Assessing Sample Quality in Conditional Generation under Compositional Shift

Berker Demirel[1], **Valentino Maiorca**[1], **Marco Fumero**[1,3], **Theofanis Karaletsos**[2], **Francesco Locatello**[1]
[1]Institute of Science and Technology Austria (ISTA)
[2]Achira
[3]Kempner Institute, Harvard University

Abstract

Conditional generators provide a natural tool for controllable generation, including settings where the desired condition is a new composition of observed attributes or experimental factors. In many applications, especially in scientific domains, such models are attractive to explore conditions for which real samples are rare, expensive, or not yet observed. However, this creates a circularity for evaluation: standard conditional quality metrics require a reference target distribution, but in the extrapolative regime that distribution is unavailable by definition. We address this problem with a post-hoc, per-sample trust score for assessing conditional samples using only the training distribution. The score combines two estimable quantities: global realism, measuring compatibility with the real data manifold, and attribute-wise faithfulness, measuring whether a sample is closer to the requested attributes than to plausible alternatives. We show that the score can recover meaningful comparisons across extrapolated generations, under a mild coverage condition on the observed attributes. These comparisons enable effective filtering, ranking, and abstention of generations and can be used directly on off-the-shelf pretrained models. In biological imaging, selected samples preserve real morphological structure better and improve downstream predictive performance, while similar gains are observed on controlled vision benchmarks. Finally, we show how the score can be applied during generation, enabling abstention before full decoding. Code is available at https://github.com/berkerdemirel/faithful-cond-gen.

1 Introduction

Conditional diffusion models can generate samples for user-specified conditions, including class labels [7, 30], attribute tuples [5], biological perturbations [22], and structured prompts [23]. This is increasingly appealing in scientific settings, where synthetic data are useful not because they reproduce what has already been measured, but because they may help explore conditions that are rare, expensive, or not yet observed [34, 2, 10]. For example, if we want to achieve certain phenotypic properties in a cell population, we would like to prioritize wet-lab testing of perturbations that are most likely to produce the desired outcome. A generative model can support this process by simulating experiments in silico before laboratory validation. The central challenge is determining whether these predictions can be trusted at all, as they must be both visually realistic and faithful to the requested condition, despite the target population being unseen.

Evaluating conditional generation faces a circularity issue: standard distributional metrics such as Fréchet Inception Distance (FID) [11] and Kernel Inception Distance (KID) [3], along with their conditional variants, score generated samples by comparison with real samples from the target distribution. Yet conditional generators are often most useful precisely in settings where such target

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

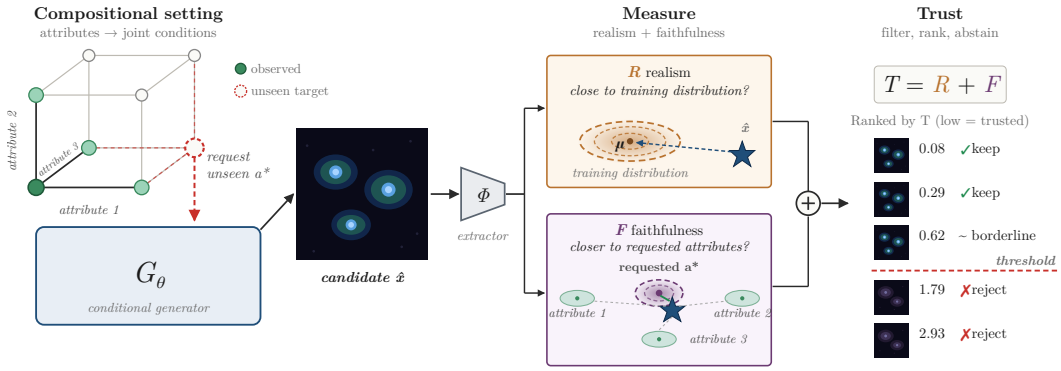


Figure 1: **Pipeline for trust scoring under compositional shift.** A conditional diffusion model G_θ is queried with an unseen joint condition $a^\star$ and produces candidate samples $\hat{x}$. Features from Φ are used to compute a realism term R, which measures proximity to the real training distribution, and a faithfulness term F, which measures alignment with the requested attribute values. Their sum $T = R + F$ provides a calibrated trust score used to rank, filter, or abstain from generated samples.

samples are unavailable. In this extrapolative regime, target-distribution metrics can still serve as controlled validation tools, but they cannot decide whether an individual generated sample should be trusted at deployment time.

We therefore ask a different question: *what sample quality information is identifiable from the real training distribution alone?* We focus on two real-data-calibrated quantities that remain observable. A sample should be *globally realistic*, meaning compatible with the feature geometry of real data, and *attribute-wise faithful*, meaning closer to the requested attribute values than to competing values. These quantities capture separable failure modes: a sample may be realistic but wrong for the requested condition, or condition-consistent but outside the real-data manifold. Existing post-hoc reliability tools typically capture only one side of this split: out-of-distribution and one-class-style scores measure global compatibility with an in-distribution set [33, 10], while prompt- or classifier-based alignment scores measure agreement with a requested condition [20, 17, 28].

Our method combines both views in a post-hoc trust score computed in the latent space of an auxiliary feature extractor, using only the generated sample and real training-set statistics. The key idea is attribute-level decomposition: instead of matching a sample to an unseen condition as a whole, we evaluate each constituent attribute independently, checking whether its requested value is favored over plausible alternatives in observed real-data contexts. Under this coverage condition, we show that attribute-wise decisions remain identifiable even when the full joint target is not. This formalizes both what the score can recover and what remains fundamentally unidentifiable without target samples.

Most notably, we demonstrate that the same real-data geometry allows our score to be evaluated before generation is complete. By aligning the diffusion model to the feature extractor space, we are able to evaluate the same trust score directly in representation space, during generation. This avoids the computational cost of repeated forward passes through the feature extractor and enables low-quality samples to be discarded before full generation.

Our contributions can be summarized as follows:

- **Reference-free trust score.** We recast per-sample quality assessment for conditional generators as a problem solvable without observing the joint target conditioning. Decomposing it into two geometric quantities, global realism and attribute-wise faithfulness, yields our post-hoc trust score computable from a pretrained generator, an encoder, and training-set statistics.

- **Attribute-level identifiability.** We show that meaningful attribute-level comparison scores can be identified even when the full joint target is not.

- **Empirical validation.** On `RxRx1` and `CelebA` our score reliably filters, ranks, and curates samples under support shift under both KID-based and downstream task-based validations.

- **Early sample rejection**. We show how the score can be lifted from a post-hoc method to the latent representations inside the generator itself via a simple geometric alignment objective, enabling quality assessment and abstention before decoding.

2 Related Work

Image quality metrics. There is active literature evaluating generative models by comparing real and generated sample sets in a feature space, including FID [11], KID [3], density/coverage [21], probabilistic precision/recall [26, 31, 13], as well as sample-level metrics closer to our setting [1]. However, these metrics require a collection of real reference samples from the target distribution being evaluated. In contrast, our method combines capabilities that no other method offers as a set: it is per-sample, post-hoc (*i.e.,* after generation, and therefore, model agnostic), and most importantly, it is designed for compositional generalization, where no real samples from the requested target condition are available.

Compositional generalization in conditional diffusion. A growing line of work studies whether conditional diffusion models can generate unseen compositions of attributes or concepts. The closest connection to our setting is [24], which analyzes when compositional abilities emerge in a synthetic discrete conditional-generation task. Related work studies the emergence of hidden compositional capabilities in concept space [25], and analyzes hierarchical compositional generalization in diffusion models through context-free grammars [8]. Other work modifies training objectives to encourage logical or compositional generation under partial support [9]. Our focus is complementary. We do not try to improve the generator or characterize when compositional generalization emerges. Instead, we ask how to evaluate and rank individual generated samples after training, especially when the requested joint condition has no real target reference set. This shifts the question from whether a model can sometimes extrapolate to whether a particular extrapolated sample should be trusted.

OOD detection and support scoring. Out-of-distribution detection and feature-space support estimation provide another line of related work, but they are primarily developed for discriminative prediction rather than as trust mechanisms for synthetic generations. Mahalanobis-based detectors measure compatibility with an in-distribution feature geometry [15, 20], distance-based methods provide nonparametric alternatives [33, 29, 27, 16], and energy-based methods score whether an input lies in a high-confidence region of a learned representation space [17]. We repurpose [15] to build a trust mechanism for conditional diffusion models. For realism, we ask whether a sample looks like real data from the domain; for faithfulness, we ask whether, within a real-data-calibrated geometry, it is closer to the requested attribute value than to its competitors. This complements unconditional support scoring by incorporating the requested conditioning, even when the requested joint target is unavailable.

Faithfulness and alignment scores for generated data. Conditional faithfulness is often evaluated with discriminative surrogates. In compositional-generation studies, probes can test whether generated samples contain requested attributes [24]. In text-to-image settings, CLIP-style alignment is widely used as a proxy for prompt adherence, including when assessing whether synthetic images are useful for recognition tasks [10]. Recent work has also moved alignment scoring into the denoising process, evaluating prompt–latent agreement before a final image is decoded [28]. These approaches are useful discriminative tests of whether a generated sample expresses a requested label, prompt, or attribute, but they are not directly image-quality metrics: an image can be classified as condition-consistent while remaining unrealistic or far from the real data geometry. Our goal is to identify what conditional quality remains estimable from the training distribution alone when the requested joint target is unavailable. This yields a decomposition into global realism and reference-anchored per-attribute faithfulness, where faithfulness is not a standalone classifier score but a conditional component of a real-data-calibrated quality score.

3 Method

We propose a post-hoc per-sample *trust score* for conditional generation when the requested joint condition may be absent from the real training support. The score is computed w.r.t. any individually generated sample, relying only on the training data and a pretrained feature extractor Φ. It combines two measurable quantities: *global realism* (*i.e.,* compatibility with the real-data feature geometry), and *attribute-wise faithfulness*, (*i.e.,* agreement with each requested attribute value relative to competing

ones). While the proposed score is applicable to any pretrained conditional generative model, in the following we assume a latent conditional diffusion model [30]. Figure 1 illustrates our pipeline.

Generative setting. Let $\mathcal{A}_1, \dots, \mathcal{A}_K$ be finite attribute-value spaces, and let $a^\star = (a_1^\star, \dots, a_K^\star) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$ denote the requested joint condition. We consider a latent diffusion model G_θ with latent space $\mathcal{Z}$, input space $\mathcal{X}$, decoder $D : \mathcal{Z} \to \mathcal{X}$, and denoising network ϵ_θ. Starting from Gaussian noise $z_T \in \mathcal{Z}$, the reverse process conditioned on $a^\star$ produces a final latent $\hat{z}_0(a^\star) \in \mathcal{Z}$, which is decoded into the generated sample $\hat{x}(a^\star) = D(\hat{z}_0(a^\star)) \in \mathcal{X}$.

The joint condition $a^\star$ may appear in training or may lie under compositional support shift, where the specific tuple is unseen but its individual attributes appear in other real training samples. Let $\Phi : \mathcal{X} \to \mathbb{R}^d$ denote the feature extractor. We score samples using normalized features $y(x) = \Phi(x)/\|\Phi(x)\|_2$. All means, covariance estimates, precision matrices, and calibration constants below are fit only on normalized real training features.

Global realism. Let μ_{real} and Σ_{real} be the empirical mean and regularized covariance of real normalized features. We define the global Mahalanobis energy

$$E_{\text{real}}(y) = (y - \mu_{\text{real}})^\top \Sigma_{\text{real}}^{-1} (y - \mu_{\text{real}}),$$

and standardize it as $R(y) = (E_{\text{real}}(y) - m_R)/s_R$, where m_R and s_R are the mean and standard deviation of E_{real} on real calibration features. Therefore, larger R indicates lower realism.

Attribute-wise faithfulness. For each attribute k and value $v \in \mathcal{A}_k$, define the real-data prototype $\eta_{k,v} = \mathbb{E}[y \mid a_k = v]$, estimated by averaging the normalized features $y(x)$ of all real training samples whose k-th attribute value is v. For each attribute k, we estimate a shared precision matrix $P_k \succ 0$ from the pooled within-value covariance of real samples and define

$$d_k(y; v) = (y - \eta_{k,v})^\top P_k (y - \eta_{k,v}).$$

For requested value $t \in \mathcal{A}_k$, the target-vs-competitor margin is

$$M_k(y; t) = d_k(y; t) - \min_{v \neq t} d_k(y; v).$$

Thus $M_k(y; t) < 0$ means that y is closer to the requested value than to every competing value of attribute k. We standardize each margin on real samples with $a_k = t$: $F_k(y; t) = (M_k(y; t) - m_{k,t})/s_{k,t}$, where $m_{k,t}$ and $s_{k,t}$ are the corresponding real calibration mean and standard deviation. The total faithfulness score is $F(y; a^\star) = \frac{1}{K} \sum_{k=1}^{K} F_k(y; a_k^\star)$. Larger F indicates weaker agreement with the requested attributes.

The final trust score is $T(y; a^\star) = R(y) + F(y; a^\star)$. The decomposition is interpretable: R detects samples that are globally far from the real feature distribution, while F detects samples that are realistic but mismatched to the requested attributes. Estimation details are given in Section C.

3.1 When can we assess sample quality in practice?

We remark that the same score T can be used both after generation and during generation; only the way we obtain the Φ-compatible features y changes. Figure 5 in Appendix A illustrates the difference.

Post-generation scoring. After generation, we decode the image and apply the feature extractor:

$$y^{\text{post}} = \frac{\Phi(D(\hat{z}_0(a^\star)))}{\|\Phi(D(\hat{z}_0(a^\star)))\|_2}.$$

We then evaluate $T(y^{\text{post}}; a^\star)$. This directly scores the generated image in the feature-space geometry, but requires the VAE decoder and a Φ pass.

During-generation scoring. While generating, we skip decoding and feature extractor encoding by mapping internal denoising representation into a Φ-compatible space. Let $r_{\ell,\tau}(a^\star) \in \mathbb{R}^{d_\ell}$ denote the representation extracted at layer ℓ and denoising step τ along the inference trajectory conditioned

on $a^\star$. We learn a *translator*, parametrized by a shallow MLP, $g_{\phi,\ell} : \mathbb{R}^{d_\ell} \to \mathbb{R}^d$ to match the pairs $\big(r_{\ell,\tau}(x_i, a_i),\ y(x_i)\big)$ obtained from real training samples x_i with observed attribute condition a_i.

$$y^{\mathrm{map}}_{\ell,\tau} = \frac{g_{\phi,\ell}(r_{\ell,\tau}(a^\star))}{\|g_{\phi,\ell}(r_{\ell,\tau}(a^\star))\|_2}.$$

At inference time, we evaluate $T(y^{\mathrm{map}}_{\ell,\tau}; a^\star)$. Thus post-generation and during-generation scoring use the same real-data-calibrated metric; they differ only in whether the feature is obtained from the decoded image or from the denoising trajectory.

Online abstention. The mapped score enables early rejection during denoising. Given a layer-step-specific threshold $\kappa_{\ell,\tau}$, we abstain when

$$T(y^{\mathrm{map}}_{\ell,\tau}; a^\star) > \kappa_{\ell,\tau}.$$

This rejects samples predicted to have low trust before completing the full denoising trajectory, decoding, and post-generation encoder pass. Details of the translator training and compute accounting are given in Section F.

4 Theory: Identifiability under missing target distributions

If the requested condition $a^\star$ is missing from the real data, then the behavior of y under that condition cannot be estimated directly from the data.

Proposition 1 (Informal: non-identifiability of missing targets)**.** *Let $S \subseteq \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$ be the observed support of the conditioning variable a, and let $a^\star \notin S$. From the observed joint distribution of (y, a) restricted to S, the conditional distribution of y at $a^\star$ is not identifiable without assumptions relating observed and unobserved conditions. Therefore, no post-hoc score using only the observed real distribution can certify full conditional fidelity to the missing target distribution.*

This motivates a weaker target: attribute-level comparisons. Instead of asking whether a sample matches the unobserved joint condition $a^\star$, we ask whether it is closer to each requested attribute value than to competing values.

Definition 1 (Reference coverage)**.** *Let a_{-k} denote all attributes except a_k. The observed support has* reference coverage *if there exists a reference condition $\bar{a}$ such that, for every attribute k and value $v \in \mathcal{A}_k$,*

$$\mathbb{P}(a_k = v,\ a_{-k} = \bar{a}_{-k}) > 0.$$

Reference coverage provides just enough conditioning variation to isolate attribute effects without relying on structural assumptions. When attributes only appear in fixed combinations, their effects are support-confounded: the data cannot tell which attribute accounts for an observed change. Reference coverage avoids this ambiguity while using the smallest possible set of observed joint conditions. We formalize this minimality result and give a support-confounding example in Sections B.1 and B.2.

Moreover, under reference coverage, we can identify a fixed-context attribute comparator. Define the reference prototype $\mu^{\mathrm{ref}}_{k,v} := \mathbb{E}[y \mid a_k = v, a_{-k} = \bar{a}_{-k}]$. For a fixed $P_k \succ 0$, define $d^{\mathrm{ref}}_k(y; v) := (y - \mu^{\mathrm{ref}}_{k,v})^\top P_k (y - \mu^{\mathrm{ref}}_{k,v})$, and for target value t,

$$M^{\mathrm{ref}}_k(y; t) := d^{\mathrm{ref}}_k(y; t) - \min_{v \neq t} d^{\mathrm{ref}}_k(y; v).$$

A negative margin means that y is closer to the requested value t than to every competitor for attribute k, in the reference context.

Proposition 2 (Point identification of the reference-anchored comparator)**.** *Assume reference coverage, let y be integrable, and fix $P_k \succ 0$. Then $\mu^{\mathrm{ref}}_{k,v}$, $d^{\mathrm{ref}}_k(y; v)$, and $M^{\mathrm{ref}}_k(y; t)$ are point-identified from the observed joint distribution of (y, a).*

See Section B.3 for the proof. Thus, missing target samples prevent full conditional evaluation, but not all attribute-level evaluation. Reference coverage identifies an anchored comparator.

Theory versus implementation. Proposition 2 establishes point identification for the reference-anchored comparator, whose prototypes $\mu^{\text{ref}}_{k,v}$ are estimated within a fixed observed context. The score used in our experiments is a slightly different object: it replaces these reference-anchored prototypes with pooled prototypes $\eta_{k,v} = \mathbb{E}[y \mid a_k = v]$, estimated across all observed contexts.

We adopt pooling because it uses substantially more real samples per attribute value and is therefore more sample-efficient in practice. Section B.4 gives sufficient conditions under which the pooled and reference-anchored comparators induce the same attribute-wise decisions, while Section B.6 evaluates their agreement empirically. Section B.9 further shows that either comparator can be evaluated in the full feature space without requiring a disentangled subspace, since its decisions depend only on prototype-difference discriminant directions.

Taken together, the theory identifies what can be recovered from incomplete conditional support: under the cardinality-minimal reference-coverage condition, the reference-anchored attribute comparator is point-identified. The pooled score used in practice is a statistically more efficient surrogate whose connection to the identified comparator is established under additional sufficient conditions and assessed empirically.

5 Experiments

We evaluate the trust score in the regime it is designed for: the scorer is calibrated on real training-support features and never uses real samples from the requested target condition at deployment time. See Section E for the implementation details.

Datasets and shift regimes. We choose the fluorescent microscopy images dataset `RxRx1` [34] as the main scientific benchmark. It contains 4 cell types and 1138 siRNA perturbations, yielding 4552 cell-type×perturbation attribute combinations (*i.e.,* conditions). After filtering by support to obtain enough real samples for reliable evaluation, we obtain a test set of 25 conditions unseen at training time and 25 seen ones; full construction details are in Section H. For a more controlled setting we used a subset of `CelebA` [18] which is a compositional testbed with 4 binary attributes (Male, Smiling, Blond_Hair, Eyeglasses), yielding 16 conditions. In this specific setting, training sees only the all-zero reference and the four single-attribute conditions, while all 16 conditions are queried at generation time. Results for models trained with the full conditional coverage are reported in Section L.

Models and scoring spaces. For each dataset and support regime, we train a vanilla conditional diffusion model (SiT-B/2 [19]) and two REPA [35] variants aligned to DINOv3 [32] or SigLIP [36]. The vanilla model is the basic deployment target of the trust score, while the REPA variants test transfer across diffusion models and provide a natural baseline for model-internal scoring due to their alignment training objective (Section 5.2). Post-generation trust scores are computed in the feature extractor representation space: DINOv3 for vanilla models and the corresponding teacher representation for REPA models. During generation, scores are computed using a translator that aligns the diffusion model features at a given layer, to SigLIP features. Additionally, we show a negative result using OpenPhenom [12] features, due to representation collapse in Section G.

Sensitivity to the feature space. The trust score depends on the auxiliary representation Φ, and its performance is not expected to be invariant to the choice of encoder. We investigate this dependence using four pretrained feature extractors: DINOv3, SigLIP, DINOv2-Small, and MAE. All four representations yield positive sample-selection improvements over random subsets. Our analysis suggests that two properties are important for reliable scoring: the representation should preserve condition-discriminative information and its covariance should retain sufficient effective rank for the regularized Mahalanobis geometry to be informative. OpenPhenom failure case illustrates that condition decodability alone is not sufficient when the covariance is highly degenerate. We provide the full encoder ablation and covariance diagnostics in Appendix M.

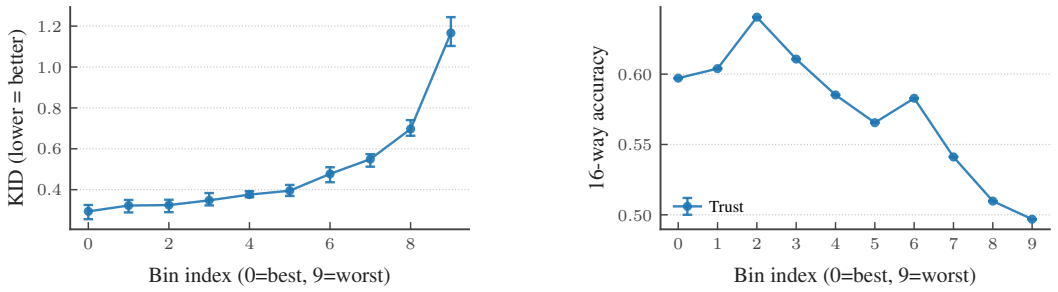
Validation metrics. Generation quality is measured via KID, which assumes access to the real samples from the unseen target conditioning as ground truth, as opposed to our trust score.

$$\Delta\text{KID} = \text{KID}(\text{gen}, \text{real}_A) - \text{KID}(\text{real}_B, \text{real}_A),$$

always evaluated in a fixed DINOv3 mean-patch feature space. On `RxRx1`, we additionally validate trust selection in CellProfiler (CP) [4] morphology space, which provides a biologically interpretable

Table 1: Support-shift post-generation results. P95-real-threshold sample selection uses a threshold set at the 95th percentile of real-sample trust scores; ΔKID is measured in DINOv3 space and compared to a condition-matched random subset. Δ% is the relative ΔKID improvement of trust-selected over random baseline (positive is better). $\rho(T)$ is the condition-level Spearman correlation between mean trust and ΔKID. Component correlations and full-support rows are in Section L.

Dataset	Model	Accept%	$\Delta\text{KID}_{\text{trust}}\downarrow$	$\Delta\text{KID}_{\text{baseline}}\downarrow$	$\Delta\%\uparrow$	$\rho(T)\uparrow$
CelebA	Vanilla	58.7	$0.224_{\pm.009}$	$0.368_{\pm.020}$	+39.1	0.96
	REPA (DINOv3)	55.4	$0.228_{\pm.010}$	$0.401_{\pm.031}$	+43.1	0.96
	REPA (SigLIP)	56.1	$0.239_{\pm.018}$	$0.423_{\pm.020}$	+43.6	0.96
RxRx1	Vanilla	4.4	$0.196_{\pm.006}$	$0.323_{\pm.013}$	+39.4	0.90
	REPA (DINOv3)	5.5	$0.197_{\pm.006}$	$0.333_{\pm.019}$	+40.8	0.90
	REPA (SigLIP)	6.1	$0.181_{\pm.003}$	$0.312_{\pm.019}$	+42.0	0.87



(a) CelebA ΔKID by trust decile.

(b) CelebA downstream 16-way accuracy.

Figure 2: **CelebA decile binning (REPA-DINOv3 held-out, DINOv3 scoring).** ΔKID increases monotonically from bin 0 (best trust) to bin 9 (worst) (*left*), and correlates with downstream classification accuracy drops (*right*). Binning results with faithfulness and realism components, together with RxRx1 DINOv3 decile curves are reported in Section L

.

per-cell measurements independent of DINOv3. Since KID depends on the representation space, we additionally evaluate whether the observed selection improvements when feature space is changed from DINOv3 to SigLIP. Both evaluation spaces yield positive improvements for the tested CelebA and RxRx1 settings (Appendix M.3).

5.1 Post-generation trust as data curation

We first evaluate the score in the post-generation setting. The central question is whether a score calibrated only on real training-support data can curate generated samples for requested conditions whose target distribution is unavailable at scoring time.

5.1.1 Real-data-calibrated filtering and condition ranking

Following the convention commonly used in the out-of-distribution detection literature with the FPR95 metric [33, 6, 20, 17], we instantiate the abstention rule from Section 3.1 by setting the acceptance threshold at the 95th percentile of real-sample trust scores and comparing the accepted subset against a condition-matched random subset of the same size. Table 1 compares the ΔKID of selected generated samples based on trust score (denoted as ΔKID_{trust}) against a condition-matched random subset of generated samples of the same size (denoted as $\Delta\text{KID}_{baseline}$). It reports the percentage of generated samples that pass the FP95-real trust threshold, and the per-condition Spearman correlation between mean trust score and ΔKID.

The trust score consistently selects better samples under support shift. On CelebA, the selected subsets improve ΔKID by 39–44% while retaining roughly 55–59% of generated samples. On RxRx1, selection is much more stringent, accepting only 4–6% of samples, but the ΔKID improvement remains strong at 39–42%. Beyond sample selection, the same table shows that per-condition mean trust provides a strong ranking of condition difficulty in both domains. Accordingly, we report the full realism/faithfulness decomposition in Section L and defer the head-to-head scorer ablation to Section K.

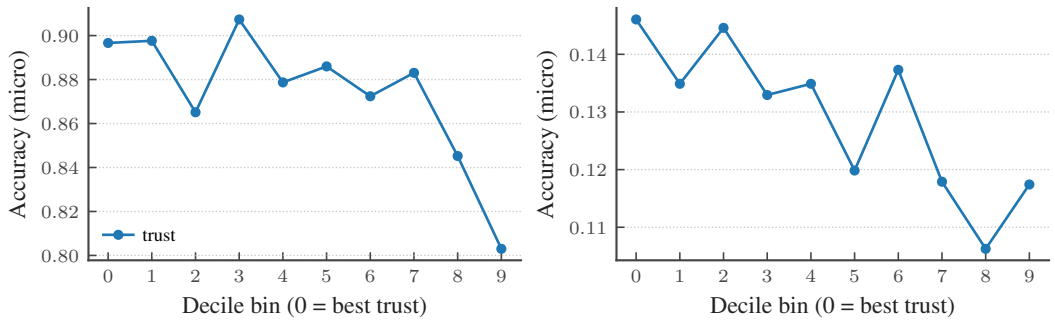


Figure 3: **Main RxRx1 CellProfiler validation (REPA-SigLIP marginal, SigLIP trust scoring).** CP-space downstream classification by trust decile. *Left*: 4-way cell-type accuracy. *Right*: 50-way condition accuracy. Trust-ranked deciles show a clear correlation with the classification performance, showing that trust ordering improves utility in an interpretable morphology space independent of the DINOv3 validation encoder.

Acceptance–quality trade-off. The low acceptance rates on RxRx1 reflect the conservative P95-real thresholding, which prioritizes sample quality over coverage. To investigate less restrictive selection, we vary the acceptance threshold. Relaxing the threshold from P95 to P99.5 increases acceptance rates, but reduces the relative ΔKID improvements. Furthermore, an independently trained siRNA classifier consistently assigns higher accuracy to accepted than rejected generations across the tested thresholds. These results demonstrate that the score supports different thresholding points, with increased coverage coming at the cost of a weaker average quality advantage. We report the complete threshold sweep in Appendix N.

Takeaway. Our score provides a practical support-shift filter and a useful condition-level ordering without using target real samples at scoring time.

5.1.2 Trust rankings track sample quality and downstream utility

We next show how our score can be employed as a full ranker, rather than a filterer. We partition generated samples into class-balanced trust deciles, train a condition classifier on each decile separately, and evaluate on held-out real data. Figure 2 shows the controlled CelebA held-out worse trust deciles have worse ΔKID and yield less useful training data for downstream condition classification.

On CelebA, ΔKID degrades by roughly $4\times$ from the best to worst decile, while downstream accuracy decreases by about 10% when training on low-quality samples according to trust. This confirms that the score is ordinally meaningful in a setting where the support shift is controlled by design. In the more realistic scientific setting, the analogous `RxRx1` DINOv3-space decile curves show the same ordering for KID and downstream classification; we report them in Figure 9 of Section L. We additionally examine how trust scores and generation quality vary with compositional distance on CelebA in Section J.

Takeaway. Across all tested settings, trust score induces a sample-level ordering that tracks both image quality and downstream utility, rather than simply serving as a thresholding rule.

5.1.3 Biological validation with CellProfiler morphology

To test whether sample selection based on trust score captures biologically meaningful morphology, we evaluate `RxRx1` generations in CellProfiler [4] (CP) feature space: interpretable per-cell measurements computed directly from segmented images. We use 621 CP features after variance thresholding, and a real-data $|z| \leq 5$ outlier filter, then standardize on real data. The full CP pipeline is in Section I.

CP-space decile utility. We repeat the decile downstream experiment of Figure 10a, on `RxRx1`, using CP feature space as opposed to DINOv3. Figure 3 shows results with the REPA-SigLIP diffusion model: better trust deciles yield better classifiers for both cell type classification (4 classes) and cell-type$\times$perturbation targets (50 classes in total) show a similar trend. The same pattern holds for all three marginal generators and both targets in Section I.

Real-spread-normalized centroid distance. For each generated sample at condition (c, s), we compute its CellProfiler distance to the real-condition centroid and standardize it by real-real variation:

$$z(x) = \frac{\|\mathrm{CP}(x) - \mu_{\mathrm{real},c,s}\|_2 - \mu_{\mathrm{RR}}(c,s)}{\sigma_{\mathrm{RR}}(c,s)}.$$

Here $\mu_{\mathrm{RR}}(c,s)$ and $\sigma_{\mathrm{RR}}(c,s)$ are estimated from random half-splits of real samples within the same condition. Thus $z = 0$ corresponds to typical real-real morphology spread, while $z > 0$ indicates excess deviation from the real morphology centroid.

Table 2 shows that trust selection improves CellProfiler morphology alignment in all six model/split settings, with bootstrap confidence intervals strictly below zero (Section I, Table 12). The seen split is nearly real-like after selection: REPA-DINOv3 trust-selected samples are only 0.024σ beyond typical real morphology. On unseen conditions, generated samples drift farther from the real distribution, but trust selection still removes 18–29% of the standardized excess distance.

We also provide a per-feature CP breakdown in Section I, where gains concentrate on biologically meaningful texture and shape axes: per-cell texture autocorrelation, image-level granularity, total cell-segmentation area, and cytoplasm shape regularity.

Table 2: Real-spread-normalized CP morphology distance on RxRx1 (kept-621 CP features). $z(x)$ is the distance to the matched real-condition centroid, standardized by its real-real morphology variation. Accept% is the share of generations passing the P95-real threshold. Bootstrap 95% CIs are in Table 12.

Model	Split	Accept%	$\bar{z}_{\mathrm{trust}}\downarrow$	$\bar{z}_{\mathrm{baseline}}\downarrow$	$\Delta\%\uparrow$
Vanilla	seen	66.3	0.055	0.111	+50
Vanilla	unseen	21.9	0.583	0.712	+18
REPA (DINOv3)	seen	61.2	0.024	0.137	+83
REPA (DINOv3)	unseen	21.0	0.535	0.755	+29
REPA (SigLIP)	seen	65.7	0.032	0.126	+75
REPA (SigLIP)	unseen	23.5	0.827	1.020	+19

Takeaway. On `RxRx1`, T-selected samples are closer to real ones in both calibrated CellProfiler morphology space and under the learned DINOv3 validation encoder. Since the metric accounts for within-condition real-real variation, gains directly reflect reductions in excess morphology deviation beyond normal biological spread.

5.2 During generation

In this section, we investigate the extent to which the score T can be evaluated at generation time before decoding so that low confidence images are never generated, directly mapping diffusion-model representations into the feature extractor space (Section 3.1).

Figure 4 evaluates the translator at intermediate denoising steps on `CelebA` held-out (250-step sampler). At $k = 83$ the filter already reaches 33% ΔKID improvement while saving 67% of denoising compute; by $k \approx 166$ it matches the post-generation oracle while still saving 34%. `RxRx1` figure, full-support rows, decile-binning confirmation, and per-step numbers are in Section L.3, Section L.1, Section L, and Section L.2.

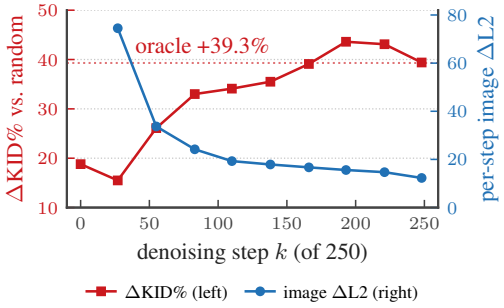


Figure 4: **Translator scoring across denoising (`CelebA`, Vanilla SiT-B/2, 250 steps).** P95-real-threshold ΔKID improvement rises as the predicted-clean trajectory settles.

We train a lightweight *translator* for this mapping, with an objective that preserves the shared-covariance Mahalanobis geometry the score relies on, allowing internal-feature trust scoring without the VAE decoder or feature extractor.

Table 3 reports the support-shift held-out result. The translator gives 25–34% ΔKID improvements on `RxRx1` with $\rho(T)$ up to 0.88, and matches post-generation filtering on `CelebA`. Here, REPA-aligned features serve as a natural baseline: REPA training aligns internal states to a pretrained teacher as part of the generator's loss, so any REPA-trained generator provides a translator at no additional cost. This baseline partly preserves the signal on `CelebA` ($\rho(T) = 0.95$, +31% ΔKID) but fails on `RxRx1`, with selection worse than random and $\rho(T) \approx 0$. Cosine distance, as used in REPA, aligns internal states to the extractor features without preserving the covariance structure needed by the Mahalanobis trust score. Thus,

Table 3: **Trust scoring during generation**. Same setup as Table 1 with the score evaluated from internal features mapped by a translator instead of post-generation SigLIP ones. The joint-aligned REPA baseline partly works on `CelebA` but fails on `RxRx1`; our translator restores filtering and ranking on both.

	Feature source	Model	Accept%	$\Delta\text{KID}_{\text{trust}}\downarrow$	$\Delta\text{KID}_{\text{baseline}}\downarrow$	$\Delta\%\uparrow$	$\rho(T)\uparrow$
`CelebA`	REPA	REPA (DINOv3)	62.7	$0.278_{\pm.015}$	$0.404_{\pm.021}$	$+31.2$	0.95
	REPA	REPA (SigLIP)	66.7	$0.293_{\pm.024}$	$0.425_{\pm.018}$	$+31.1$	0.95
	translator	Vanilla	23.6	$0.195_{\pm.008}$	$0.378_{\pm.023}$	$+48.4$	0.83
	translator	REPA (DINOv3)	48.4	$0.250_{\pm.016}$	$0.397_{\pm.017}$	$+37.2$	0.92
	translator	REPA (SigLIP)	17.2	$0.308_{\pm.015}$	$0.409_{\pm.015}$	$+24.8$	0.87
`RxRx1`	REPA	REPA (DINOv3)	1.4	$0.537_{\pm.032}$	$0.329_{\pm.037}$	-63.4	0.09
	REPA	REPA (SigLIP)	3.9	$0.564_{\pm.022}$	$0.293_{\pm.023}$	-92.3	0.13
	translator	Vanilla	38.4	$0.260_{\pm.003}$	$0.346_{\pm.016}$	$+24.9$	0.78
	translator	REPA (DINOv3)	62.3	$0.246_{\pm.003}$	$0.372_{\pm.037}$	$+33.8$	0.88
	translator	REPA (SigLIP)	60.5	$0.235_{\pm.011}$	$0.329_{\pm.018}$	$+28.7$	0.70

REPA features are not a reliable geometry for trust scoring. We further assess how well the translator preserves post-hoc sample rankings in Section F.2.

Takeaway. Preserving the trust geometry during generation enables early scoring; the same map further allows the score to be evaluated well before final decoding, increasing computational benefits.

6 Conclusion

In this work, we studied per-sample quality assessment for conditional generation under compositional shift, where real samples from the requested target condition are unavailable. We showed that full conditional fidelity to a missing target distribution is not identifiable from the observed training distribution alone, but that global realism and attribute-wise faithfulness remain estimable. We introduced a post-hoc trust score combining a pooled Mahalanobis realism term with shared-covariance attribute margins. Under reference coverage, reference-anchored attribute comparisons are identifiable. Empirically, the score supports filtering, ranking, and synthetic-data curation, improving KID, downstream utility, and RxRx1 CellProfiler morphology alignment. The same real-data geometry can also be used during generation by mapping denoising states into the pretrained-encoder feature space, enabling early abstention when the trust geometry is preserved. A limitation of the current theory is that it focuses on discrete attributes under sufficient coverage conditions. Future work should extend the framework to continuous conditioning variables, characterize weaker identifiability conditions, and use the score to drift generations toward higher-trust regions rather than only abstaining from poor ones.

Acknowledgements

This work was supported by the Chan Zuckerberg Initiative (CZI) through its AI Residency Program. We are grateful to CZI for the opportunity to take part in the program, and to the CZI AI Infrastructure Team for providing support with the GPU cluster used to train our models. MF was supported by the MSCA IST-Bridge fellowship which has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101034413.

References

[1] 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 *International conference on machine learning*, pages 290–306. PMLR, 2022.

[2] 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.

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

[4] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7:R100, 2006.

[5] Berker Demirel, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Morphgen: Controllable and morphologically plausible generative cell-imaging. *arXiv preprint arXiv:2510.01298*, 2025.

[6] Berker Demirel, Marco Fumero, and Francesco Locatello. Out-of-distribution detection with relative angles. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

[8] Alessandro Favero, Antonio Sclocchi, Francesco Cagnetta, Pascal Frossard, and Matthieu Wyart. How compositional generalization and creativity improve as diffusion models are trained. *arXiv preprint arXiv:2502.12089*, 2025.

[9] Sachit Gaudi, Gautam Sreekumar, and Vishnu Boddeti. Coind: Enabling logical compositions in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[10] 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.

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

[12] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11757–11768, 2024.

[13] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

[14] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[15] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[16] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. In *Forty-first International Conference on Machine Learning*, 2024.

[17] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015.

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

[20] Maximilian Müller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization. In *Forty-second International Conference on Machine Learning*, 2025.

[21] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.

[22] Zeinab Navidi, Jun Ma, Esteban Miglietta, Le Liu, Anne E Carpenter, Beth A Cimini, Benjamin Haibe-Kains, and Bo Wang. Morphodiff: Cellular morphology painting with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[23] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022.

[24] Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[25] Core F Park, Maya Okawa, Andrew Lee, Hidenori Tanaka, and Ekdeep S Lubana. Emergence of hidden capabilities: Exploring learning dynamics in concept space. *Advances in Neural Information Processing Systems*, 37:84698–84729, 2024.

[26] Dogyun Park and Suhyun Kim. Probabilistic precision and recall towards reliable evaluation of generative models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20099–20109, 2023.

[27] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1695, 2023.

[28] Vasco Ramos, Regev Cohen, Idan Szpektor, and Joao Magalhaes. Beyond the noise: Aligning prompts with latent representations in diffusion models. *arXiv preprint arXiv:2512.08505*, 2025.

[29] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

[30] 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*, pages 10684–10695, 2022.

[31] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

[32] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[33] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pages 20827–20840. PMLR, 2022.

[34] Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, et al. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4285–4294, 2023.

[35] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025.

[36] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

A Model Pipeline

Figure 5 contrasts post-generation and during-generation trust scoring. Post-generation scoring first completes denoising, decodes the final latent into an image, and applies the feature extractor Φ before evaluating the trust score. During-generation scoring instead maps an intermediate diffusion representation h_t into the same Φ-compatible space using a learned translator T, bypassing decoding and feature extraction. Thus, both settings use the same calibrated trust geometry, while during-generation scoring enables assessment before full generation is completed.

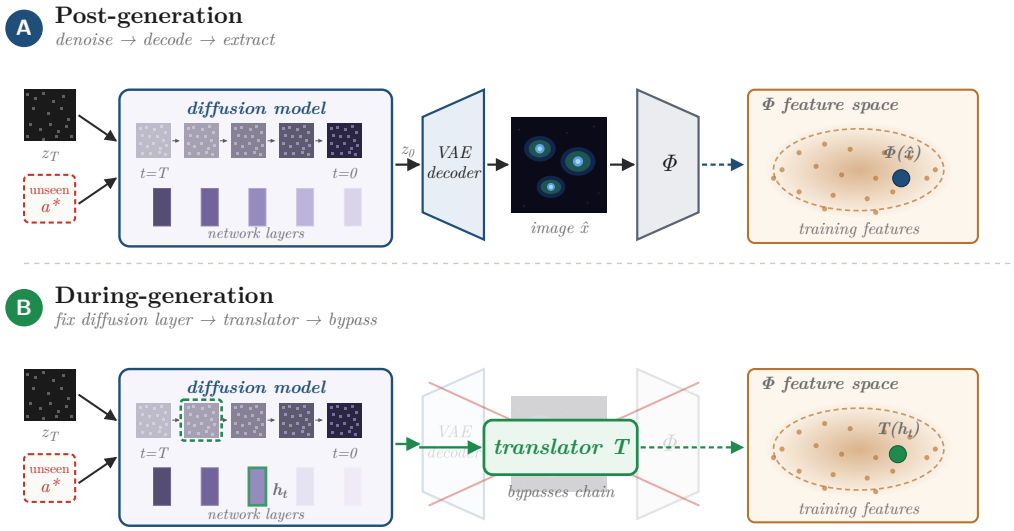


Figure 5: *Post-generation and during-generation trust scoring.* In the post-generation setting, the diffusion model is run to completion, the final latent is decoded into an image, and the feature extractor Φ maps the generated sample into the calibrated feature space where the trust score is evaluated. In the during-generation setting, scoring bypasses decoding and feature extraction: an intermediate denoising representation h_t from a fixed diffusion layer, and selected on a possibly early diffusion timestep t, is mapped by a learned translator T into the same Φ-compatible feature space. Both pipelines therefore use the same real-data-calibrated trust geometry, but during-generation scoring enables early assessment and abstention before full generation is completed.

B Additional theoretical details

B.1 Minimality for unconfounded attribute coverage

Proposition 3 (Reference coverage is minimal)**.** *Let $S \subseteq \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$ be an observed support. Define the one-attribute contrast graph G_S to have vertices $a \in S$, with an edge between two vertices if they differ in exactly one attribute. Without structural assumptions on $\mathbb{E}[y \mid a]$, suppose S permits unconfounded identification of every attribute value effect, in the sense that G_S contains a connected subgraph whose vertices cover every value in each $\mathcal{A}_k$. Then*

$$|S| \geq 1 + \sum_{k=1}^{K} (|\mathcal{A}_k| - 1) = N_{\text{ref}}.$$

Reference coverage attains this bound, since it contains exactly N_{ref} joint conditionings. Hence reference coverage is cardinality-minimal.

Proof of Proposition 3. Let

$$m(a) := \mathbb{E}[y \mid a]$$

denote the conditional mean at attribute configuration a. With no structural assumptions, $m(a)$ is an arbitrary function of the full joint attribute configuration.

A model-free contrast can be attributed to a single attribute only when the two conditions being compared differ in that attribute alone. If two observed conditions differ in multiple attributes, then the resulting mean difference can be decomposed in multiple ways among the changed attributes and their interactions. Thus, without structural assumptions, such a contrast is support confounded.

Define the one-attribute contrast graph G_S of an observed support $S \subseteq \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$ as follows. The vertices are the observed joint conditions $a \in S$. Two vertices are connected by an edge labeled k if they differ only in coordinate k. Such an edge gives an unconfounded contrast for attribute k.

Suppose an observed support identifies unconfounded effects for all attribute values. Then its one-attribute contrast graph must contain a connected subgraph whose vertices cover every attribute value: for every attribute k and every value $v \in \mathcal{A}_k$, some vertex in the subgraph must satisfy $a_k = v$. Otherwise, some value would either be unobserved in any one-attribute comparison or would be observed only through joint changes with other attributes, leaving its individual contribution non-identifiable.

Now consider any connected subgraph T of G_S that covers all attribute values. Let R be a spanning tree of T, and choose an arbitrary root condition $a^{(0)}$. The root covers exactly one value of each attribute, hence K attribute values in total. Each additional vertex in the tree differs from its parent in exactly one coordinate, so it can introduce at most one previously unseen attribute value. The total number of attribute values that must be covered is

$$\sum_{k=1}^{K} |\mathcal{A}_k|.$$

After the root, the number of remaining attribute values that must be introduced is therefore

$$\sum_{k=1}^{K} |\mathcal{A}_k| - K = \sum_{k=1}^{K} (|\mathcal{A}_k| - 1).$$

Since each new observed joint condition can introduce at most one such value, any support that identifies all attribute values through unconfounded one-attribute contrasts must contain at least

$$1 + \sum_{k=1}^{K} (|\mathcal{A}_k| - 1) = N_{\mathrm{ref}}$$

observed joint conditions.

Reference coverage attains this lower bound. Indeed, for a reference condition $\bar{a}$, it observes the baseline condition $\bar{a}$ and, for every attribute k and every non-reference value $v \neq \bar{a}_k$, the one-attribute deviation

$$(v, \bar{a}_{-k}).$$

Thus it contains exactly

$$1 + \sum_{k=1}^{K} (|\mathcal{A}_k| - 1) = N_{\mathrm{ref}}$$

joint conditions. These conditions form a star-shaped one-attribute contrast graph centered at $\bar{a}$, and the graph covers every attribute value. Therefore reference coverage is cardinality-minimal.

This minimal design is not unique: other supports with N_{ref} conditions may also cover all attribute values through unconfounded one-attribute contrasts. Reference coverage is the shared-context construction that attains the minimum. □

B.2 Support confounding without reference coverage

Proposition 4 (Support confounding destroys attribute attribution)**.** *There exist observed supports for which attribute-specific contributions are not determined by the observed distribution. In particular, with four binary attributes and observed support*

$$S = \{0000, 1111\},$$

the observed data distribution determines only the joint contrast between 0000 *and* 1111*; it does not determine any unique decomposition of that contrast into attribute-specific contributions.*

Proof of Proposition 4. It suffices to consider a scalar feature coordinate. Let the observed support be

$$S = \{0000, 1111\},$$

and write

$$m_0 := \mathbb{E}[y \mid a = 0000], \qquad m_1 := \mathbb{E}[y \mid a = 1111].$$

The observed data distribution determines m_0 and m_1, hence only the joint difference $m_1 - m_0$.

Now consider an additive representation on the observed support,

$$\mathbb{E}[y \mid a] = \beta_0 + \sum_{j=1}^{4} g_j(a_j).$$

On the two observed cells, this implies

$$m_1 - m_0 = \sum_{j=1}^{4} \big(g_j(1) - g_j(0)\big).$$

Only the sum is determined; the individual differences are not.

For example, the following two choices agree on the observed support but assign different attribute-1 contributions.

Model A. Set

$$g_1(1) - g_1(0) = m_1 - m_0, \qquad g_j(1) - g_j(0) = 0 \quad (j = 2, 3, 4).$$

Model B. Set

$$g_j(1) - g_j(0) = \frac{m_1 - m_0}{4} \qquad (j = 1, 2, 3, 4).$$

Both models reproduce exactly the same observed conditional means at 0000 and 1111, but they attribute different portions of the observed change to attribute 1. Hence the observed distribution on S does not determine any unique attribute-specific contribution. This is the support confounding claimed in the proposition. □

B.3 Identification and well-definedness of anchored objects

We collect here the basic support-side consequences used in the main text.

Proof of Proposition 2. Fix an attribute k and values $u, v \in \mathcal{A}_k$. Under reference coverage,

$$\mathbb{P}(a_k = v,\ a_{-k} = \bar{a}_{-k}) > 0 \qquad \text{and} \qquad \mathbb{P}(a_k = u,\ a_{-k} = \bar{a}_{-k}) > 0.$$

Hence the conditional expectations

$$\mu^{\mathrm{ref}}_{k,v} = \mathbb{E}[y \mid a_k = v,\ a_{-k} = \bar{a}_{-k}] \qquad \text{and} \qquad \mu^{\mathrm{ref}}_{k,u} = \mathbb{E}[y \mid a_k = u,\ a_{-k} = \bar{a}_{-k}]$$

are functions of the observed joint distribution of (y, a) on positive-probability events, and are therefore point-identified whenever y is integrable. Their difference

$$\Delta^{\mathrm{ref}}_k(v, u) = \mu^{\mathrm{ref}}_{k,v} - \mu^{\mathrm{ref}}_{k,u}$$

is then point-identified as well. Since P_k is fixed, the oracle distances are deterministic functions of the identified prototypes. The oracle margin is a finite minimum over identified distances, and is therefore also identified. □

Proposition 5 (Well-definedness under reference coverage)**.** *Assume the support has reference coverage with reference condition $\bar{a}$. Then for every attribute $k \in \{1, \ldots, K\}$ and every value $v \in \mathcal{A}_k$,*

$$\mathbb{P}(a_k = v,\ a_{-k} = \bar{a}_{-k}) > 0 \qquad \textit{and} \qquad \mathbb{P}(a_k = v) > 0.$$

Consequently, whenever y is integrable, both

$$\mu^{\mathrm{ref}}_{k,v} := \mathbb{E}[y \mid a_k = v,\ a_{-k} = \bar{a}_{-k}] \qquad \textit{and} \qquad \eta_{k,v} := \mathbb{E}[y \mid a_k = v]$$

are well-defined. If in addition $P_k \succ 0$, then for every $y \in \mathbb{R}^d$,

$$d_k^{\mathrm{ref}}(y; v) := (y - \mu_{k,v}^{\mathrm{ref}})^\top P_k (y - \mu_{k,v}^{\mathrm{ref}}), \qquad d_k(y; v) := (y - \eta_{k,v})^\top P_k (y - \eta_{k,v})$$

and the corresponding margins

$$M_k^{\mathrm{ref}}(y; t) := d_k^{\mathrm{ref}}(y; t) - \min_{v \neq t} d_k^{\mathrm{ref}}(y; v), \qquad M_k(y; t) := d_k(y; t) - \min_{v \neq t} d_k(y; v)$$

are well-defined for all targets $t \in \mathcal{A}_k$.

Proof. Reference coverage states directly that

$$\mathbb{P}(a_k = v,\ a_{-k} = \bar{a}_{-k}) > 0$$

for every k and v. Marginalizing over a_{-k} then gives

$$\mathbb{P}(a_k = v) \geq \mathbb{P}(a_k = v,\ a_{-k} = \bar{a}_{-k}) > 0.$$

Hence both conditioning events have positive probability. Since y is integrable, the conditional expectations defining $\mu_{k,v}^{\mathrm{ref}}$ and $\eta_{k,v}$ are well-defined. If $P_k \succ 0$, then the quadratic forms $d_k^{\mathrm{ref}}(y; v)$ and $d_k(y; v)$ are finite for every $y \in \mathbb{R}^d$, and since $\mathcal{A}_k$ is finite, the minima over $v \neq t$ are attained and finite. Therefore the margins are well-defined. □

Remark 1 (Identification versus estimation)**.** *Reference coverage is a support-side condition ensuring that the fixed-context comparison $\mu_{k,v}^{\mathrm{ref}}$ is identified and well-defined. It does not by itself guarantee low-variance estimation in finite samples. Finite-sample issues enter through how many real samples populate each required slice, and are separate from the population-level argument.*

B.4 Pooled and reference prototypes

The main identifiability result concerns the reference-anchored prototypes $\mu_{k,v}^{\mathrm{ref}}$. The implemented score uses pooled prototypes $\eta_{k,v}$ because they are more sample-efficient: they use all real samples with $a_k = v$, rather than only samples in the anchored cell $(a_k = v, a_{-k} = \bar{a}_{-k})$. This appendix gives sufficient conditions under which this lower-variance pooled comparator agrees with the reference-anchored comparator.

Conditional-mean decomposition and pooling assumptions. The pooled prototype for attribute k and value v is

$$\eta_{k,v} := \mathbb{E}[y \mid a_k = v].$$

For the pooled-reference analysis, write the conditional mean as

$$\mathbb{E}[y \mid a_k = v,\ a_{-k} = b] = s_{k,v} + c_k(b) + \delta_{k,v}(b),$$

where $s_{k,v}$ is the value-associated signal, $c_k(b)$ is the remaining-context contribution, and $\delta_{k,v}(b)$ is a value-dependent interaction residual. This decomposition is only an analysis device; the score does not require estimating these terms.

We use the following two assumptions:

$$\mathbb{E}[c_k(a_{-k}) \mid a_k = v] = \bar{c}_k \qquad \text{for all } v \in \mathcal{A}_k, \tag{A1}$$

and

$$\mathbb{E}[\delta_{k,v}(a_{-k}) \mid a_k = v] = 0 \qquad \text{for all } v \in \mathcal{A}_k. \tag{A2}$$

A1 is the substantive context-balance condition: after pooling over observed contexts, the average non-k context contribution does not depend on the value of attribute k. A2 is a centering convention rather than an additional support assumption: any value-specific average interaction bias can be absorbed into $s_{k,v}$.

Taking expectation over $a_{-k} \mid a_k = v$, we obtain

$$\eta_{k,v} = s_{k,v} + \mathbb{E}[c_k(a_{-k}) \mid a_k = v] + \mathbb{E}[\delta_{k,v}(a_{-k}) \mid a_k = v].$$

Under A1–A2, this reduces to

$$\eta_{k,v} = s_{k,v} + \bar{c}_k.$$

At the reference context,

$$\mu_{k,v}^{\mathrm{ref}} = s_{k,v} + c_k(\bar{a}_{-k}) + \delta_{k,v}(\bar{a}_{-k}).$$

Thus pooled prototypes share the same value signal $s_{k,v}$ as the reference prototypes, but differ by a shared context shift and a value-dependent reference leakage term. The next lemma makes this relation explicit.

Lemma 1 (Pooled-reference decomposition)**.** *Under A1–A2,*

$$\eta_{k,v} = \mu^{\mathrm{ref}}_{k,v} + \gamma^{\mathrm{ref}}_{k} - e^{\mathrm{ref}}_{k,v},$$

where

$$\gamma^{\mathrm{ref}}_{k} = \bar{c}_k - c_k(\bar{a}_{-k}), \qquad e^{\mathrm{ref}}_{k,v} = \delta_{k,v}(\bar{a}_{-k}).$$

Proof. By the conditional-mean decomposition,

$$\eta_{k,v} = s_{k,v} + \mathbb{E}[c_k(a_{-k}) \mid a_k = v] + \mathbb{E}[\delta_{k,v}(a_{-k}) \mid a_k = v].$$

Using A1–A2 gives

$$\eta_{k,v} = s_{k,v} + \bar{c}_k.$$

At the reference context,

$$\mu^{\mathrm{ref}}_{k,v} = s_{k,v} + c_k(\bar{a}_{-k}) + \delta_{k,v}(\bar{a}_{-k}).$$

Subtracting yields

$$\eta_{k,v} = \mu^{\mathrm{ref}}_{k,v} + \bar{c}_k - c_k(\bar{a}_{-k}) - \delta_{k,v}(\bar{a}_{-k}).$$

□

B.5 Pooled-reference perturbation bound

Proposition 6 (Pooled margin perturbation, value-specific split form)**.** *Assume A1–A2 and let*

$$\tilde{y} = y - \gamma^{\mathrm{ref}}_{k}.$$

Suppose that for each $v \in \mathcal{A}_k$,

$$\|e^{\mathrm{ref}}_{k,v}\|_{P_k} \le \varepsilon^{\mathrm{ref}}_{k,v}.$$

Define

$$\rho^{\mathrm{ref}}_{k,v}(y) := \|\tilde{y} - \mu^{\mathrm{ref}}_{k,v}\|_{P_k}, \qquad b^{\mathrm{ref}}_{k,v}(y) := 2\varepsilon^{\mathrm{ref}}_{k,v}\rho^{\mathrm{ref}}_{k,v}(y) + (\varepsilon^{\mathrm{ref}}_{k,v})^2.$$

For a target value t, define

$$B^{\mathrm{ref}}_{k,\mathrm{split}}(y;t) := b^{\mathrm{ref}}_{k,t}(y) + \max_{v \neq t} b^{\mathrm{ref}}_{k,v}(y).$$

Then

$$\left|M_k(y;t) - M^{\mathrm{ref}}_{k}(\tilde{y};t)\right| \le B^{\mathrm{ref}}_{k,\mathrm{split}}(y;t).$$

Proof. By Lemma 1,

$$y - \eta_{k,v} = \tilde{y} - \mu^{\mathrm{ref}}_{k,v} + e^{\mathrm{ref}}_{k,v}.$$

Therefore

$$d_k(y;v) = \|\tilde{y} - \mu^{\mathrm{ref}}_{k,v} + e^{\mathrm{ref}}_{k,v}\|^2_{P_k},$$

while

$$d^{\mathrm{ref}}_{k}(\tilde{y};v) = \|\tilde{y} - \mu^{\mathrm{ref}}_{k,v}\|^2_{P_k}.$$

Define

$$\Delta_v := d_k(y;v) - d^{\mathrm{ref}}_{k}(\tilde{y};v).$$

Expanding the square gives

$$\Delta_v = 2\left\langle \tilde{y} - \mu^{\mathrm{ref}}_{k,v}, e^{\mathrm{ref}}_{k,v} \right\rangle_{P_k} + \|e^{\mathrm{ref}}_{k,v}\|^2_{P_k}.$$

By Cauchy–Schwarz and the value-specific leakage bound,

$$|\Delta_v| \le 2\varepsilon^{\mathrm{ref}}_{k,v}\|\tilde{y} - \mu^{\mathrm{ref}}_{k,v}\|_{P_k} + (\varepsilon^{\mathrm{ref}}_{k,v})^2 = b^{\mathrm{ref}}_{k,v}(y).$$

Therefore

$$|\Delta_t| \le b^{\mathrm{ref}}_{k,t}(y).$$

For the competitor term, using

$$\left|\min_{v \neq t} a_v - \min_{v \neq t} b_v\right| \le \max_{v \neq t} |a_v - b_v|,$$

we obtain

$$\left|\min_{v\neq t} d_k(y;v) - \min_{v\neq t} d_k^{\mathrm{ref}}(\tilde{y};v)\right| \leq \max_{v\neq t} b_{k,v}^{\mathrm{ref}}(y).$$

Adding target and competitor errors gives

$$\left|M_k(y;t) - M_k^{\mathrm{ref}}(\tilde{y};t)\right| \leq b_{k,t}^{\mathrm{ref}}(y) + \max_{v\neq t} b_{k,v}^{\mathrm{ref}}(y) = B_{k,\mathrm{split}}^{\mathrm{ref}}(y;t).$$

□

Corollary 1 (Buffered pooled-comparator decision equivalence)**.** *Under the assumptions of Proposition 6,*

$$\left|M_k^{\mathrm{ref}}(\tilde{y};t)\right| > B_{k,\mathrm{split}}^{\mathrm{ref}}(y;t) \quad \Longrightarrow \quad \operatorname{sign} M_k(y;t) = \operatorname{sign} M_k^{\mathrm{ref}}(\tilde{y};t).$$

The same conclusion holds with $M_k(y;t)$ and $M_k^{\mathrm{ref}}(\tilde{y};t)$ interchanged.

Proof. Proposition 6 gives

$$\left|M_k(y;t) - M_k^{\mathrm{ref}}(\tilde{y};t)\right| \leq B_{k,\mathrm{split}}^{\mathrm{ref}}(y;t).$$

Thus if either margin has magnitude larger than $B_{k,\mathrm{split}}^{\mathrm{ref}}(y;t)$, the other margin has the same sign. This proves the claim. □

B.6 Empirical decision agreement (Corollary 1)

The value-specific split certificate is

$$B_{k,\mathrm{split}}^{\mathrm{ref}}(y;t) = b_{k,t}(y) + \max_{v\neq t} b_{k,v}(y), \qquad b_{k,v}(y) = 2\varepsilon_{k,v}^{\mathrm{ref}}\|\tilde{y} - \mu_{k,v}^{\mathrm{ref}}\|_{P_k} + (\varepsilon_{k,v}^{\mathrm{ref}})^2.$$

When

$$|M_k^{\mathrm{ref}}(\tilde{y};t)| > B_{k,\mathrm{split}}^{\mathrm{ref}}(y;t),$$

Corollary 1 certifies that the pooled and reference-anchored margin signs agree.

We instantiate this diagnostic on held-out real samples from both datasets, where the reference-anchored comparator can be evaluated. For each attribute k, we estimate $\eta_{k,v}$, P_k, and $\mu_{k,v}^{\mathrm{ref}}$ using the same calibration protocol as the trust-score experiments, and evaluate on the canonical held-out evaluation splits.

For each sample, let $a_k^\star(y)$ denote the target value. We compare the deployed pooled margin

$$M_k(y;a_k^\star) = d_k(y;a_k^\star) - \min_{v\neq a_k^\star} d_k(y;v)$$

with the corresponding reference-anchored margin

$$M_k^{\mathrm{ref}}(\tilde{y};a_k^\star) = d_k^{\mathrm{ref}}(\tilde{y};a_k^\star) - \min_{v\neq a_k^\star} d_k^{\mathrm{ref}}(\tilde{y};v).$$

The certificate used in Table 4 is

$$B_{k,\mathrm{split}}^{\mathrm{ref}}(y;a_k^\star) = b_{k,a_k^\star}(y) + \max_{v\neq a_k^\star} b_{k,v}(y).$$

Certified agreement. Across every row with nonzero certified support, the certified agreement rate is 1.000. This is the empirical behavior predicted by Corollary 1: whenever the reference margin exceeds the split-radius perturbation band, the pooled and reference-anchored decisions have matching signs. The certificate is strongest for CelebA Male, where 86.2% of samples are certified, and for RxRx1 cell type, where 42.1% are certified.

Agreement beyond the certified regime. The certificate is sufficient, not necessary. Thus low certified coverage does not imply poor pooled/reference agreement. On CelebA, agreement remains very high for all attributes, ranging from 0.990 to 1.000, even when the certified fraction is small. RxRx1 cell type also has high agreement, 0.997, with substantial certified coverage. The main failure case is the finer-grained RxRx1 sirna decision: the median reference margin is only 1.2, while the median split bound is 16.6, so no samples are certified and overall agreement drops to 0.803.

Table 4: Empirical pooled/reference sign certification using the split bound $B^{\mathrm{ref}}_{k,\mathrm{split}}(y;t) = b_{k,t}(y) + \max_{v\neq t} b_{k,v}(y)$, $\quad b_{k,v}(y) = 2\varepsilon^{\mathrm{ref}}_{k,v}\|\tilde{y} - \mu^{\mathrm{ref}}_{k,v}\|_{P_k} + (\varepsilon^{\mathrm{ref}}_{k,v})^2$. med B_{split} is the median certificate radius, and Pr[certified] is the fraction of samples satisfying $|M^{\mathrm{ref}}_k(\tilde{y};t)| > B^{\mathrm{ref}}_{k,\mathrm{split}}(y;t)$. Agree(all) and Agree(cert.) report sign agreement before and after certification.

Dataset	Attribute	N_{eval}	V	$\varepsilon^{\mathrm{ref}}_k$	$\lvert M^{\mathrm{ref}}\rvert^{\mathrm{med}}$	med B_{split}	Pr[certified]	Agree(all)	Agree(cert.)
CelebA	Male	162770	2	0.79	74.2	53.0	**0.862**	**1.000**	**1.000**
CelebA	Smiling	162770	2	0.70	22.6	44.8	0.007	0.998	**1.000**
CelebA	Blond_Hair	162770	2	0.78	21.6	49.7	0.008	0.994	**1.000**
CelebA	Eyeglasses	162770	2	0.73	13.3	46.0	< 0.001	0.990	**1.000**
RxRx1	cell_type_id	758	4	0.79	14.0	15.2	0.421	0.997	**1.000**
RxRx1	sirna_id	640	20	1.04	1.2	16.6	0.000	0.803	—

Why certification is hard in low-margin regimes. The split bound is tighter than the earlier one-radius bound, but it can still be conservative when the reference margin is small relative to the leakage-weighted prototype radii. This is most visible for Smiling, Blond_Hair, and Eyeglasses, where empirical agreement is high but few samples have margins large enough to clear the certificate. In the RxRx1 sirna setting, the issue is more severe: many sirna prototypes are close in the reference geometry, so the deployed margin is often small compared with the perturbation band. This is exactly the regime where the theory predicts pooled scoring should be least reference-anchored comparator-like.

B.7 Empirical A1 context-balance diagnostic

A1 ($\mathbb{E}[c_k(a_{-k}) \mid a_k = v] = \bar{c}_k$) is the load-bearing condition that lets pooled scoring stand in for the reference-anchored comparator. By Lemma 1, under A1+A2 the difference $\eta_{k,v} - \mu^{\mathrm{ref}}_{k,v} = \gamma^{\mathrm{ref}}_k - e^{\mathrm{ref}}_{k,v}$ is v-independent up to the leakage $e^{\mathrm{ref}}_{k,v} = \delta_{k,v}(\bar{a}_{-k})$. We therefore use $\varepsilon^{\mathrm{ref}}_k = \max_v \|e^{\mathrm{ref}}_{k,v}\|_{P_k}$ as an operational A1-violation scale and compare it to the inter-prototype reference scale $\|\mu^{\mathrm{ref}}_{k,u} - \mu^{\mathrm{ref}}_{k,t}\|_{P_k}$, which estimates the value-signal gap $\|s_{k,u} - s_{k,t}\|_{P_k}$ in Lemma 1. A ratio $\ll 1$ means that any A1 violation is small relative to the attribute signal in the reference geometry; a ratio $\gtrsim 1$ indicates that leakage is large enough to obscure at least one value-pair comparison. Estimation reuses exactly the fitting and evaluation pools of Table 4; Table 5 reports the resulting per-attribute leakage-to-gap ratios.

Table 5: Empirical A1 diagnostic. "min gap" and "median gap" are the minimum and median pairwise P_k-Mahalanobis distances $\|\mu^{\mathrm{ref}}_{k,u} - \mu^{\mathrm{ref}}_{k,t}\|_{P_k}$ over $u \neq t$; Smaller ε/gap indicates greater compatibility with the A1 regime. Bold flags the regime where leakage exceeds the closest value-pair gap.

Dataset	Attribute	V	$\varepsilon^{\mathrm{ref}}_k$	min gap	median gap	ε/min	ε/median
CelebA	Male	2	0.79	8.81	8.81	0.090	0.090
CelebA	Smiling	2	0.70	4.99	4.99	0.141	0.141
CelebA	Blond_Hair	2	0.78	4.86	4.86	0.160	0.160
CelebA	Eyeglasses	2	0.73	3.84	3.84	0.190	0.190
RxRx1	cell_type_id	4	0.79	3.96	5.28	0.199	0.149
RxRx1	sirna_id	20	1.04	0.90	1.64	**1.157**	0.635

Small-leakage regimes. On all four CelebA attributes, the leakage-to-gap ratio is below 0.20. RxRx1 cell type is also in a small-leakage regime, with ε/median gap $= 0.149$ and ε/min gap $= 0.199$. These are precisely the settings where empirical pooled-vs-comparator sign agreement is near-perfect: agreement ranges from 0.990 to 1.000 on CelebA and is 0.997 on RxRx1 cell type (Table 4).

Large-leakage regime. The only regime where leakage exceeds the closest value-pair gap is RxRx1 sirna, with ε/min gap $= 1.157$. Its median leakage-to-gap ratio remains below one, ε/median gap $=$

0.635, but the closest sirna prototypes are sufficiently close that the perturbation band dominates many margins. This matches Table 4: RxRx1 sirna has median reference margin 1.2, median split certificate radius 16.6, no certified samples, and substantially lower pooled/reference agreement 0.803.

Takeaway. The A1 diagnostic aligns with the empirical certification results. Small leakage-to-gap ratios coincide with high pooled/reference agreement and, for CelebA Male and RxRx1 cell type, substantial certified coverage. The only ratio exceeding one occurs for the closest RxRx1 sirna pair, exactly where certification fails and empirical agreement is weakest. Thus A1 is not merely a formal assumption: its violation scale is measurable and predicts when pooled scoring is a reliable reference-comparator stand-in.

B.8 Exact translation equivalence

Corollary 2 (Exact translation equivalence)**.** *Suppose*

$$\eta_{k,v} = \mu^{\mathrm{ref}}_{k,v} + \gamma \qquad \textit{for all } v \in \mathcal{A}_k.$$

Then, for every y and every target t,

$$M_k(y;t) = M^{\mathrm{ref}}_k(y-\gamma;t).$$

Proof. For every v,

$$y - \eta_{k,v} = y - \gamma - \mu^{\mathrm{ref}}_{k,v}.$$

Therefore

$$d_k(y;v) = d^{\mathrm{ref}}_k(y-\gamma;v)$$

for every v. Taking the target distance minus the minimum competitor distance gives the result. □

B.9 Full-feature Mahalanobis discriminants

Proposition 7 (Pairwise decisions depend only on discriminant directions)**.** *Fix an attribute k, a positive definite matrix P_k, and prototypes*

$$\{m_{k,v} : v \in \mathcal{A}_k\} \subset \mathbb{R}^d.$$

Define

$$d^{(m)}_k(y;v) := (y - m_{k,v})^\top P_k (y - m_{k,v}).$$

For any two values $t, u \in \mathcal{A}_k$,

$$d^{(m)}_k(y;t) - d^{(m)}_k(y;u) = 2(m_{k,u} - m_{k,t})^\top P_k y + m_{k,t}^\top P_k m_{k,t} - m_{k,u}^\top P_k m_{k,u}.$$

Thus the sign of the pairwise preference between t and u depends on y only through the scalar projection

$$(P_k(m_{k,u} - m_{k,t}))^\top y.$$

Equivalently, only the component of y along the span of the pairwise discriminant directions

$$\operatorname{span}\{P_k(m_{k,u} - m_{k,t}) : u \neq t\}$$

can affect the target-vs-competitor comparison.

Proof of Proposition 7. Fix an attribute k, prototypes

$$\{m_{k,v} : v \in \mathcal{A}_k\},$$

and a positive definite matrix P_k. For any two values $t, u \in \mathcal{A}_k$, expand the two Mahalanobis distances:

$$d^{(m)}_k(y;t) = (y - m_{k,t})^\top P_k (y - m_{k,t}) = y^\top P_k y - 2m_{k,t}^\top P_k y + m_{k,t}^\top P_k m_{k,t},$$

and

$$d^{(m)}_k(y;u) = (y - m_{k,u})^\top P_k (y - m_{k,u}) = y^\top P_k y - 2m_{k,u}^\top P_k y + m_{k,u}^\top P_k m_{k,u}.$$

Subtracting cancels the common quadratic term $y^\top P_k y$, giving

$$d_k^{(m)}(y;t) - d_k^{(m)}(y;u) = 2(m_{k,u} - m_{k,t})^\top P_k y + m_{k,t}^\top P_k m_{k,t} - m_{k,u}^\top P_k m_{k,u}.$$

Thus the pairwise decision between t and u depends on y only through

$$(P_k(m_{k,u} - m_{k,t}))^\top y,$$

using symmetry of P_k. Therefore components of y orthogonal to the span of the discriminant directions

$$\operatorname{span}\{P_k(m_{k,u} - m_{k,t}) : u, t \in \mathcal{A}_k,\ u \neq t\}$$

cannot affect any pairwise Mahalanobis decision for attribute k. This proves the proposition. □

B.10 Robustness details

This subsection records an additional sufficient condition under which residual context variation does not change the attribute-wise Mahalanobis decision. It is not needed for the pooled-reference identification result above, but helps explain why full-feature scoring can remain stable when residual variation is small in the relevant discriminant directions.

Proposition 8 (Robustness to residual variation in discriminant directions)**.** *Fix an attribute k, a target value $t \in \mathcal{A}_k$, and suppose the ideal pooled prototypes satisfy*

$$\eta_{k,v} = \bar{c}_k + s_{k,v}.$$

Let a generated sample targeted at value t have normalized feature

$$y = \bar{c}_k + s_{k,t} + \lambda,$$

where $\lambda \in \mathbb{R}^d$ collects residual context dependence, imperfect disentanglement, and noise. Then for every competitor $u \neq t$,

$$d_k(y;t) - d_k(y;u) = 2(s_{k,u} - s_{k,t})^\top P_k \lambda - \|s_{k,u} - s_{k,t}\|_{P_k}^2,$$

where $\|z\|_{P_k}^2 := z^\top P_k z$. Consequently, if

$$2 \max_{u \neq t} \left|(s_{k,u} - s_{k,t})^\top P_k \lambda\right| < \min_{u \neq t} \|s_{k,u} - s_{k,t}\|_{P_k}^2,$$

then

$$M_k(y;t) < 0.$$

Proof. Since $\eta_{k,t} = \bar{c}_k + s_{k,t}$, we have

$$y - \eta_{k,t} = \bar{c}_k + s_{k,t} + \lambda - (\bar{c}_k + s_{k,t}) = \lambda,$$

and therefore

$$d_k(y;t) = \lambda^\top P_k \lambda.$$

For a competitor $u \neq t$,

$$y - \eta_{k,u} = \bar{c}_k + s_{k,t} + \lambda - (\bar{c}_k + s_{k,u}) = s_{k,t} + \lambda - s_{k,u}.$$

Hence

$$\begin{aligned} d_k(y;u) &= (s_{k,t} + \lambda - s_{k,u})^\top P_k (s_{k,t} + \lambda - s_{k,u}) \\ &= \lambda^\top P_k \lambda - 2(s_{k,u} - s_{k,t})^\top P_k \lambda + \|s_{k,u} - s_{k,t}\|_{P_k}^2. \end{aligned}$$

Subtracting gives

$$d_k(y;t) - d_k(y;u) = 2(s_{k,u} - s_{k,t})^\top P_k \lambda - \|s_{k,u} - s_{k,t}\|_{P_k}^2.$$

If

$$2\left|(s_{k,u} - s_{k,t})^\top P_k \lambda\right| < \|s_{k,u} - s_{k,t}\|_{P_k}^2$$

for every $u \neq t$, then each pairwise difference $d_k(y;t) - d_k(y;u)$ is strictly negative. Therefore the target value t beats every competitor under the Mahalanobis comparison, so

$$M_k(y;t) = d_k(y;t) - \min_{u \neq t} d_k(y;u) < 0.$$

□

C Estimation details

This appendix records the finite-sample estimators used to instantiate the population-level objects from the main text.

C.1 Calibration details for score normalization constants

Let $\mathcal{D}_{\mathrm{real}}$ denote the real-data split used to fit and calibrate the score model, and let $\mathcal{D}_{\mathrm{cal}} \subseteq \mathcal{D}_{\mathrm{real}}$ denote the subset used for calibration. In practice, $\mathcal{D}_{\mathrm{cal}}$ may be a held-out calibration split or the same real-data split used for fitting when data is limited.

For realism, after fitting $(\mu_{\mathrm{real}}, \Sigma_{\mathrm{real}})$, we compute

$$\widehat{E}_{\mathrm{real}}(y_i) := (y_i - \widehat{\mu}_{\mathrm{real}})^\top \widehat{\Sigma}_{\mathrm{real}}^{-1}(y_i - \widehat{\mu}_{\mathrm{real}}) \qquad \text{for } y_i \in \mathcal{D}_{\mathrm{cal}},$$

and estimate

$$\hat{m}_R := \frac{1}{|\mathcal{D}_{\mathrm{cal}}|} \sum_{y_i \in \mathcal{D}_{\mathrm{cal}}} \widehat{E}_{\mathrm{real}}(y_i), \qquad \hat{s}_R := \sqrt{\frac{1}{|\mathcal{D}_{\mathrm{cal}}| - 1} \sum_{y_i \in \mathcal{D}_{\mathrm{cal}}} \big(\widehat{E}_{\mathrm{real}}(y_i) - \hat{m}_R\big)^2}.$$

The calibrated realism score is then

$$\widehat{R}(y) := \frac{\widehat{E}_{\mathrm{real}}(y) - \hat{m}_R}{\hat{s}_R + \varepsilon_R},$$

where $\varepsilon_R > 0$ is a small numerical stabilizer.

For faithfulness, fix an attribute k and target value $t \in \mathcal{A}_k$. For each calibration sample y_i satisfying $a_{i,k} = t$, compute the fitted margin

$$\widehat{M}_k(y_i; t) := \widehat{d}_k(y_i; t) - \min_{v \neq t} \widehat{d}_k(y_i; v).$$

Then estimate

$$\hat{m}_{k,t} := \frac{1}{|\mathcal{D}_{k,t}^{\mathrm{cal}}|} \sum_{y_i \in \mathcal{D}_{k,t}^{\mathrm{cal}}} \widehat{M}_k(y_i; t),$$

and

$$\hat{s}_{k,t} := \sqrt{\frac{1}{|\mathcal{D}_{k,t}^{\mathrm{cal}}| - 1} \sum_{y_i \in \mathcal{D}_{k,t}^{\mathrm{cal}}} \big(\widehat{M}_k(y_i; t) - \hat{m}_{k,t}\big)^2},$$

where

$$\mathcal{D}_{k,t}^{\mathrm{cal}} := \{y_i \in \mathcal{D}_{\mathrm{cal}} : a_{i,k} = t\}.$$

The calibrated attribute score becomes

$$\widehat{F}_k(y; t) := \frac{\widehat{M}_k(y; t) - \hat{m}_{k,t}}{\hat{s}_{k,t} + \varepsilon_F},$$

with $\varepsilon_F > 0$ a small numerical stabilizer.

In practice, if a class-conditional calibration set is very small, one may use a minimum-count threshold before reporting calibrated scores, pool variance estimates across values of the same attribute, or clip the denominator away from zero. These are finite-sample stabilization choices and do not affect the population-level theory above.

C.2 Shared-covariance estimation and shrinkage

For each attribute k, let

$$\widehat{\eta}_{k,v} = \frac{1}{n_{k,v}} \sum_{i : a_{i,k} = v} y_i$$

denote the empirical prototype for value $v \in \mathcal{A}_k$, where $n_{k,v}$ is the number of real samples with $a_{i,k} = v$.

The pooled within-value covariance is

$$\widehat{\Sigma}_k^{\text{within}} := \frac{1}{N_k - |\mathcal{A}_k|} \sum_{v \in \mathcal{A}_k} \sum_{i: a_{i,k} = v} (y_i - \widehat{\eta}_{k,v})(y_i - \widehat{\eta}_{k,v})^\top,$$

where $N_k = \sum_{v \in \mathcal{A}_k} n_{k,v}$.

To improve conditioning in high dimension, we use shrinkage toward an isotropic target:

$$\widehat{\Sigma}_k^{\text{shr}} := (1 - \alpha_k)\widehat{\Sigma}_k^{\text{within}} + \alpha_k \tau_k I, \qquad \tau_k := \frac{1}{d} \operatorname{tr}\left(\widehat{\Sigma}_k^{\text{within}}\right),$$

with shrinkage coefficient $\alpha_k \in [0, 1]$. The corresponding precision estimate is

$$\widehat{P}_k := \left(\widehat{\Sigma}_k^{\text{shr}} + \varepsilon_P I\right)^{-1},$$

where $\varepsilon_P > 0$ is a small numerical ridge added for inversion stability.

The resulting fitted quadratic distance is

$$\widehat{d}_k(y; v) := (y - \widehat{\eta}_{k,v})^\top \widehat{P}_k (y - \widehat{\eta}_{k,v}).$$

The same construction can be used for the global realism model by replacing the class-conditional pooled covariance with the empirical covariance of all real normalized features.

C.3 Finite-sample considerations

Several finite-sample issues are worth noting.

First, reference coverage is a population support condition. In finite data, the corresponding anchored slices may exist but contain few examples. This is one reason we use pooled prototypes in the practical method even when the theory is stated relative to a reference-anchored comparator.

Second, class imbalance affects both prototype estimation and calibration. For heavily imbalanced attributes, the empirical prototypes $\widehat{\eta}_{k,v}$ and calibration moments $(\hat{m}_{k,t}, \hat{s}_{k,t})$ may have different variances across values. This motivates reporting class counts and, if needed, applying minimum-support thresholds or pooled-variance stabilizers.

Finally, the shrinkage level α_k trades bias for conditioning. Stronger shrinkage produces more stable inverses in high dimension and small sample regimes, while weaker shrinkage preserves finer geometry when sufficient data are available.

D Broader impacts

Positive impacts. The proposed trust score is intended to make conditional generative models more deployable in scientific settings where target real samples are unavailable, particularly biological imaging where wet-lab validation of *in silico* predictions is expensive. By providing a per-sample, per-condition reliability estimate, the score can prevent unreliable synthetic generations from being treated as evidence in downstream analyses, reducing the risk of misleading scientific conclusions and saving experimental resources by prioritizing high-trust predictions for validation.

Potential negative impacts. Any post-hoc filtering rule can be misused to cherry-pick generations that look plausible without being faithful to the requested condition — our score mitigates but does not eliminate this risk, and we encourage users to report acceptance rates and per-condition trust calibration alongside any selected sample sets.

Dual-use considerations. The score itself is a diagnostic and does not improve the generative capabilities of the underlying model; it does not lower the barrier to producing higher-quality synthetic faces or biological images relative to running the underlying generator alone.

E Implementation details

Architecture. SiT-B/2 [19] backbone (12 layers, 768 hidden dim, 12 heads), patch size 2 over a 32×32 VAE latent (256 patches), stable-diffusion SD-VAE-MSE encoder/decoder [30] (frozen). Conditioning is per-attribute embedding summed with the timestep embedding. The REPA projector is a 3-layer MLP $768 \to 2048 \to 2048 \to d_h$ inserted at encoder depth $\ell = 8$, where $d_h \in \{1024, 1152, 384\}$ for DINOv3 / SigLIP / OpenPhenom respectively.

Training. AdamW with $\mathrm{lr} = 10^{-4}$, weight decay 0, 1000 linear warmup steps, gradient clip 1.0, $400{,}000$ optimizer steps, global batch size 128 (16 per GPU $\times$ 8 GPUs, DDP, `bf16` mixed precision), EMA on the generator with decay 0.9999. Class-conditional dropout $p = 0.1$ for classifier-free-guidance training; for marginal/held-out runs we additionally use the compositional regularizer with weight 1.0 applied to 25% of batches at $t > 0.7$.

Data and preprocessing. `CelebA`: 256×256 centre-cropped RGB, random horizontal flip, $[-1, 1]$ normalization; conditioning on the 4 binary attributes MALE / SMILING / BLOND_HAIR / EYEGLASSES. `RxRx1`: 512×512 6-channel raw fluorescence resized to 256×256 via numpy bilinear, no flips, no per-channel normalization (raw uint16 scaled to $[0, 1]$); conditioning on (cell_type_id, sirna_id).

Sampling. 250-step DDIM-style integration of the SiT flow with $t_{\mathrm{cutoff}} = 0.04$; classifier-free-guidance scale 1.0. We draw 1000 samples per condition for `CelebA` (16 conditions) and 100 per condition for `RxRx1` (50 conditions in the canonical subset).

Trust-score calibration. Encoder features are taken as the mean over patch tokens (DINOv3 / SigLIP) or as the model's pooled output (OpenPhenom). Covariance estimates use Ledoit–Wolf shrinkage [14] with $\varepsilon_P = 10^{-5}$ ridge added before inversion. Calibration constants $(\hat{m}_R, \hat{s}_R, \hat{m}_{k,t}, \hat{s}_{k,t})$ are estimated on the same real training-support split used for fitting; full formulas are in Section C.

F During-generation translator and compute details

The translator is trained to preserve the aspects of the feature extractor representation used by the Mahalanobis trust score. A standard cosine-alignment loss is sufficient to align dominant directions, but it need not preserve the residual geometry that controls whitened distances and prototype comparisons. This can distort the relative scale of directions that are weak in cosine similarity but important after covariance normalization. We therefore train $g_{\phi,\ell}$ with a whitened-geometry objective: a tempered feature-matching term in feature-whitened coordinates, together with mean and covariance matching penalties in normalized feature space. The whitening transform is computed from the real training features, using the same covariance geometry as the trust score. This makes the mapped features compatible with the post-generation scoring rule, so that any difference between post-generation and during-generation evaluation reflects feature acquisition rather than a change in metric.

Specifically, we train the translator from an internal model feature $z_{\ell,\tau}(x)$ to the Φ feature space. Let

$$p_\phi(x) = g_{\phi,\ell}(z_{\ell,\tau}(x)), \qquad \bar{p}_\phi(x) = \mathrm{norm}(p_\phi(x)), \qquad \bar{y}(x) = \mathrm{norm}(\Phi(x)).$$

Let μ_t and Σ_t denote the empirical mean and covariance of normalized pretrained features on real training data. We use

$$\mathcal{L}_{\mathrm{whitgeom}} = \lambda_{\mathrm{whit}} \mathbb{E}_x \left[\left\| W \left(\bar{p}_\phi(x) - \bar{y}(x) \right) \right\|_2^2 \right] + \lambda_{\mathrm{geom}} \left(\frac{\|\mu_p - \mu_t\|_2^2}{\|\mu_t\|_2^2 + \epsilon} + \frac{\|\Sigma_p - \Sigma_t\|_F^2}{\|\Sigma_t\|_F^2 + \epsilon} \right),$$

where μ_p and Σ_p are the batch mean and covariance of mapped normalized features. The matrix W is a tempered whitening transform computed from the feature covariance, with exponent $\gamma = 0.75$ and condition-number cap $\kappa = 1000$. In all experiments we use

$$\lambda_{\mathrm{whit}} = 1.0, \qquad \lambda_{\mathrm{geom}} = 0.1.$$

F.1 Compute save from early abstention

Let $C_{\mathrm{den}}(\ell, \tau)$ denote the cost of running the generator up to layer-step pair (ℓ, τ), C_{dec} the decoder cost, C_{enc} the Φ cost, and C_{g_ϕ} the translator cost. Post-generation scoring costs

$$C_{\mathrm{post}} = C_{\mathrm{den}}(\mathrm{full}) + C_{\mathrm{dec}} + C_{\mathrm{enc}} + C_{\mathrm{score}},$$

whereas during-generation scoring costs

$$C^{\mathrm{map}}_{\ell,\tau} = C_{\mathrm{den}}(\ell, \tau) + C_{g_\phi} + C_{\mathrm{score}}.$$

Thus early abstention avoids the remaining denoising computation, the VAE decoder, and the feature extractor.

F.2 Translator–post-hoc ranking agreement

We examine whether the during-generation translator preserves the ranking induced by the post-generation trust score. We evaluate the same CelebA generations using three scoring representations: translated features in SigLIP space, post-hoc SigLIP features and post-hoc DINOv3 features.

Table 6: Spearman correlations between translated and post-hoc trust scores on CelebA generations.

Score	Translator	SigLIP	DINOv3
Translator	1.000	0.674	0.504
Post-hoc SigLIP	0.674	1.000	0.765
Post-hoc DINOv3	0.504	0.765	1.000

The translated trust score correlates more strongly with post-hoc SigLIP, its target feature space, than with post-hoc DINOv3. This suggests that the translator preserves part of the intended post-hoc ranking rather than learning an entirely unrelated ordering.

F.3 Hardware, runtime, and total compute

All experiments were run on a single internal cluster node with 8×NVIDIA A100 80GB GPUs (DDP, NCCL backend, mixed-precision `bf16`). Generator training uses 400,000 optimizer steps with `lr=1e-4`, `warmup_steps=1000`, `gradient_clip_val=1.0`, and global batch size 128 (16 per GPU $\times$ 8 GPUs).

Per-experiment cost. Generations are performed with the 250-step DDIM-style trajectory: 1000 samples per condition for CelebA (16 conditions $\Rightarrow$ 16k samples) and 100 per condition for RxRx1 (50 conditions $\Rightarrow$ 5k samples). Per-step generator FLOPs are $\approx$ 46 GFLOPs (SiT-B/2, 256 patches), so a full 250-step trajectory is $\approx$ 11.5 TFLOPs per sample; the REPA projector adds $\approx$ 4 GFLOPs (<0.04% overhead).

Trust-evaluation cost. Once features are cached, all reported trust evaluations run on a single GPU in under an hour per (model, scoring space, regime).

G OpenPhenom feature diagnostic

While OpenPhenom-aligned REPA models achieve the best generation quality on RxRx1 as measured by post-hoc DINOv3 KID (Table 18), trust scoring in the OpenPhenom-aligned feature space fails ($\rho \approx 0.05$). We conducted a systematic diagnostic to understand this failure.

Root cause: collapsed within-class covariance. OpenPhenom features have an effective within-class dimensionality of $\approx$4 in 2304 dimensions, compared to $\approx$16 for DINOv3. The top principal component captures 61.5% of within-class variance, and this dominant direction is orthogonal to both cell type and perturbation identity. The precision matrix therefore has condition number $\approx$21M, amplifying $\approx$2300 near-zero variance directions.

Table 7: Proposition 8 robustness margin test across encoders. Violation rate = fraction of samples where noise exceeds signal in the Mahalanobis margin. N/S = noise-to-signal ratio.

	cell_type attribute			siRNA attribute		
Encoder	Violation	N/S	Slack	Violation	N/S	Slack
DINOv3	0.0%	0.10	56.4	69.5%	1.21	−4.7
SigLIP	0.0%	0.07	63.1	70.5%	1.26	−5.3
OpenPhenom	0.7%	0.18	32.5	77.8%	1.48	−6.3

Table 8: Effect of OpenPhenom-style preprocessing on DINOv3 features. Cross-space = OP trust vs. DINOv3 KID; same-space = DINOv3 trust vs. DINOv3 KID. IN = InstanceNorm.

DINOv3 preprocessing variant	ρ(cross-space)	ρ(same-space)
Standard (to_rgb)	−0.54	0.93
+ InstanceNorm	−0.60	0.80
Dual 3ch (no IN)	−0.65	0.77
Dual 3ch + IN	+0.16	0.43
Per-channel (no IN)	−0.40	0.43
Per-channel + IN	+0.12	0.56

Mahalanobis scoring destroys signal. The precision-weighted signal-to-noise ratio (Mahalanobis SNR) is 0.11 for OpenPhenom vs. 0.24 for DINOv3. The Mahalanobis/Euclidean SNR ratio is 0.73 for OpenPhenom (signal-destructive) vs. 1.55 for DINOv3 (signal-enhancing). This places OpenPhenom features in the weak-robustness regime of Proposition 8: the noise term $2\max_{u\neq t}|(\eta_{k,u}-\eta_{k,t})^\top P_k\lambda|$ exceeds the signal term $\min_{u\neq t}\|\eta_{k,u}-\eta_{k,t}\|^2_{P_k}$ for 77.8% of samples on the siRNA attribute (Table 7).

Preprocessing drives the geometry. An isolation experiment applying OpenPhenom-style preprocessing (InstanceNorm + per-channel encoding) to DINOv3 eliminates the cross-space anti-correlation (ρ: $-0.54 \to +0.12$) but simultaneously destroys DINOv3's same-space scoring power (ρ: $0.93 \to 0.43$); see Table 8. This confirms that the issue is a fundamental consequence of OpenPhenom's channel-agnostic architecture: InstanceNorm removes per-image intensity variation, and per-channel processing prevents cross-channel interaction, concentrating variance into non-semantic directions.

Contrastive models are robust. A control experiment with post-hoc SigLIP features on the same RxRx1 images achieves $\rho = 0.80$ cross-space (SigLIP trust vs. DINOv3 KID), despite SigLIP having even higher pairwise cosine similarity (0.979 vs. 0.95 for OpenPhenom). The problem is isolated to OpenPhenom's MAE feature geometry, not to high feature similarity per se (Table 9).

Feature geometry comparison. Table 10 summarizes the within-class covariance geometry that underlies the scoring failure.

Practical implication. OpenPhenom features are discriminative for classification ($\eta^2 \approx 0.41$ for cell type, comparable to DINOv3's 0.36) but violate the geometric assumptions required for Mahalanobis-based distributional scoring. This highlights a key distinction: classification requires only hyperplane separability in a few discriminative dimensions, while trust scoring requires meaningful geometry across the full feature space. For trust evaluation on RxRx1, we recommend using post-hoc contrastive features (DINOv3 or SigLIP) rather than MAE-based domain-specific encoders.

H RxRx1 50-condition evaluation subset

All RxRx1 results in the main text are reported on a balanced 50-condition subset (25 seen + 25 unseen, covering all 4 cell types) rather than the full 4552-condition space. This appendix documents why the subset is constructed the way it is, and why a naive "top-n by sample count, stratified by cell type" construction fails.

Table 9: Cross-encoder comparison on RxRx1 (50 conditions, same generated images from REPA-OpenPhenom model). Same-space = trust and KID in the same encoder; cross-space = encoder trust vs. DINOv3 KID.

	Same-space		Cross-space (vs. DINOv3 KID)	
Encoder	ρ(trust)	ρ(real.)	ρ(trust)	ρ(real.)
DINOv3	0.93	0.94	—	—
SigLIP	0.79	0.65	0.80	0.84
OpenPhenom	0.45	0.40	−0.54	−0.48

Table 10: Within-class covariance geometry across encoders (RxRx1 real features).

Metric	DINOv3	SigLIP	OpenPhenom
Feature dimension	1024	1152	2304
Effective dim (within-class)	16.0	12.2	4.0
Top-1 PC fraction	0.19	0.23	0.62
Precision condition number	5.1M	1.5M	21.1M
Euclidean SNR	0.15	0.11	0.15
Mahalanobis SNR	0.24	0.28	0.11
Mahal/Euclid ratio	1.55	2.52	0.73

Constraints. A usable subset must satisfy three requirements simultaneously. (i) *Sample sufficiency*: each retained condition must have enough real samples for a per-condition KID bootstrap to be stable. (ii) *Cell-type coverage*: all four cell types must appear in the subset, so that trust-vs-ΔKID correlations are not driven by a single biological context. (iii) *Discriminability*: each condition must be distinguishable from the rest under a class-balanced linear probe on real feature space, otherwise trust scoring has nothing to rank, and apparent correlations are dominated by class-imbalance artifacts rather than genuine per-condition structure. The held-out training regime (Section 5) adds a fourth constraint: the test set should include enough unseen conditionings for it to be stress-testing, the 100 most-frequent cell type $\times$ perturbation pairs removed from training.

Failure mode of the naive construction. A first-pass subset was built by selecting the top-n conditions by real-sample count, stratified across cell types (7/6/6/6 seen across ct=0/1/2/3 and 9/8/8 unseen across ct=0/1/2). A cross-validated class-balanced linear probe on real DINOv3 features, run on this subset, exposed three problems.

1. *Filler rows.* 13 of the 25 seen conditions (all ct=0 and ct=2 non-controls) had only n=14 real samples each. This is below any stable per-condition KID bootstrap. Under a class-balanced probe on the 50-class subproblem, these rows collapse to top-1 $\approx$0—the probe cannot distinguish them from each other or from the rest of the catalog. They contribute no usable ranking signal.
2. *No unseen functional diversity.* The unseen part of the test set contained only control wells, even though several non-control held-out perturbations were available in held-out set of the diffusion training. The support-shift test therefore covered a single class of perturbations rather than the diversity the held-out set was designed to provide.
3. *Imbalance-probe confound.* A class-imbalanced probe resulted in a class-frequency artifact: controls dominate the per-class sample counts. Under a class-balanced probe restricted to the 50-class subproblem, the controls collapse and the genuinely discriminable rows are the non-control siRNAs together with a smaller subset of controls. Any subset that does not prioritize these rows is measuring imbalance, not trust

Corrected construction. The subset used throughout the paper is built to satisfy all four constraints by explicitly targeting discriminable rows under a class-balanced probe, rather than largest-n rows.

- *Seen arm (25 conditions, $n \geq 23$, median n=32).* 20 ct=1 non-control siRNAs selected by real-sample count (all at n=32), plus 5 ct=3 controls at $n \approx$ 22–24. The ct=1 non-controls carry the per-condition signal identified by the class-balanced probe; the ct=3 rows satisfy the cell-type-coverage constraint (see below).

- *Unseen arm (25 conditions, $n \geq 27$, median $n \approx 52$).* 8 ct=0 controls (largest by n), 7 ct=1 non-control held-out siRNAs (349, 350, 351, 352, 363, 364, 401), 2 ct=1 controls (largest by n), and 8 ct=2 controls (largest by n). This is the first configuration in which the unseen arm contains non-control perturbations, giving the support-shift evaluation functional rather than purely positional variety.

Cell-type coverage notes. Two cell-type asymmetries are worth making explicit.

- *ct=0 and ct=2 appear only in the unseen arm.* Their non-control conditions cap at n=14, which fails the sample-sufficiency constraint, and all of their high-n controls are listed in `RXRX1_HELDOUT_PAIRS`, so they cannot populate the seen arm. Subset-level cell-type coverage (each cell type appears somewhere in the 50 conditions) is therefore retained, even though arm-level coverage is not—the alternative would require reinstating the 13 filler rows.
- *ct=3 is always retained.* ct=3 has no held-out pairs at all (so it cannot populate the unseen arm) and its non-control wells max out at n=6 (unusable). The only ct=3 data with sufficient sample counts are the 31 control wells (`sirna_id` 1108–1138, max n=24). The seen arm therefore includes the 5 largest ct=3 controls unconditionally, as the only way to honor the all-cell-types constraint.

Relationship to the naive subset. 29 of the 50 final conditions are also present in the naive construction (18 in the unseen arm, 11 in the seen arm), and no condition migrates between arms. The corrected subset is therefore a targeted sharpening of the original—dropping filler rows and adding discriminable rows—rather than a rebuild, so results under the two constructions remain directly comparable on their shared conditions.

I CellProfiler validation details on RxRx1

This appendix details the CellProfiler (CP) feature pipeline used for the morphology-space validation in Section 5.1.3 and reports the per-feature correlation breakdown and uninformative-feature sanity check.

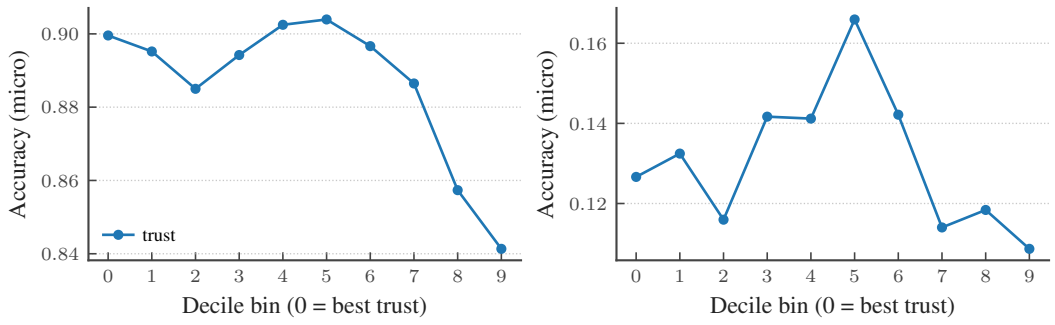
Decile-binning headline numbers. Table 11 summarizes the curves of Figure 6 as the bin-0 vs. bin-9 micro-accuracy (the headline contrast referenced in the main text), per (model, target).

Table 11: CP-space decile downstream classification on RxRx1 (Figure 6 headline numbers; 50-condition subset, marginal models, SigLIP scoring). Micro-accuracy of bin-0 vs. bin-9 classifiers trained on per-decile gen samples and tested on the canonical-50 real pool. "Trust spread" = bin 0 − bin 9; positive means higher-trust deciles produce more useful training data. Random-baseline micro-accuracy is flat (std ≤ 0.005) and matches the bin-0/bin-9 mean.

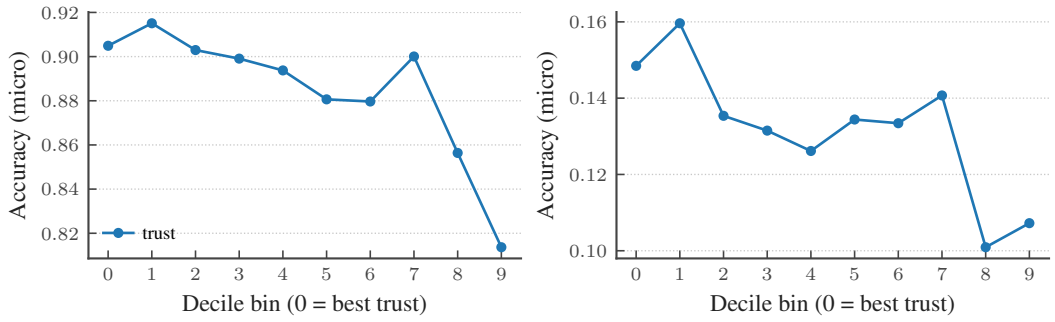
Model	Target	bin 0	bin 9	Trust spread	Random mean
Vanilla	celltype	0.900	0.841	+0.058	0.893
Vanilla	combo	0.127	0.109	+0.018	0.130
REPA (DINOv3)	celltype	0.905	0.814	+0.091	0.887
REPA (DINOv3)	combo	0.148	0.107	+0.041	0.132
REPA (SigLIP)	celltype	0.897	0.803	+0.094	0.887
REPA (SigLIP)	combo	0.146	0.117	+0.029	0.131

Per-sample centroid distance (full table). Table 12 reproduces Table 2 from the main text with the trust-accepted sample count n_{trust} and the bootstrap 95% CI on Δd (2000 reps, stratified). Negative Δd with a CI bounded above by zero is read as "trust-selected samples are morphologically closer to real"; the vanilla-unseen row is the only cell whose CI grazes zero.

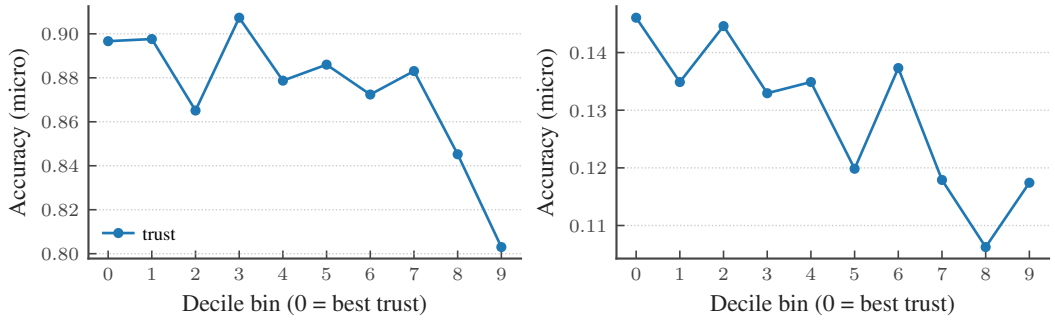
Feature pipeline. Starting from the raw `Image.csv` produced by CellProfiler over the canonical-50 real subset, we apply the following deterministic pipeline. (i) Drop metadata columns (regex matches against `metadata`, `filename`, `pathname`, `imagenumber`, `objectnumber`, `executiontime`, `moduleerror`, `series`, `frame`, `group_*`, `url`), leaving 2467 numeric features. (ii) Drop NaN



(a) Vanilla marginal.



(b) REPA-DINOv3 marginal.



(c) REPA-SigLIP marginal.

Figure 6: **Full CP-space decile downstream classification on RxRx1 (kept-621 features, SigLIP trust scoring).** Each row is one marginal model; *left*: 4-way celltype accuracy by decile, *right*: 50-way combo accuracy. Solid blue: trust-ranked decile. The REPA-SigLIP row is highlighted in the main text.

rows and restrict to the canonical 50 cell×siRNA pairs (2061 real samples). (iii) Apply a variance threshold (var $< 10^{-5}$) and an outlier filter (drop any column with a real-data $|z| > 5$), giving the *kept-621* set used throughout the centroid and per-feature correlation analyses. (iv) Real-fit `StandardScaler` is reused across all readouts.

For the uninformative-feature sanity check we use a less aggressive variant — *unfiltered* — which keeps all 2415 columns that survive metadata removal and zero-variance pruning, so genuinely noise-only features can surface.

Per-feature correlation construction. For each (model, arm, feature f),

$$r_{\text{trust}}(f) = \text{Pearson}\big(\mu_{\text{trust}}(f),\, \mu_{\text{real}}(f)\big), \qquad r_{\text{rand}}(f) = \text{Pearson}\big(\mu_{\text{rand}}(f),\, \mu_{\text{real}}(f)\big),$$

where $\mu_\bullet(f)$ is the per-combo mean of f over the corresponding gen subset (trust-accepted, or combo-stratified random averaged over 5 seeds), and the Pearson correlation is taken across the

Table 12: Full version of Table 2: per-sample centroid distance on RxRx1 (kept-621 CP features) with trust-accepted counts and bootstrap CIs. Random pool is condition-matched (combo-stratified) and matched in size to n_{trust} per arm.

Model	Arm	n_{trust}	Accept%	$\bar{d}_{\text{trust}}$	$\bar{d}_{\text{baseline}}$	Δd	95% CI
Vanilla	seen	1657	66.3	18.63	18.93	-0.300	$[-0.570, -0.020]$
Vanilla	unseen	547	21.9	20.92	21.52	-0.601	$[-1.211, +0.028]$
REPA (DINOv3)	seen	1530	61.2	18.31	18.90	-0.586	$[-0.854, -0.314]$
REPA (DINOv3)	unseen	524	21.0	20.19	21.21	-1.018	$[-1.663, -0.394]$
REPA (SigLIP)	seen	1640	65.7	18.38	18.88	-0.496	$[-0.750, -0.241]$
REPA (SigLIP)	unseen	588	23.5	22.13	22.97	-0.844	$[-1.489, -0.167]$

Table 13: Top-5 features by r_{trust} on REPA-DINOv3, RxRx1 unseen arm (kept-621). $\Delta = r_{\text{trust}} - r_{\text{rand}}$.

Feature	r_{trust}	r_{rand}	Δ
`Median_Cells_Texture_Correlation_OrigGray_5_02_256`	$+0.937$	$+0.926$	$+0.011$
`Median_Cells_Texture_Correlation_OrigGray_10_02_256`	$+0.935$	$+0.836$	$+0.099$
`Median_Cytoplasm_Texture_Correlation_OrigGray_5_02_256`	$+0.934$	$+0.863$	$+0.071$
`Mean_Cells_Texture_Correlation_OrigGray_5_03_256`	$+0.934$	$+0.893$	$+0.041$
`Mean_Cells_Texture_Correlation_OrigGray_5_01_256`	$+0.932$	$+0.921$	$+0.011$

arm's combos. Higher r means the feature's per-combo profile in gen tracks real well; $\Delta(f) := r_{\text{trust}}(f) - r_{\text{rand}}(f) > 0$ means trust selection sharpens cross-combo correspondence on that feature.

Top-correlated features under trust (kept-621). Across all six (model, arm) cells the same families recur at the top of r_{trust} ($r \geq 0.90$): per-cell and per-cytoplasm texture autocorrelation at scales 3/5/10 (`Cells_Texture_Correlation_OrigGray_*`), image-level granularity (`Granularity_2/4_OrigGray`), total cell-segmentation area (`AreaOccupied_AreaOccupied_IdentifySecondaryObjects`), and cytoplasm shape regularity variance (`StDev_Cytoplasm_AreaShape_FormFactor`). Table 13 gives a representative example for the REPA-DINOv3 unseen arm.

Uninformative-feature sanity check (unfiltered). Repeating the analysis on the unfiltered 2415-feature set surfaces features for which even random gen has no per-combo correspondence with real. Sorting by $|r_{\text{rand}}|$ ascending isolates these. Table 14 shows the bot-10 for the vanilla unseen arm; the same pattern appears across all six (model, arm) cells.

Table 14: Bot-10 features by $|r_{\text{rand}}|$ on Vanilla, RxRx1 unseen arm (unfiltered 2415-feature set). Random gen has no per-combo correspondence with real on these features ($|r_{\text{rand}}| \approx 0$), consistent with their interpretation as uninformative under any selection.

Feature	r_{trust}	r_{rand}	Family
`Mean_IdentifyPrimaryObjects_Children_..._Count`	-0.000	-0.000	tautological count
`StDev_IdentifyPrimaryObjects_Children_..._Count`	-0.000	-0.000	tautological count
`StDev_Nuclei_Texture_Variance_OrigGray_5_01_256`	-0.086	$+0.001$	texture variability
`Median_Nuclei_RadialDistribution_RadialCV_..._4of4`	-0.047	$+0.001$	outermost-shell radial CV
`Median_Cells_Granularity_15_OrigGray`	-0.262	-0.001	very-coarse granularity
`StDev_Nuclei_AreaShape_Zernike_9_7`	-0.269	-0.002	high-order Zernike
`Median_Cytoplasm_Texture_InfoMeas1_OrigGray_10_01_256`	$+0.052$	-0.004	texture moment
`Median_Cytoplasm_AreaShape_FormFactor`	$+0.127$	-0.004	shape regularity
`Mean_Cells_AreaShape_Zernike_7_1`	$+0.282$	-0.004	high-order Zernike
`StDev_Cells_Neighbors_AngleBetweenNeighbors_5`	$+0.080$	$+0.005$	neighbour angle

The recurring uninformative families across all six (model, arm) cells are: tautological segmentation counts (every primary object has exactly one secondary child after segmentation, so the column is constant), very-coarse-scale granularity (dominated by well-level lighting), high-order Zernike moments (sample-noise-dominated rotational shape), outermost-ring radial CVs, and absolute object

orientation (intrinsically arbitrary). None of these should correlate with biology, and $r_{\text{rand}} \approx 0$ confirms they don't.

A side observation: a handful of features that are uninformative under random become *strongly* anti-correlated under trust ($r_{\text{trust}} \ll 0$ while $|r_{\text{rand}}| \approx 0$). The clearest example is `Mean_Cytoplasm_Location_MaxIntensity_Y_OrigGray` on REPA-SigLIP unseen ($r_{\text{rand}} = +0.005$, $r_{\text{trust}} = -0.475$): image-level position features that random gen treats as noise become systematically biased by trust selection. This is the same pose/position bias visible in the kept-621 anti-correlation list.

Table 15: Per-condition mean trust scores and ΔKID on CelebA under compositional shift. Labels indicate the four binary attributes (Male, Smiling, Blond Hair, Eyeglasses). Lower values indicate better quality.

Label (M·Sm·Bl·Ey)	Set	Mean Trust	ΔKID
0100	Seen	-0.37	0.190
0010	Seen	-0.22	0.221
0000	Seen	0.50	0.211
0110	Unseen	0.59	0.297
1000	Seen	0.92	0.302
0001	Seen	0.93	0.354
0011	Unseen	1.16	0.341
0101	Unseen	1.32	0.337
1100	Unseen	2.56	0.751
0111	Unseen	2.65	0.684
1010	Unseen	2.85	0.595
1001	Unseen	3.08	0.706
1110	Unseen	4.37	1.303
1101	Unseen	4.44	1.406
1011	Unseen	4.72	1.045
1111	Unseen	6.18	1.662

J Graded compositional shift on CelebA

A natural follow-up question to the post-generation ordering result is whether unseen conditionings should be treated as uniformly unreliable. CelebA held-out gives this question a sharp form: each queried condition has a natural notion of shift severity given by its minimum Hamming distance to the observed training support, and we can ask whether the per-condition mean trust score follows the same gradient.

Figure 7 and table 15 show the answer. The five seen conditions cluster at low ΔKID and low trust; unseen conditions spread outward as the Hamming distance from support grows. Crucially, this is not a binary seen vs.unseen effect: some unseen conditions at small Hamming distance achieve quality comparable to seen ones, and the score correctly assigns them correspondingly favorable trust. Conversely, larger-distance compositions are both harder to generate and assigned worse trust.

This behavior matters for the intended use case. Under compositional generalization, the relevant question is not simply whether a request is outside training support, but how severe that shift is and whether the resulting sample should still be trusted. The score captures this graded structure directly, rather than collapsing all unseen combinations into a single failure category.

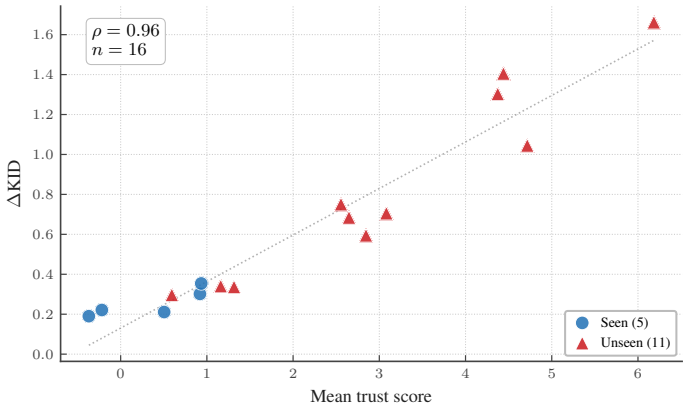


Figure 7: **CelebA stress-test (REPA DINOv3, DINOv3 scoring): per-condition trust vs. ΔKID.** Each point is one of 16 conditions. The 5 seen conditions cluster at low ΔKID; unseen conditions with more active attributes drift further from training support and exhibit higher ΔKID. Trust scores track this degradation ($\rho = 0.96$).

K Scorer-design ablation

Table 16 compares our calibrated Mahalanobis trust score (Section 3) against three alternative scorers on the same DINOv3 mean-patch feature space, under the identical P95-real-threshold selection protocol used for Table 1. *Linear probe* fits one logistic regression per attribute on real DINOv3 features and scores a generated sample by the summed free energy $-\sum_a \mathrm{logsumexp}(z_a(x))$. *kNN (per-attr.)* scores by the sum, over attributes, of cosine distance from the sample to its 5-th nearest neighbour inside the target-class real subset. *CLIP* (CelebA-only) scores by $-\cos(\text{image}, \text{prompt})$ over a fixed template across the 16 attribute combos.

The picture is consistent with the takeaway in Section 5.1.1. All three alternative scorers produce near-degenerate P95-real thresholds, the generated trust scores are almost entirely below the 95 th-percentile real threshold, yielding $\approx 100\%$ acceptance. So they are unusable as abstention rules even when their ranking happens to be informative. kNN recovers a Mahalanobis-comparable condition-level correlation but cannot translate it into a calibrated filter; the linear probe and CLIP alignment collapse on both axes. Our Mahalanobis score is the only construction here that yields both non-degenerate acceptance and substantial KID improvement, which is what the deployment-time abstention story requires. The Mahalanobis rows in Table 16 are reproduced from Tables 1 and 17; held-out rows match the main text and full-support rows are kept here as sanity checks.

Table 16: Scorer ablation: P95-real-threshold sample selection across four alternative trust scorers on the same DINOv3 meanpatch feature space. Threshold, KID computation (k=500 shuffled subsampling), and random baseline are identical to Table 1. *Mahalanobis* (ours) is the global-energy + per-attribute margin scorer in Section 3. ρ(T,ΔKID) is the Spearman correlation between per-condition mean trust score and per-condition ΔKID (n=16 for CelebA, n=50 for RxRx1). All three ablation scorers produce near-degenerate P95-real thresholds (gen scores sit almost entirely below the 95 th-percentile real threshold), yielding $\approx 100\%$ acceptance; kNN still recovers Mahalanobis-level ranking correlation, while the linear probe and CLIP alignment collapse on both axes. Mahalanobis held-out rows are reproduced from Table 1; full-support rows are kept here as sanity checks.

Model	Setting	Accept%	$\text{KID}_{\text{trust}}\downarrow$	$\text{KID}_{\text{baseline}}\downarrow$	$\Delta\%\uparrow$	ρ(T)$\uparrow$
CelebA — Mahalanobis (ours)						
Vanilla	full	85.6	$0.266_{\pm.005}$	$0.288_{\pm.012}$	+7.5	+0.87
Vanilla	held-out	58.7	$0.224_{\pm.009}$	$0.368_{\pm.020}$	+39.1	+0.96
REPA (DINOv3)	full	87.9	$0.163_{\pm.005}$	$0.194_{\pm.004}$	+15.6	+0.83
REPA (DINOv3)	held-out	55.4	$0.228_{\pm.010}$	$0.401_{\pm.031}$	+43.1	+0.96
REPA (SigLIP)	full	85.2	$0.225_{\pm.009}$	$0.265_{\pm.008}$	+15.1	+0.88
REPA (SigLIP)	held-out	56.1	$0.239_{\pm.018}$	$0.423_{\pm.020}$	+43.6	+0.97
CelebA — Linear probe (per-attribute energy)						
Vanilla	full	100.0	$0.292_{\pm.011}$	$0.294_{\pm.008}$	+0.9	−0.29
Vanilla	held-out	100.0	$0.350_{\pm.012}$	$0.343_{\pm.033}$	−2.2	−0.04
REPA (DINOv3)	full	100.0	$0.192_{\pm.008}$	$0.194_{\pm.003}$	+1.0	−0.27
REPA (DINOv3)	held-out	100.0	$0.388_{\pm.015}$	$0.409_{\pm.020}$	+5.1	+0.21
REPA (SigLIP)	full	100.0	$0.256_{\pm.013}$	$0.256_{\pm.013}$	+0.0	−0.29
REPA (SigLIP)	held-out	100.0	$0.391_{\pm.016}$	$0.417_{\pm.033}$	+6.2	+0.34
CelebA — kNN (per-attr., 5-th NN summed)						
Vanilla	full	99.9	$0.292_{\pm.013}$	$0.302_{\pm.012}$	+3.2	+0.89
Vanilla	held-out	99.8	$0.357_{\pm.023}$	$0.358_{\pm.021}$	+0.3	+0.92
REPA (DINOv3)	full	99.9	$0.188_{\pm.007}$	$0.187_{\pm.010}$	−0.5	+0.83
REPA (DINOv3)	held-out	99.7	$0.403_{\pm.012}$	$0.396_{\pm.026}$	−1.7	+0.88
REPA (SigLIP)	full	99.9	$0.263_{\pm.011}$	$0.263_{\pm.010}$	+0.2	+0.91
REPA (SigLIP)	held-out	99.8	$0.424_{\pm.020}$	$0.411_{\pm.024}$	−3.0	+0.87
CelebA — CLIP alignment (16 joint-combo prompts)						
Vanilla	full	100.0	$0.649_{\pm.013}$	$0.656_{\pm.008}$	+1.0	−0.55
Vanilla	held-out	100.0	$0.729_{\pm.016}$	$0.720_{\pm.024}$	−1.3	+0.60
REPA (DINOv3)	full	100.0	$0.625_{\pm.012}$	$0.603_{\pm.015}$	−3.6	−0.48
REPA (DINOv3)	held-out	100.0	$0.770_{\pm.020}$	$0.775_{\pm.034}$	+0.7	+0.77
REPA (SigLIP)	full	100.0	$0.631_{\pm.025}$	$0.613_{\pm.011}$	−2.8	−0.65
REPA (SigLIP)	held-out	100.0	$0.733_{\pm.011}$	$0.751_{\pm.030}$	+2.5	+0.73
RxRx1 (50-condition subset) — Mahalanobis (ours)						
Vanilla	full	14.9	$0.202_{\pm.003}$	$0.307_{\pm.018}$	+34.2	+0.87
Vanilla	held-out	4.4	$0.196_{\pm.006}$	$0.323_{\pm.013}$	+39.4	+0.90
REPA (DINOv3)	full	21.7	$0.175_{\pm.004}$	$0.276_{\pm.013}$	+36.7	+0.83
REPA (DINOv3)	held-out	5.5	$0.197_{\pm.006}$	$0.333_{\pm.019}$	+40.8	+0.90
REPA (SigLIP)	full	18.4	$0.190_{\pm.002}$	$0.301_{\pm.007}$	+37.1	+0.84
REPA (SigLIP)	held-out	6.1	$0.181_{\pm.003}$	$0.312_{\pm.019}$	+42.0	+0.87
RxRx1 — Linear probe (per-attribute energy)						
Vanilla	full	100.0	$0.304_{\pm.015}$	$0.312_{\pm.015}$	+2.4	−0.09
Vanilla	held-out	100.0	$0.329_{\pm.020}$	$0.330_{\pm.023}$	+0.5	+0.08
REPA (DINOv3)	full	100.0	$0.278_{\pm.019}$	$0.274_{\pm.019}$	−1.3	+0.15
REPA (DINOv3)	held-out	100.0	$0.351_{\pm.019}$	$0.354_{\pm.013}$	+0.8	+0.10
REPA (SigLIP)	full	100.0	$0.303_{\pm.011}$	$0.311_{\pm.011}$	+2.7	+0.01
REPA (SigLIP)	held-out	100.0	$0.330_{\pm.025}$	$0.318_{\pm.018}$	−3.9	+0.08
RxRx1 — kNN (per-attr., 5-th NN summed)						
Vanilla	full	100.0	$0.304_{\pm.015}$	$0.312_{\pm.015}$	+2.4	+0.91
Vanilla	held-out	100.0	$0.329_{\pm.020}$	$0.330_{\pm.023}$	+0.5	+0.88
REPA (DINOv3)	full	100.0	$0.278_{\pm.019}$	$0.274_{\pm.019}$	−1.3	+0.92
REPA (DINOv3)	held-out	100.0	$0.351_{\pm.019}$	$0.354_{\pm.013}$	+0.8	+0.91
REPA (SigLIP)	full	100.0	$0.303_{\pm.011}$	$0.311_{\pm.011}$	+2.7	+0.90
REPA (SigLIP)	held-out	100.0	$0.330_{\pm.025}$	$0.318_{\pm.018}$	−3.9	+0.89

L Detailed results

This appendix collects the auxiliary results referenced from Sections 5.1.1, 5.1.2 and 5.2: full-support sanity-check rows for the P95-real-threshold selection, condition-level Spearman correlations of the trust score with its realism and faithfulness components, the trust/realism/faithfulness decile decomposition on `CelebA`, the analogous DINOv3 decile binning on `RxRx1`, and the during-generation decile binning under the translator.

Full-support sanity checks. Table 17 extends Table 1 with the full-support training-distribution rows. Relative KID gains shrink under full support, where the generator already covers the target conditions, but remain positive on both datasets and all three generators (+7–16% on `CelebA`, +34–37% on `RxRx1`). The held-out rows are the deployment-relevant ones and are reproduced compactly in Table 1.

Table 17: Complete P95-real-threshold post-generation sample selection (DINOv3 / SigLIP trust scoring). Full-support rows are sanity checks; held-out rows are the main deployment-relevant setting and are reproduced compactly in Table 1. Threshold set at the 95th percentile of real-sample trust scores; for marginal models, only seen-combo real samples are used for calibration. KID measured in DINOv3 space via shuffled subsampling (k=500, repeated for CIs). Random baseline: random subset of the same size, each compared against its condition-matched real pool. $\Delta\%$ = relative KID improvement of trust-selected over random (positive = better).

Model	Setting	Accept%	$\text{KID}_{\text{trust}}\downarrow$	$\text{KID}_{\text{baseline}}\downarrow$	$\Delta\%\uparrow$
CelebA					
Vanilla	full	85.6	$0.266_{\pm.005}$	$0.288_{\pm.012}$	+7.5
Vanilla	held-out	58.7	$0.224_{\pm.009}$	$0.368_{\pm.020}$	+39.1
REPA (DINOv3)	full	87.9	$0.163_{\pm.005}$	$0.194_{\pm.004}$	+15.6
REPA (DINOv3)	held-out	55.4	$0.228_{\pm.010}$	$0.401_{\pm.031}$	+43.1
REPA (SigLIP)	full	85.2	$0.225_{\pm.009}$	$0.265_{\pm.008}$	+15.1
REPA (SigLIP)	held-out	56.1	$0.239_{\pm.018}$	$0.423_{\pm.020}$	+43.6
RxRx1 (50-condition subset)					
Vanilla	full	14.9	$0.202_{\pm.003}$	$0.307_{\pm.018}$	+34.2
Vanilla	held-out	4.4	$0.196_{\pm.006}$	$0.323_{\pm.013}$	+39.4
REPA (DINOv3)	full	21.7	$0.175_{\pm.004}$	$0.276_{\pm.013}$	+36.7
REPA (DINOv3)	held-out	5.5	$0.197_{\pm.006}$	$0.333_{\pm.019}$	+40.8
REPA (SigLIP)	full	18.4	$0.190_{\pm.002}$	$0.301_{\pm.007}$	+37.1
REPA (SigLIP)	held-out	6.1	$0.181_{\pm.003}$	$0.312_{\pm.019}$	+42.0

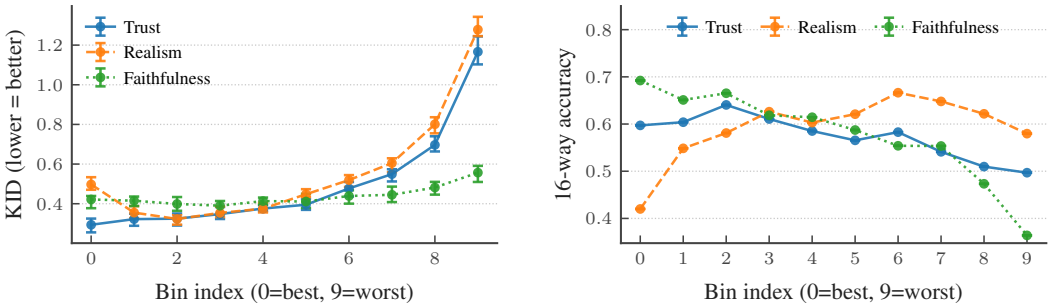
Component-level condition correlations. Table 18 reports the per-condition Spearman correlations of trust, realism, and faithfulness against the held-out ΔKID. Trust is the strongest predictor across both datasets and all three generators ($\rho(T) \geq 0.88$). The decomposition shows that on `RxRx1` the realism signal carries most of the predictive power, consistent with the held-out perturbations producing manifold drift that the generator cannot reach; on `CelebA` both faithfulness and realism contribute, with faithfulness being the slightly stronger of the two on the controlled compositional split.

Table 18: Condition-level Spearman correlations under the harder support-shift regime: CelebA held-out and RxRx1 held-out (DINOv3/SigLIP scoring, marginal models). $\rho(T)$ = trust, ρ(R) = realism, ρ(F) = faithfulness. The main text reports only $\rho(T)$ in Table 1.

	CelebA (n=16)			RxRx1 (n=50)		
Model	$\rho(T)\uparrow$	ρ(R)$\uparrow$	ρ(F)$\uparrow$	$\rho(T)\uparrow$	ρ(R)$\uparrow$	ρ(F)$\uparrow$
Vanilla	0.96	0.69	0.84	0.90	0.88	0.68
REPA (DINOv3)	0.96	0.71	0.86	0.90	0.88	0.75
REPA (SigLIP)	0.96	0.76	0.84	0.88	0.86	0.75

Trust/realism/faithfulness decile decomposition (`CelebA`). Figure 8 re-plots the main-text Figure 2 with all three score components. Realism alone does not track downstream condition accuracy,

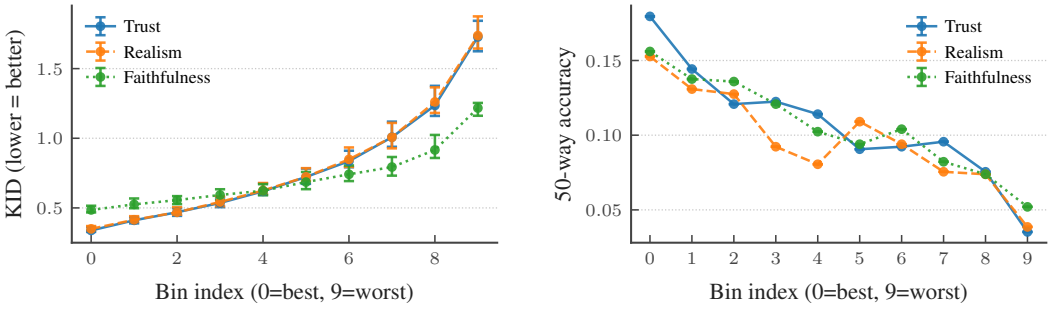
consistent with realism being a manifold-quality rather than a condition-aware signal; trust combines the two and drives the ordering on both panels.



(a) CelebA ΔKID by decile. (b) CelebA downstream 16-way accuracy.

Figure 8: **Trust / realism / faithfulness decomposition for the main-text Figure 2.** Same setting as the main figure (REPA-DINOv3 held-out, DINOv3 scoring), but plotting all three score components. Realism does not track downstream condition accuracy, which is consistent with realism not being a condition-aware signal; trust drives the ordering on both panels.

`RxRx1` **DINOv3 decile binning.** Figure 9 shows the analogous DINOv3-space decile binning on the `RxRx1` 50-condition subset. Better trust deciles produce both lower ΔKID and stronger downstream classifiers, mirroring the `CelebA` ordering. The main-text discussion emphasizes the CellProfiler morphology validation in Section 5.1.3 because CP features are independent of the learned DINOv3 validation encoder used for KID.



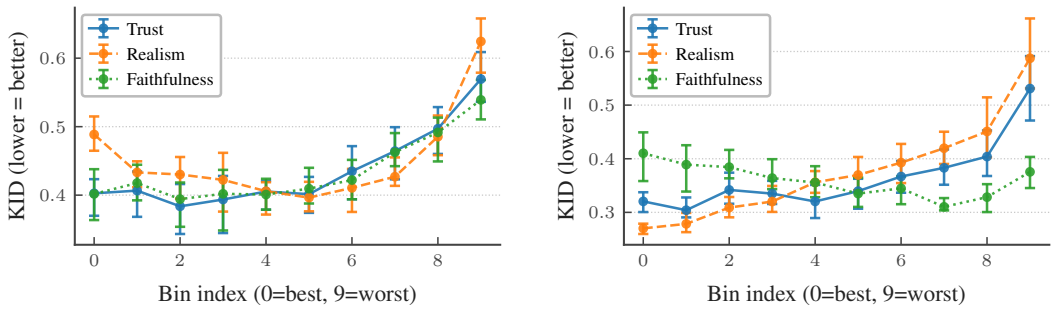
(a) RxRx1 ΔKID by trust decile. (b) RxRx1 downstream classification.

Figure 9: **RxRx1 DINOv3 decile binning (REPA-DINOv3 held-out, DINOv3 scoring, 50-condition subset).** These learned-encoder trends support the same ordering story as CelebA, but the main text emphasizes the CellProfiler morphology validation because it is independent of the learned DINOv3 validation encoder.

During-generation decile binning. Figure 10 confirms that the translator features reproduce the same monotonic trust ordering on both `CelebA` and `RxRx1` without decoding the sample, complementing the support-shift evidence in Table 3.

L.1 Complete during-generation table

Table 19 reports the full during-generation table, including full-support sanity-check rows that are omitted from the main text. The main text keeps only the held-out rows because those are aligned with the missing-target evaluation setting.



(a) CelebA ΔKID by decile.

(b) RxRx1 ΔKID by decile.

Figure 10: **During-generation decile binning after translation.** The translator features recover the monotonic trust trend without decoding the sample and re-encoding it through the feature extractor.

Table 19: Complete non-DINOv3 trust scoring: model-internal REPA features and our translator. *Left*: P95-real-threshold selection (KID in DINOv3 space, random baseline with per-draw condition-matched real). *Right*: condition-level Spearman correlations. Full-support rows are sanity checks; held-out rows are summarized in Table 3.

		P95-real Selection				Correlations		
Model	Setting	Accept%	$\text{KID}_{\text{trust}}\downarrow$	$\text{KID}_{\text{baseline}}\downarrow$	$\Delta\%\uparrow$	$\rho(T)\uparrow$	$\rho(\text{R})\uparrow$	$\rho(\text{F})\uparrow$
CelebA — REPA aligned_mean (during generation)								
REPA (DINOv3)	full	97.8	$0.187_{\pm.013}$	$0.195_{\pm.010}$	+4.1	0.73	0.74	0.29
REPA (DINOv3)	held-out	62.7	$0.278_{\pm.015}$	$0.404_{\pm.021}$	+31.2	0.95	0.64	0.91
REPA (SigLIP)	full	97.2	$0.254_{\pm.008}$	$0.258_{\pm.012}$	+1.6	0.81	0.80	0.53
REPA (SigLIP)	held-out	66.7	$0.293_{\pm.024}$	$0.425_{\pm.018}$	+31.1	0.95	0.56	0.89
CelebA, translator (during generation)								
Vanilla	full	32.6	$0.254_{\pm.015}$	$0.299_{\pm.007}$	+15.2	0.61	0.29	0.57
Vanilla	held-out	23.6	$0.195_{\pm.008}$	$0.378_{\pm.023}$	+48.4	0.83	0.74	0.82
REPA (DINOv3)	full	89.2	$0.188_{\pm.010}$	$0.197_{\pm.005}$	+4.4	0.78	0.64	0.71
REPA (DINOv3)	held-out	48.4	$0.250_{\pm.016}$	$0.397_{\pm.017}$	+37.2	0.92	0.66	0.91
REPA (SigLIP)	full	87.2	$0.248_{\pm.009}$	$0.263_{\pm.011}$	+5.7	0.56	0.29	0.52
REPA (SigLIP)	held-out	17.2	$0.308_{\pm.015}$	$0.409_{\pm.015}$	+24.8	0.88	0.83	0.86
RxRx1 — REPA aligned_mean (during generation)								
REPA (DINOv3)	full	26.4	$0.378_{\pm.018}$	$0.283_{\pm.005}$	−33.4	−0.01	0.10	−0.27
REPA (DINOv3)	held-out	1.4	$0.537_{\pm.032}$	$0.329_{\pm.037}$	−63.4	0.09	0.05	0.19
REPA (SigLIP)	full	20.6	$0.628_{\pm.029}$	$0.315_{\pm.006}$	−99.5	−0.13	−0.08	−0.38
REPA (SigLIP)	held-out	3.9	$0.564_{\pm.022}$	$0.293_{\pm.023}$	−92.3	0.13	−0.11	0.55
RxRx1, translator (during generation)								
Vanilla	full	37.0	$0.277_{\pm.007}$	$0.290_{\pm.017}$	+4.7	0.70	0.36	0.67
Vanilla	held-out	38.4	$0.260_{\pm.003}$	$0.346_{\pm.016}$	+24.9	0.78	0.31	0.76
REPA (DINOv3)	full	82.7	$0.276_{\pm.017}$	$0.281_{\pm.016}$	+1.9	0.18	0.67	−0.17
REPA (DINOv3)	held-out	62.3	$0.246_{\pm.003}$	$0.372_{\pm.037}$	+33.8	0.88	0.47	0.80
REPA (SigLIP)	full	80.3	$0.290_{\pm.015}$	$0.316_{\pm.016}$	+8.1	0.65	0.68	0.00
REPA (SigLIP)	held-out	60.5	$0.235_{\pm.011}$	$0.329_{\pm.018}$	+28.7	0.70	0.31	0.68

L.2 Fine-grained timestep (full table)

Table 20 gives the full per-k numbers summarized by Figure 4 in the main text.

Scoring step	ρ_{trust}	$\text{KID}_{\text{trust}}\downarrow$	$\text{KID}_{\text{baseline}}\downarrow$	ΔKID%$\uparrow$ (accept)	Image Δ L2	Steps saved
k=0 ($t\approx$1.00)	0.21	0.290	0.358	+18.8% (19%)	–	$\approx$100%
k=27 ($t\approx$0.90)	0.08	0.309	0.365	+15.5% (19%)	74.5	$\approx$89%
k=55 ($t\approx$0.79)	0.37	0.263	0.356	+26.1% (19%)	33.7	$\approx$78%
k=83 ($t\approx$0.68)	0.58	0.240	0.357	+33.0% (19%)	24.2	$\approx$67%
k=110 ($t\approx$0.58)	0.65	0.234	0.355	+34.1% (21%)	19.3	$\approx$56%
k=138 ($t\approx$0.47)	0.73	0.223	0.346	+35.5% (21%)	17.9	$\approx$45%
k=166 ($t\approx$0.36)	0.75	0.208	0.342	+39.1% (21%)	16.7	$\approx$34%
k=193 ($t\approx$0.26)	0.79	0.207	0.367	+43.6% (22%)	15.6	$\approx$23%
k=221 ($t\approx$0.15)	0.78	0.202	0.356	+43.1% (23%)	14.7	$\approx$12%
k=248 ($t\approx$0.04)	0.79	0.210	0.347	+39.4% (24%)	12.3	$\approx$1%
Post generation	0.96	0.221	0.365	+39.3% (59%)	–	—

Table 20: **Fine-grained timestep ablation for cheap internal trust scoring on CelebA held-out** (Vanilla SiT-B/2, 250-step sampler). At each scoring step k, the predicted clean latent $\hat{x}_0$ is passed through the translator into SigLIP space (no VAE decode, no encoder pass) before Mahalanobis scoring. Spearman ρ_{trust} is the condition-level correlation against ΔKID. KID_{acc} and $\text{KID}_{\text{baseline}}$ are computed on the P95-real-threshold-accepted set and random matched-size subsets, respectively. "Image Δ L2" is the mean L2 distance between consecutively captured VAE-decoded $\hat{x}_0$ images, quantifying how much the predicted clean image still moves between capture steps (undefined at k=0). "Steps saved" is the fraction of sampler steps avoided by abstaining at k. The final row is the post-generation DINOv3 oracle (reproduced from Table 16, Vanilla held-out). The ranking signal grows monotonically with k, and P95-real-threshold ΔKID matches or exceeds the post-generation oracle from $k\approx$166 onward while still saving $\geq$34% of generation compute.

L.3 RxRx1 timestep ablation

The same fine-grained timestep ablation on `RxRx1` held-out is shown in Figure 11 for the Vanilla SiT-B/2 generator with the translator scoring path and the same 250-step sampler used for `CelebA`. The qualitative picture matches Figure 4: the trust signal is already useful well before the final decoded sample, and the per-step L2 change in the predicted-clean image $\hat{x}_0$ falls monotonically as the trajectory settles ($\sim$ 7.0 at $k=27$ down to $\sim$ 4.6 at $k=248$).

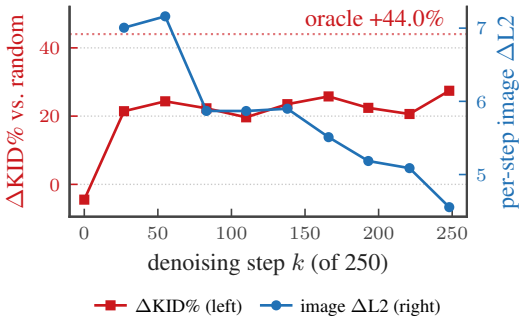


Figure 11: `RxRx1` **timestep ablation (Vanilla SiT B/2, 250-step sampler, translator features).** Red squares: P95-real ΔKID% of the FPR95-accepted subset against a condition-matched random subset, evaluated on intermediate predicted-clean latents $\hat{x}_0(k)$ projected through the translator. Blue circles: per-step L2 change in the VAE-decoded $\hat{x}_0$. Dashed red line: post-generation DINOv3 oracle (+44.0%). Compared to Figure 4, the during generation signal saturates earlier ($k\approx$27) but plateaus below the post-generation oracle, reflecting the harder support shift on `RxRx1`.

Two differences are worth flagging. First, on `RxRx1` the during-generation ΔKID saturates very early — by k = 27 the FPR95-selected subset already reaches roughly +21% over a condition-matched random subset, and subsequent steps oscillate in the +20% to +27% band without a further monotonic trend. Second, the during-generation score does not match the post-generation DINOv3 oracle (+44.0%, dashed line): on this harder support-shift regime the mapped internal feature captures roughly 60% of the discriminative power that the full feature extractor pass recovers after decoding. This gap is consistent with the regime in which Mahalanobis-based scoring is intrinsically harder on `RxRx1` (smaller per-attribute margins on the sirna axis, Tables 4 and 5); the during-generation extension preserves a useful trust ordering but does not recover the full precision of the post-generation oracle. The practical compute trade-off is therefore sharper than on `CelebA`: abstaining as early as $k=27$ saves $\approx$89% of the denoising trajectory while retaining most of the during-generation ΔKID signal, but full post-generation scoring remains the higher-precision choice when compute is not the bottleneck.

M Encoder sensitivity and evaluation-space robustness

M.1 Sensitivity to the scoring encoder

We investigate the sensitivity of the proposed trust score to the auxiliary feature extractor used for scoring. In addition to DINOv3 and SigLIP, we evaluate DINOv2-Small and MAE as alternative representations. For each encoder, we recompute the real-data statistics and post-generation trust scores, using the same P95-real selection protocol. We report relative ΔKID improvement over condition-matched random selection.

Table 21: Relative ΔKID improvement (%) of trust-selected samples over condition-matched random selection for different scoring encoders. KID is evaluated in DINOv3 feature space. Higher is better.

Setting	DINOv3	SigLIP	DINOv2-S	MAE
CelebA (full)	16.2	8.4	7.4	3.0
CelebA (marginal)	25.9	8.7	18.6	7.1
RxRx1 (full)	39.7	18.4	14.1	5.6
RxRx1 (marginal)	40.4	24.4	12.2	16.3

Table 21 shows that all four evaluated feature spaces yield positive selection improvements across both datasets and training regimes. However, the magnitude of the improvement varies substantially with the encoder, indicating that representation choice is an important practical factor.

M.2 Covariance geometry and encoder suitability

For Mahalanobis-based scoring, condition discrimination alone is not sufficient. The feature covariance must also retain enough effective rank for the estimated precision matrix to capture meaningful data geometry. We present participation ratio as a proxy for usefulness of the ℓ_2-normalized real features

$$\mathrm{PR}(\Sigma) = \frac{(\mathrm{tr}\,\Sigma)^2}{\mathrm{tr}(\Sigma^2)}, \tag{1}$$

together with the fraction of total variance explained by the ten largest eigenvalues.

Table 22: Covariance geometry of normalized real RxRx1 features. PR denotes the participation ratio, and Top-10 var. denotes the fraction of variance explained by the ten leading eigen-directions.

Feature space	PR	Top-10 var.
DINOv2-S	18.7	0.59
SigLIP	14.3	0.66
DINOv3	10.1	0.71
MAE	8.8	0.73
OpenPhenom	2.6	0.93

OpenPhenom provides an informative failure case. Although its representation retains condition-discriminative information, its covariance is strongly concentrated in a small number of directions. Consequently, many empirical eigenvalues are close to zero, making the regularized precision matrix highly dependent on the shrinkage target. The resulting Mahalanobis distances can therefore become dominated by regularization rather than meaningful feature-space geometry.

These results motivate two complementary diagnostics for choosing a scoring representation: condition decodability, evaluated through probes trained on real data and effective covariance rank, evaluated through spectral statistics. The results show latter is particularly important when the precision matrix is estimated from limited real data.

M.3 Sensitivity to the KID evaluation space

The previous ablation varies the representation used to compute the trust score while keeping the KID evaluation space fixed. We now investigate whether the observed sample-selection improvements depend on the feature space used to evaluate generation quality.

We compare relative ΔKID improvements computed using DINOv3 and SigLIP feature spaces. The results are reported in Table 23.

Table 23: Relative ΔKID improvement (%) of trust selection over condition-matched random selection using different KID evaluation spaces. Higher is better.

Setting	DINOv3 KID	SigLIP KID
RxRx1 (full)	5.2	25.3
RxRx1 (marginal)	4.2	14.8
CelebA (full)	9.5	6.9
CelebA (marginal)	18.4	23.0

Across both datasets and training regimes, trust-based selection improves the evaluation metric under both feature spaces. The magnitude of the improvement depends on the evaluation encoder, but its direction is consistent. This provides evidence that the selection improvements are not limited to DINOv3 KID.

N Acceptance–quality trade-off on RxRx1

Seeing the low acceptance rate for RxRx1, we investigate how the acceptance threshold affects the quality and coverage of generated samples. The experiment uses the marginally trained vanilla generator and 25 held-out conditions, with 2,500 generated samples in total.

We accept a sample when its trust score falls below a specified percentile of the trust-score distribution of real/seen samples. We vary this percentile from P95 to P99.5 and measure the relative ΔKID improvement over a condition-matched random subset of equal size.

To assess condition faithfulness independently of the trust score, we additionally evaluate accepted and rejected samples using a SigLIP-based siRNA classifier trained only on real samples.

Table 24: Acceptance–quality trade-off on 25 unseen RxRx1 conditions (2,500 generated samples). ΔKID gain is the relative improvement over condition-matched random selection. Classifier accuracy provides an independent proxy for siRNA faithfulness.

Threshold	Accept (%)	ΔKID gain (%)	Acc. accepted	Acc. rejected
P95	3.2	34.3	0.156	0.088
P97	4.5	25.7	0.158	0.087
P99	9.2	11.7	0.152	0.084
P99.5	12.2	10.0	0.138	0.084

Table 24 shows that relaxing the threshold increases the fraction of accepted generations, while progressively reducing the relative improvement in ΔKID.

Additionally, at every thresholding point, the independent classifier achieves higher accuracy on accepted than rejected samples. This indicates that the selection rule also captures differences in perturbation faithfulness.

The low acceptance rate is therefore best interpreted as a conservative decision rule. The score can support less selective filtering, but this comes with a weaker average quality advantage.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Conclusion section discusses potential limitations and future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Paper provides the full set of assumptions together with their proofs for its theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We give a detailed explanation of the proposed algorithm that is needed to reproduce the main experimental results, together with the source code in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Details on the dataset are discussed in the experiments section and the code is released through the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The experiments section and Appendix E specify all the details necessary to understand the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports confidence intervals whenever applicable.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix F.3 discusses the compute resources in detail.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix D discusses both positive and negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The owners of the used assets are explicitly credited through citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the full training, sampling, scoring, and mapper-training pipeline in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as part of the core methodology; any use was limited to writing, editing and code formatting assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.