\text{equiv}}(m) = \frac{K_{\text{upd}} B_{\text{train}}}{|D| + m},$$

where K_{upd} is the update budget and B_{train} is the batch size. This avoids describing comparisons as both equal-epoch and equal-computation experiments when their dataset sizes differ.

B.4.6 Target risks and downstream metrics

Classification will report top-1 accuracy and mean NLL. Segmentation will report mIoU over the 150 ADE20K classes and image-averaged pixel NLL,

$$R_{\text{seg}} = \frac{1}{|E|} \sum_{i \in E} \frac{1}{|\Omega_i|} \sum_{u \in \Omega_i} -\log p_\theta(y_{iu} \mid x_i, u),$$

where Ω_i contains valid, non-ignored pixels. Equal image weighting will also be used for target-risk derivatives and synthetic-data scoring.

Detection will report VOC 2007 11-point AP50 and the fixed-weight composite loss

$$R_{\mathrm{det}} = L_{\mathrm{RPN,cls}} + L_{\mathrm{RPN,box}} + L_{\mathrm{RoI,cls}} + L_{\mathrm{RoI,box}}.$$

All four coefficients will equal one and remain unchanged across methods, candidate sets, and addition counts. Loss normalization, anchor matching, proposal limits, and RoI sampling will be fixed in the evaluation configuration. Classification terms will use cross-entropy and box terms smooth-L_1 loss with transition parameter one. Composite-loss reduction will be distinguished from AP50 improvement; it is not an NLL in the same units as the classification and segmentation risks.

For training seed r, observed utility will be

$$U_r(S \mid D) = R_E(\theta_{D,r}) - R_E(\theta_{D\cup S,r}).$$

Both terms will use the same evaluation images, and positive utility will indicate lower target risk.

B.4.7 Controlled comparisons

Annotation errors. We will fix a 250-image, class-balanced CIFAR-10 synthetic set and change labels at rates $\{0, 0.1, 0.2, 0.3, 0.4\}$. Corrupted subsets will be nested, and selected labels will be changed by the fixed cyclic map $y \mapsto (y + 1) \bmod 10$. Images, current data, addition count, and training configurations will remain unchanged. The intended-label and corrupted-label versions will be scored and trained separately.

Current-data coverage. We will construct two 1,000-image CIFAR-10 current sets with 100 images per class. Using frozen DINOv2 features, a target-relevant region within each class will be defined from T_{fit} before examining downstream outcomes. The coverage-gap set will contain 10 images per class from that region; the coverage-filled set will contain 50. Replacing 40 images per class will preserve total size and class balance.

Both current sets will receive identical synthetic images, labels, and nested addition orders. Each observed gain will be measured relative to its own current-only baseline. Region construction and candidate ordering will not use evaluation-set predictions or measured augmentation gains.

Target-risk definition. On ADE20K, we will form 1,000-image candidate sets with source-B fractions $\{0, 0.25, 0.5, 0.75, 1\}$. Source A will use basic prompts and ControlNet strength 0.6; source B will use detailed prompts and strength 1.2. Each trained model's identical outputs will be evaluated under ordinary pixel NLL and a prespecified class-weighted NLL.

The 15 least frequent classes in D will receive weight four; other classes will receive weight one. Class identifiers will resolve frequency ties. Weighted NLL will be normalized by the sum of pixel weights within each image and then averaged across images. Only the target-risk definition will change between the two evaluations; images, annotations, model parameters, and outputs will remain fixed. The same weights will be used when recomputing IGRT target derivatives on T_{score}.

B.4.8 Reporting uncertainty and selection regret

Downstream performance will be summarized by the mean and sample standard deviation over the five training seeds. For utility curves, standard deviations will be computed from the paired differences U_r, not from the two losses independently. These intervals will not be labeled confidence intervals.

Prediction MAE will be computed across all prespecified candidate-set–size pairs. Sign agreement will compare predicted and measured utility using threshold zero, with exact ties counted as agreement only when both signs are zero. Candidate-set Spearman correlations will be computed separately at each fixed addition count, with average ranks for ties.

At fixed m, set-selection regret will be

$$\max_{S\in\mathcal{C}_m} \bar{U}(S, m) - \bar{U}(\widehat{S}_m, m),$$

where $\widehat{S}_m$ is selected without access to evaluation outcomes. For a fixed candidate source, size-selection regret will be

$$\max_{m \in \mathcal{M}} \bar{U}(S, m) - \bar{U}(S, \widehat{m}).$$

Here $\bar{U}$ averages paired utility over training seeds, and $\mathcal{M}$ includes zero. Observed outcomes will be used only to compute these evaluation metrics, never to choose $\widehat{S}_m$ or $\widehat{m}$.

B.5 Numerical validation protocol

Data and reference learner. Figure 2 evaluates the risk expansion in Theorem 1 using Gaussian linear regression. We set $p = 8$, $\theta_T = \mathbf{1}_8/\sqrt{8}$, and generate observations from the target, current, and candidate distributions as

$$\begin{aligned} X_T &\sim \mathcal{N}(0, I_8), & Y_T &= X_T^\top \theta_T + \varepsilon_T, \\ X_D &\sim \mathcal{N}(0, \Sigma_D), & Y_D &= X_D^\top \theta_T + \varepsilon_D, \\ X_S &\sim \mathcal{N}(0, I_8), & Y_S &= X_S^\top (\theta_T + \delta e_1) + \varepsilon_S, \end{aligned}$$

where $\Sigma_D = \mathrm{diag}(\gamma I_4, I_4)$ and e_1 is the first coordinate vector. The noises are independent of the inputs, with $\varepsilon_T, \varepsilon_D \sim \mathcal{N}(0, 1)$ and $\varepsilon_S \sim \mathcal{N}(0, \sigma_S^2)$. Observations are independent within and across sources. Each training set contains exactly $N/2$ current and $N/2$ candidate observations, giving augmentation weight $\alpha = 1/2$. We fit the reference likelihood $p_\theta(y \mid x) = \mathcal{N}(y; x^\top \theta, 1)$ by unpenalized least squares, without an intercept:

$$\widehat{\theta} = (X_D^\top X_D + X_S^\top X_S)^{-1}(X_D^\top Y_D + X_S^\top Y_S),$$

where X_D, X_S now denote the corresponding design matrices. The target excess NLL of each fitted model is evaluated exactly as

$$Z = R_T(\widehat{\theta}) - R_T(\theta_T) = \frac{1}{2}\|\widehat{\theta} - \theta_T\|_2^2,$$

so evaluation introduces no finite-test-set noise.

Population-moment prediction. We compute the approximation in Equation 7 using known population moments at θ_T. For this model,

$$\begin{aligned} G_T &= I_8, & b_\alpha &= -\frac{\delta}{2} e_1, & H_\alpha &= \frac{1}{2}(\Sigma_D + I_8), \\ J_D &= \Sigma_D, & J_S &= (\sigma_S^2 + \delta^2) I_8 + \delta^2 e_1 e_1^\top, & \mathcal{J}_{P_\alpha, \pi} &= \frac{1}{2}(J_D + J_S). \end{aligned}$$

The covariance is centered within each source because source counts are fixed; it therefore excludes between-source score-mean variation. Solid curves evaluate the resulting risk approximation on a dense horizontal grid, without fitting to the Monte Carlo outcomes.

Parameter sweeps. All three panels use $\sigma_S^2 \in \{1, 1.5, 2\}$, with the same color assignment. The remaining settings are:

- **Sample size:** $N \in \{256, 512, 1024, 2048, 4096, 8192\}$, with $\delta = 0$ and $\gamma = 1$.
- **Label shift:** $\delta \in \{0, 0.05, 0.10, 0.15, 0.20, 0.25\}$, with $N = 1024$ and $\gamma = 1$.
- **Current coverage:** $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$, with $N = 1024$ and $\delta = 0$.

Thus, δ changes the candidate labeling mechanism, whereas γ changes current-data coverage while keeping the candidate and target distributions fixed.

Monte Carlo estimation. For each unique configuration, we independently generate and refit $R = 500$ training datasets. Configurations shared across panels reuse the same replicates, giving 48 unique configurations and 24,000 least-squares fits. The base random seed is 2026092404. Points report the empirical mean $\bar{Z}$; error bars and shading show the approximate 95% Monte Carlo interval

$$\bar{Z} \pm 1.96 \frac{s_Z}{\sqrt{R}}, \qquad s_Z^2 = \frac{1}{R-1}\sum_{r=1}^{R}(Z_r - \bar{Z})^2.$$

These intervals quantify uncertainty in the estimated expected excess NLL, rather than variability of individual fits. The vertical axes display NLL in units of 10^{-3}.

B.6 IMAGE-SCALE PROTOCOL AND IMPLEMENTATION

Same-candidate context comparison (Figure 1). Both current sets contain 100 CIFAR-10 images per class. The filled set replaces 40 images per class to increase coverage along fixed training-side DINO feature directions. From a pool of 4,000 images generated with Stable Diffusion 1.4 and Z-Image-Turbo, we construct class-balanced, nested additions of 50, 100, and 250 images in the selected feature region. Their images, labels, and order are identical across the two current sets. We use frozen, normalized 384-dimensional DINOv2 features and fit multinomial classifiers with ridge coefficient 0.01, an unpenalized intercept, and deterministic L-BFGS optimization. A separate real-data set fits the reference model; 1,000 target images supply the risk statistics. The full IGRT predictions, including the Godambe term, are saved before augmented classifier fitting. Measured gains subtract each condition's own current-only NLL on the same 8,000 held-out test images.

Supplementary fixed-feature CIFAR-10 validation. This independent supplementary experiment evaluates risk estimates on a fixed candidate panel, separately from the downstream selection experiments in Table 1. It uses frozen 384-dimensional DINOv2 ViT-S/14 features without PCA and a ridge-regularized multinomial logistic reference model. Only the classifier is refitted for each augmented dataset; the backbone remains fixed. The target-fitting, target-risk, current-set, and held-out evaluation splits contain 1,000 class-balanced images each and are mutually disjoint; mechanism-specific candidate sets contain 800 images, with label mismatch reusing the exact clean-redundant images before changing labels. Five candidate mechanisms—clean complementary, clean redundant, same-image cyclic label mismatch, EDM base, and EDM native stress—are crossed with three mutually exclusive candidate blocks and exact added counts 50, 100, and 250. For the 1,000-image current set, the realized fractions are therefore $50/1050 = 0.047619$, $100/1100 = 0.090909$, and $250/1250 = 0.2$. Prefixes are nested within a block, but the three maximum-dose blocks are disjoint. The first three mechanisms form the 27-cell control panel; the two EDM mechanisms form 18 pilot-informed secondary cells.

Every predefined cell is evaluated, rather than selecting one candidate set at a fixed budget (setting A) or selecting 1,000 images from a shared pool (setting B). Dose selection compares only the additions of 50, 100, and 250 images within each mechanism–block group. The fixed-feature refits, candidate panel, and addition counts define this supplementary protocol; they are not the training configurations or augmentation budgets of the main selection table.

All 45 IGRT estimates use the splits and fixed-feature model defined above. Matrix-free float64 conjugate-gradient solves use 256 common Hutchinson probes. In every cell, the numerical criterion required the larger of the two $1.96\times$ standard-error half-widths, for IGRT and its internal Fisher comparator, to be at most 5×10^{-4}; the observed maximum was 1.822×10^{-4}. The change in target NLL is $\mathrm{NLL}(D) - \mathrm{NLL}(D \oplus S)$, so positive values are beneficial. The primary evaluation statistic is mean absolute prediction error over the fixed 27-cell control panel, with a 0.002 equivalence margin. The one-sided 95th percentile of its aligned bootstrap distribution is 0.00181. The full 45-cell correlations, NLL-region agreement for internal ablations, and dose selection in Tables 5–6 are descriptive finite-panel summaries, not inference over i.i.d. cells or generator populations.

We quantify held-out evaluation uncertainty with 5,000 shared class-stratified bootstrap resamples (seed 20279822), resampling the same held-out evaluation examples for every cell. The candidate blocks and nested doses remain fixed design strata; they are not resampled as independent replicates. Positive, practically equivalent, and harmful regions are defined by the change in target NLL above 0.002, inside $[-0.002, 0.002]$, and below -0.002, respectively. The three control tests and two paired mechanism contrasts use centered-bootstrap tests with Holm correction. This resampling protocol quantifies held-out evaluation uncertainty conditional on the fixed 45-cell panel.

Supplementary-panel comparator definitions. The external comparators in Table 5 are evaluated on the fixed panel defined above. Scores are computed on each augmented mixture and oriented so that larger values indicate a more favorable change. We report correlations and within-block dose choices, not calibrated NLL estimates or utility-sign predictions. Clean-FID uses Clean-FID v0.1.35 "clean" 2048-dimensional Inception features and the exact low-rank Bures identity (Heusel et al., 2017). CLIP-CMMD uses the official OpenAI ViT-L/14@336 QuickGELU checkpoint and the biased RBF MMD estimator with $\sigma = 10$ and scale 1000 (Jayasumana et al., 2024). DINO-FD uses the same fixed 384-dimensional feature representation as IGRT, sample covariance with one degree-of-freedom

Table 5: Supplementary fixed-feature CIFAR-10 validation on 45 predefined cells. Spearman and Pearson coefficients compare each oriented score with the observed reduction in target NLL. Dose accuracy records agreement with the observed best addition count among 50, 100, and 250 within each mechanism–block group. External scores are used for diagnostic comparison and dose ranking, not calibrated utility prediction.

Evaluator	Family	Spearman	Pearson	Dose accuracy
IGRT	Godambe risk expansion	0.9966	0.9998	0.867
CLIP-CMMD	distribution	0.8043	0.2138	0.733
LEEP	transferability	0.7686	0.9507	0.600
CLIP label cosine	label compatibility	0.7544	0.9821	0.467
PRDC recall	support	0.7283	0.4791	0.667
DINO-FD	distribution	0.7162	0.3369	0.600
CLIP label log probability	label compatibility	0.6913	0.9811	0.533
CLIP label accuracy	label compatibility	0.6878	0.9807	0.467
PRDC coverage	support	0.6867	0.3662	0.400
Density response ($k = 5$)	local density	0.6842	0.5063	0.333
PRDC density	support	0.6527	0.4422	0.600
DINO label log probability	label compatibility	0.6440	0.9823	0.533
DINO label accuracy	label compatibility	0.6281	0.9806	0.467
LogME	transferability	0.6009	0.9668	0.400
DINO-KID	distribution	0.5296	0.2883	0.533
Clean-FID	distribution	0.4923	0.1068	0.600
Inception-KID	distribution	0.4327	0.2177	0.667
PRDC precision	support	0.2992	0.2981	0.467

correction, and Gaussian Bures distance. PRDC uses Euclidean five-nearest-neighbor precision, recall, density, and coverage in that DINO feature representation (Naeem et al., 2020). LEEP uses the reference-model class posterior as its source-label posterior (Nguyen et al., 2020). For FID, KID, CMMD, and DINO-FD, the oriented comparator score is the current-to-target distance minus the augmented-to-target distance; for PRDC, LEEP, LogME, and label alignment, it is augmented score minus current score.

Additional comparators use degree-three polynomial KID with ten deterministic matched subsets of at most 100 images, one-versus-rest LogME on standardized DINO features (Bińkowski et al., 2018; You et al., 2021), DINO/CLIP true-label probability and accuracy, multi-scale DINO radius response at $k \in \{1, 5, 10, 20\}$, and nearest-neighbor replay with duplicate threshold 10^{-6}. The duplicate fraction is tied across cells and is recorded rather than used as a rank predictor.

Estimator ablations and the influence baseline. Table 6 reports estimator ablations and the one-step influence baseline on the same supplementary fixed-feature panel. In Table 6, "information equality" keeps the observed bias but forces $J = H$; "sandwich only" removes directed pseudo-true displacement; and "Fisher only" additionally imposes correct-specification score mean and information equality. The one-step row adapts the influence approximation of Koh & Liang (2017) to augmentation: $\widehat{U}_{\mathrm{IF}} = \alpha g_T(\theta_D)^\top A_D^{-1}\{b_S(\theta_D) - b_D(\theta_D)\}$. Here θ_D is fitted on the current set, and A_D is the current-set training Hessian with ridge and numerical damping. The gradient difference accounts for the change from the current distribution to the augmented mixture. The baseline uses an inverse-Hessian solve and keeps the linear change in target loss. The $G = I$ ablation replaces the target-loss Hessian by the identity.

On this panel, full IGRT has NLL MAE 0.00090, compared with 0.00077 for directed displacement only. Its region agreement is higher (0.978 versus 0.933), while dose accuracy is identical (0.867). Thus, including sampling covariance does not improve every prediction metric under this protocol.

Table 6: Estimator ablations and the one-step influence baseline on the 45-cell supplementary fixed-feature CIFAR-10 panel. Spearman and Pearson coefficients compare each estimate with the observed change in target NLL; NLL MAE is reported directly because every row uses the same units. Region agreement uses the ± 0.002 equivalence region, and dose accuracy is computed within each mechanism–block group.

Estimator	Spearman	Pearson	NLL MAE	Region agreement	Dose accuracy
Full IGRT	0.9966	0.9998	0.00090	0.978	0.867
Directed displacement only	0.9958	0.9999	0.00077	0.933	0.867
Observed bias, force $J = H$	0.9951	0.9998	0.00097	0.956	0.867
One-step influence	0.9939	0.9971	0.00392	0.956	0.867
Replace target-loss Hessian by I	0.7722	0.8085	1.07853	0.600	0.733
Sandwich only	0.5437	0.3959	0.03048	0.156	0.400
Fisher only	-0.2511	-0.2847	0.03064	0.156	0.400

B.7 Synthetic data examples

Figures 5–8 show synthetic data examples for classification, semantic segmentation, and object detection, including the associated training supervision.

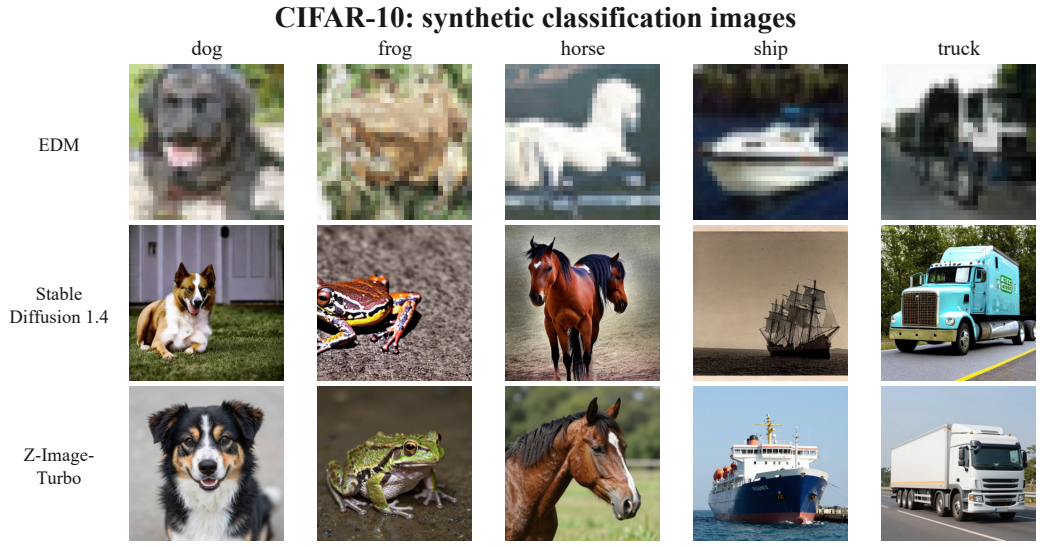

Figure 5: **Synthetic classification examples for CIFAR-10.** Rows show EDM, Stable Diffusion 1.4, and Z-Image-Turbo outputs. Columns correspond to the indicated classes.

B.8 Candidate-set ranking at fixed augmentation budgets

Table 1 evaluates the performance of the selected sets. Here we examine whether each method correctly orders the full candidate menu. For each dataset, we use the candidate sets and augmentation budget from the candidate-set selection experiment. We hold the current data, learner, and training protocol fixed, and compare method scores with observed utility averaged over matched training seeds. Observed utility is the reduction in held-out task loss: NLL for classification, image-averaged pixel NLL for segmentation, and the fixed-weight composite loss for detection.

We measure ranking agreement using Spearman's correlation. All scores are oriented so that larger values indicate greater predicted benefit; ties receive average ranks. Correlations are computed within each fixed-budget comparison, without pooling different augmentation sizes. This isolates the ability

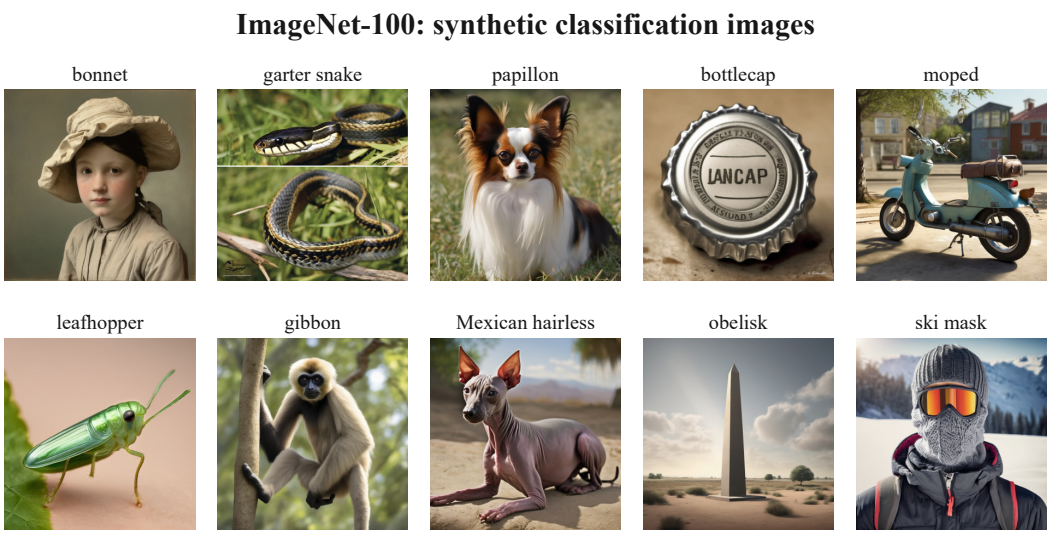

Figure 6: **Synthetic classification examples for ImageNet-100.** Each image is shown with its associated class label.

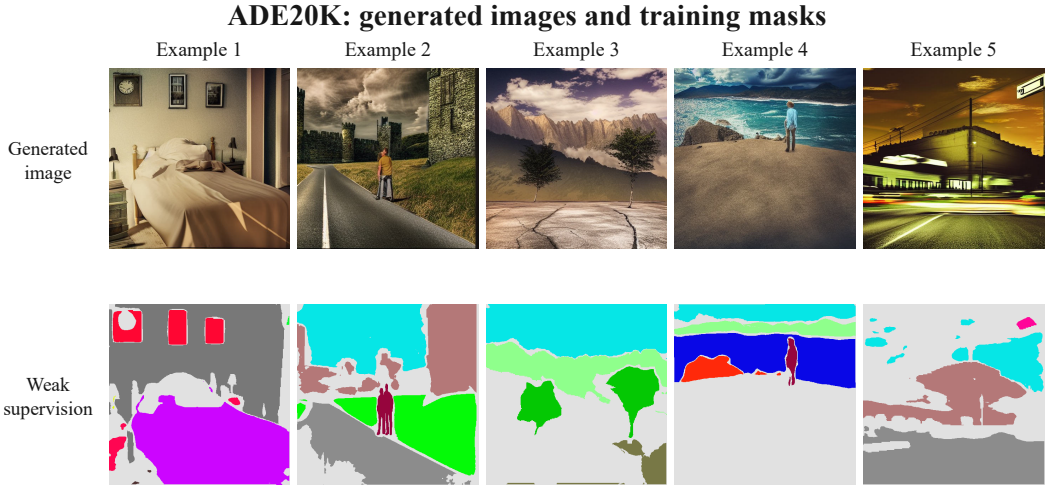

Figure 7: **Synthetic segmentation examples for ADE20K.** The top row shows generated images; the bottom row shows the corresponding supervision masks used for training. Colors distinguish semantic classes.

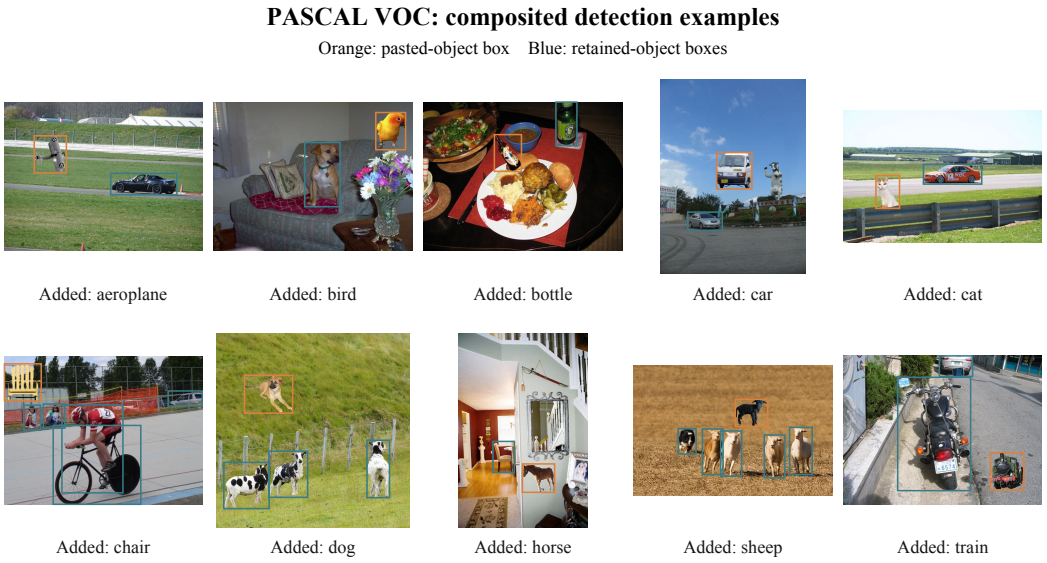

Figure 8: **Detection examples for PASCAL VOC.** Examples are selected for foreground completeness and visibility from the existing group with pasted-cutout long sides set to 25% of the background short side. Orange boxes mark pasted objects, and blue boxes mark retained objects in the background images. Labels below each example identify the added object class.

Table 7: Candidate-set ranking at fixed augmentation budgets. Entries report Spearman's correlation between oriented method scores and observed reductions in held-out task loss (higher is better). Each column uses a common menu of K candidate sets, each containing m added images.

Method	CIFAR-10	ImageNet-100	ADE20K	VOC
Added images m	1000	1000	1000	1000
Candidate sets K	10	8	10	8
FID	$0.18_{\pm 0.23}$	$0.31_{\pm 0.19}$	$-0.08_{\pm 0.26}$	$0.12_{\pm 0.22}$
CMMD	$0.29_{\pm 0.17}$	$0.14_{\pm 0.25}$	$0.16_{\pm 0.21}$	$0.24_{\pm 0.18}$
FD-DINOv2	$0.36_{\pm 0.20}$	$0.19_{\pm 0.22}$	$0.11_{\pm 0.28}$	$0.21_{\pm 0.24}$
PRDC Recall	$0.25_{\pm 0.18}$	$0.38_{\pm 0.16}$	$0.34_{\pm 0.19}$	$0.07_{\pm 0.27}$
First-order influence	$0.47_{\pm 0.19}$	$0.43_{\pm 0.23}$	$0.53_{\pm 0.17}$	$0.39_{\pm 0.21}$
SOURCE	$0.64_{\pm 0.14}$	$0.58_{\pm 0.18}$	$0.59_{\pm 0.15}$	$0.72_{\pm 0.11}$
LWP	$0.42_{\pm 0.16}$	$0.23_{\pm 0.27}$	$0.28_{\pm 0.24}$	$0.18_{\pm 0.29}$
IGRT **(Ours)**	$0.79_{\pm 0.09}$	$0.63_{\pm 0.15}$	$0.57_{\pm 0.20}$	$0.69_{\pm 0.13}$

to distinguish candidate sets from the effect of adding more data. A constant score has undefined correlation and is reported as `n/a`.

B.9 Feature-space visualization of sample selection

Figure 9 complements the downstream evaluation with a visualization of selection behavior. In this comparison, IGRT has a higher fraction of samples marked as gap-supported (25.8%) than the other methods (11.2–13.4%), illustrating differences in selection relative to current-data coverage.

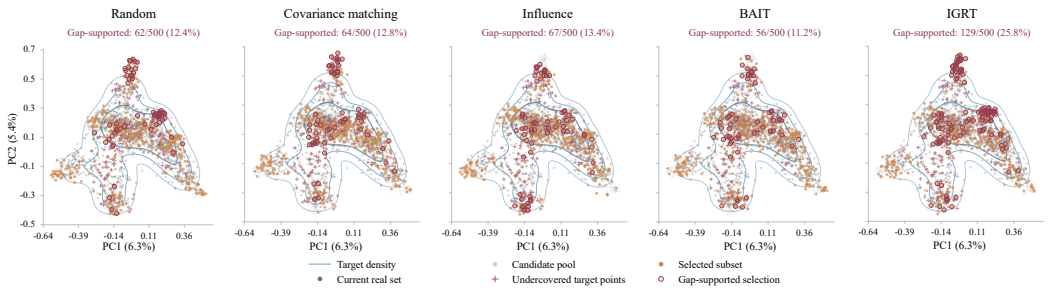

Figure 9: **Feature-space visualization of selected subsets.** Panels compare random sampling, covariance matching, influence, BAIT, and IGRT in the same PCA view. Blue contours show target density; dark and light gray points indicate the current real set and candidate pool, respectively. Orange points show selected samples, red crosses mark undercovered target points, and red circles highlight gap-supported selections. Panel annotations report gap-supported counts and fractions. Axis percentages indicate explained variance.

C Learning Response and Target NLL

We derive prediction-space identities and relate a local residual-to-prediction approximation to the implemented IGRT score.

C.1 Identities in prediction space

First fix the current set, candidate set, and training randomness. For squared loss, let $F_0(T)$ denote raw predictions, $r_T = Y_T - F_0(T)$, and $\Delta F_T = F_{D\oplus S}(T) - F_0(T)$. For a target set of size q,

$$\frac{1}{q}\|Y_T - F_0(T)\|^2 - \frac{1}{q}\|Y_T - F_0(T) - \Delta F_T\|^2 = \frac{2}{q} r_T^\top \Delta F_T - \frac{1}{q}\|\Delta F_T\|^2. \tag{43}$$

This is an exact identity, irrespective of the training procedure. Alignment with target residuals helps only if it exceeds the quadratic cost of the prediction change. With a factor $1/2$ in the loss, both terms are halved.

For classification, F instead contains logits, $P(F)$ contains softmax probabilities, and $r_T = Y_T - P(F_0)$. Define empirical target NLL by

$$\mathcal{L}_T(F) = \frac{1}{q}\sum_{i=1}^{q}\left\{\log\sum_{c=1}^{C}\exp(F_{ic}) - F_{i,y_i}\right\}. \tag{44}$$

Its gradient at F_0 is $-r_T/q$, and its Hessian is $W(F_0)/q$, where $W(F)$ is block diagonal with blocks $\mathrm{diag}(p_i) - p_i p_i^\top$. Taylor's integral identity gives

$$\begin{aligned}\mathcal{L}_T(F_0) - \mathcal{L}_T(F_0 + \Delta F_T) = \frac{1}{q} r_T^\top \Delta F_T \\ - \frac{1}{q}\int_0^1 (1-t)\Delta F_T^\top W(F_0 + t\Delta F_T)\Delta F_T \, dt.\end{aligned} \tag{45}$$

Replacing the curvature along the path by $W(F_0)$ yields Equation 5. If the third derivative of $\mathcal{L}_T$ has operator norm at most M along this path, the absolute remainder is at most $M\|\Delta F_T\|^3/6$. Expected changes over training randomness require averaging the whole identity; the baseline residual and prediction change need not be independent.

C.2 A local residual-to-prediction operator

Let θ_0 be a current-set reference fit, and let L_D, L_S denote unpenalized average NLL on the two sets. Consider training objectives

$$\begin{aligned}L_0^{\mathrm{tr}}(\theta) &= L_D(\theta) + \tfrac{\lambda_0}{2}\theta^\top R\theta, \\ L_\alpha^{\mathrm{tr}}(\theta) &= (1-\alpha)L_D(\theta) + \alpha L_S(\theta) + \tfrac{\lambda_\alpha}{2}\theta^\top R\theta.\end{aligned} \tag{46}$$

Here $R \succeq 0$ is the penalty matrix. Let $e_D = \nabla L_0^{\mathrm{tr}}(\theta_0)$ measure imperfect stationarity, and define the logit Jacobian $Q_B = \partial F_\theta(B)/\partial\theta^\top|_{\theta_0}$. For C classes and p parameters, $Q_T \in \mathbb{R}^{qC\times p}$ and $Q_S \in \mathbb{R}^{mC\times p}$. The softmax chain rule gives

$$\nabla L_S(\theta_0) = -\frac{1}{m} Q_S^\top r_S, \qquad r_S = Y_S - P_{\theta_0}(S). \tag{47}$$

Consequently, the augmented gradient at the current reference point is

$$\nabla L_\alpha^{\mathrm{tr}}(\theta_0) = -\frac{\alpha}{m} Q_S^\top r_S + (1-\alpha)e_D + \{\lambda_\alpha - (1-\alpha)\lambda_0\}R\theta_0. \tag{48}$$

Set $A_\alpha = \nabla^2 L_\alpha^{\mathrm{tr}}(\theta_0) + \lambda_{\mathrm{num}} D_T$. Assume that it is invertible and that the predicted update stays within the local neighborhood. A damped Newton approximation and a first-order expansion of logits then give

$$\widehat{\Delta F_T} \approx \widehat{\mathcal{T}}_{\mathcal{A},D,S} r_S + c_{\mathcal{A},D,S}, \tag{49}$$

with

$$\widehat{\mathcal{T}}_{\mathcal{A},D,S} = \frac{\alpha}{m} Q_T A_\alpha^{-1} Q_S^\top, \tag{50}$$

$$c_{\mathcal{A},D,S} = -Q_T A_\alpha^{-1} \left[(1-\alpha)e_D + \{\lambda_\alpha - (1-\alpha)\lambda_0\} R\theta_0\right]. \tag{51}$$

The (i, j) block of this operator maps the class-label residual on candidate j to a change in the logits of target sample i. Its entries need not be positive. A small image distance or a large candidate residual therefore does not determine the sign of the target-NLL change.

When $\lambda_0 = \lambda_\alpha = 0$ and $e_D = 0$, $c = 0$ and the expression reduces to $\widehat{\Delta F_T} = \widehat{\mathcal{T}}_{\mathcal{A},D,S} r_S$. With fixed average-loss ridge, $\lambda_\alpha = \lambda_0 = \lambda$, even a stationary baseline retains the correction $-\alpha\lambda Q_T A_\alpha^{-1} R\theta_0$. Numerical damping affects the solve but is not added to the gradient.

The inverse Hessian also assumes a local fitted solution, not an arbitrary training budget. With constant Hessian H and step size η, K local gradient steps instead use the response

$$\eta \sum_{k=0}^{K-1} (I - \eta H)^k \tag{52}$$

in place of H^{-1}. Convergence to that inverse requires appropriate stability conditions. Changes in the backbone, batch normalization, or training trajectory require further modeling.

C.3 Connection to the implemented Godambe calculation

The original protocol uses a target-fit reference point $\widehat{\theta}_T$, rather than assuming that θ_0 is a current-set stationary point. It separately estimates $\widehat{\delta}_{P_D}$ and $\widehat{\delta}_{P_\alpha}$ from the exact mixture moments. At this shared point, define

$$\Sigma_P^\theta = \frac{1}{n_P} A_P^{-1} \mathcal{J}_{P,\pi} A_P^{-\top}, \qquad \mu_P^F = Q_T \widehat{\delta}_P, \qquad \Sigma_P^F = Q_T \Sigma_P^\theta Q_T^\top. \tag{53}$$

These are local displacement and data-sampling covariance in target-logit space. The difference in mean displacements gives Equation 16.

Let g_F and G_F be the gradient and Hessian of target NLL with respect to logits at the same reference point. For affine logits, $g_T = Q_T^\top g_F$ and $G_T = Q_T^\top G_F Q_T$. The existing IGRT calculation can therefore be written as

$$\begin{aligned} \widehat{\mathcal{U}}_\alpha = {} & g_F^\top (\mu_{P_D}^F - \mu_{P_\alpha}^F) \\ & + \tfrac{1}{2} \left\{ (\mu_{P_D}^F)^\top G_F \mu_{P_D}^F - (\mu_{P_\alpha}^F)^\top G_F \mu_{P_\alpha}^F \right\} \\ & + \tfrac{1}{2} \operatorname{tr}\{G_F (\Sigma_{P_D}^F - \Sigma_{P_\alpha}^F)\}. \end{aligned} \tag{54}$$

This is an algebraic rewriting of the existing quadratic score, not a separate estimator. The implementation needs contractions of these quantities, not a materialized sample-by-sample response matrix. For nonlinear logits, the parameter Hessian additionally contains $\sum_a (g_F)_a \nabla_\theta^2 \{F_\theta(T)\}_a$; the affine identity cannot be assumed.

The sandwich covariance averages over data sampling under the specified source-count policy. It does not, by itself, quantify optimization-seed variability conditional on fixed data or variation from representation learning.