MEASUREMENT-ALIGNED DIFFUSION: CLOSING THE LOOP BETWEEN GENERATION AND CATALOG MEASUREMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Using a generative model as a scientific instrument requires more than plausible images: catalog controls must be meaningful, behaviorally faithful, and testable against real observations. We use an eight-dimensional catalog vector—redshift, multiband fluxes, ellipticity, effective radius, and Sérsic index—as an interpretable interface to a conditional diffusion model, aligned by a frozen, non-adaptive measurement proxy. On the test set of $2{,}145$ real images, the proxy has mean $R^2 = 0.925$ and degrades in the extreme Sérsic tail. With that proxy fixed, six one-coordinate counterfactual paths track the requested axis with Spearman correlation at least 0.96, and leave-group ablations raise the error on the removed axes, most strongly for shape and scale. On 100 matched catalog conditions, alignment lowers the closed-loop error of this same proxy by 56.4% when sampling from noise and by 58.4% in image-to-image sampling, relative to an unaligned diffusion baseline. Accepted samples then retrieve real galaxies. In leave-one-out rare-object search over $46{,}130$ catalog images, Recall@500 is 0.588, compared with 0.052 for catalog-space nearest neighbors. Screening $3{,}128{,}494$ cutouts with 461 prototypes, two blinded experts mark 79.5% of their judgments on 100 pairs drawn from the top-50 neighbors as visually similar. Closed-loop scores measure consistency with the fixed proxy; retrieval and expert judgments test the outputs on real survey images. The same interface can then query future wide-field surveys by rendering a stated catalog hypothesis as a prototype and matching it to real cutouts.

1 INTRODUCTION

Conditional generators can produce images that look realistic while ignoring, mixing, or silently dropping the controls they were asked to obey (Ho et al., 2020; Song et al., 2021; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Ho & Salimans, 2022). When those controls are opaque latents, this may be only a quality failure. When they are named physical quantities, the failure is deeper: a sample that does not obey its condition cannot serve as scientific evidence (Roscher et al., 2020; Doshi-Velez & Kim, 2017; Rudin, 2019). Latent traversals and post-hoc attributions are likewise hard to state, reproduce, or falsify (Lipton, 2018; Adebayo et al., 2018; Jacovi & Goldberg, 2020).

The learning question we study is therefore behavioral and checkable: given an explicit, semantically labeled condition, can a generator be trained and tested so that (i) each named coordinate is used specifically, (ii) changing one coordinate induces the intended measurable change, and (iii) accepted samples retrieve coherent real observations rather than remaining in-distribution fakes. We treat this as a closed loop among generation, measurement, and retrieval. Following concept-based and intervention-oriented interpretability (Kim et al., 2018; Koh et al., 2020; Wachter et al., 2018; Goyal et al., 2019), we require three properties: *semantic grounding* (every control is a catalog measurement), *counterfactual faithfulness* (a coordinate change is recovered by a frozen proxy, with limited leakage on held groups), and *query utility* (a low-density condition yields prototypes that rank real images). Freezing the measurement network is essential: if the evaluator is updated on generated images, closed-loop error can fall because the proxy adapts, not because the generator improved.

Galaxy imaging is a demanding test bed for this loop, not the claim itself. Wide-field surveys provide paired RGB cutouts and catalog measurements (Dey et al., 2019; Lintott et al., 2008;

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

Willett et al., 2013; Walmsley et al., 2023; Dieleman et al., 2015; Domínguez Sánchez et al., 2018). Let $\mathbf{c} = (z, F_g, F_r, F_z, e_1, e_2, R_e, n_{\text{Sérsic}}) \in \mathbb{R}^8$ denote redshift, three-band fluxes, two ellipticity components, effective radius, and Sérsic index (Sérsic, 1963). These quantities are inferred from cutouts, not official spectroscopic redshifts or pipeline photometry; three-band photometric redshift is only a coarse proxy. Prior galaxy generators condition on physical or morphological labels and then score fidelity after sampling (Lanusse et al., 2021; Smith et al., 2022; Yunus et al., 2024). They do not freeze a measurement network into training, test whether named controls are used specifically, or turn accepted generations into survey queries. Morphology-conditioned diffusion has been used to synthesize rare classes for classifier augmentation (Ma et al., 2025); we instead retrieve real cutouts.

Approach. We contribute a *method plus an evaluation protocol* for named-control conditional generation. A diffusion generator with classifier-free guidance and a super-resolution network (Ho et al., 2020; Ho & Salimans, 2022; Liang et al., 2021) maps $\mathbf{c}$ to a 224×224 image. A frozen proxy trained only on real images supplies a non-adaptive alignment loss; accepted generations become prototypes that query a survey archive. We evaluate proxy reliability, group ablations, six counterfactual paths, alignment versus an independently trained cascade, leave-one-out retrieval, and blinded expert review. Closed-loop MAE uses the *same* frozen proxy as the training loss, so it measures proxy consistency rather than astrophysical truth; retrieval and experts are the external tests. The claim is behavioral: we do not assert mechanistic transparency of the U-Net.

Contributions.

- **Frozen-proxy alignment for conditional diffusion.** We jointly fine-tune a conditional DDPM and super-resolution module with a SmoothL1 consistency loss through a frozen measurement network, preventing co-adaptation between generator and evaluator. Alignment reduces closed-loop error of that same proxy by 56.4% (de novo) and 58.4% (img2img) relative to an independently trained cascade.
- **Tests that named controls are used specifically.** Group ablations (brightness, color, shape, scale) raise error mainly on the ablated group, and six one-coordinate counterfactual paths recover requested measurements with Spearman correlation at least 0.96 in well-supported ranges.
- **Prototypes as inspectable queries over real data.** Accepted samples retrieve held-out rare objects with Recall@500 of 0.588 versus 0.052 for catalog-space nearest neighbors, and blinded experts judge 79.5% of prototype–neighbor pairs from $3{,}128{,}494$ survey cutouts as similar.

2 Related Work

Conditional diffusion. Denoising diffusion and score-based models generate high-quality images by reversing a noise process (Ho et al., 2020; Song et al., 2021; Nichol & Dhariwal, 2021). Classifier and classifier-free guidance strengthen the effect of a condition at sampling time (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). Guidance improves visual match to a label or embedding, but does not by itself test whether each named coordinate is used, nor whether a downstream measurement of the sample recovers that coordinate. Our generator uses classifier-free guidance, then adds a frozen measurement loss so that photometric and morphological controls remain checkable after sampling.

Concepts, interventions, and control. Concept bottleneck models and TCAV treat named attributes as an interface to predictions (Koh et al., 2020; Kim et al., 2018). Counterfactual methods ask what happens if one input is changed while others are held fixed (Wachter et al., 2018; Goyal et al., 2019). We apply the same idea to a generator: group ablations test specificity, and one-coordinate paths test whether recovered measurements track the requested change. Interpretability caveats still apply (Lipton, 2018; Adebayo et al., 2018; Jacovi & Goldberg, 2020; Rudin, 2019): passing these tests does not prove that the U-Net stores a causal physical model.

Scientific image generation. Galaxy GANs and diffusion models have been conditioned on redshift, photometry, or morphology and then scored with FID or parameter error after the fact (Lanusse et al., 2021; Smith et al., 2022; Yunus et al., 2024; Ma et al., 2025). Related work in scientific

ML emphasizes that visual realism is not scientific validity (Roscher et al., 2020; Doshi-Velez & Kim, 2017). We differ in three design choices: the measurement network is frozen during generator updates; control use is tested by ablation and counterfactuals; accepted samples are used to retrieve real survey cutouts rather than to augment a classifier.

Self-supervised vision and retrieval. Self-supervised ViTs provide transferable image embeddings (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). We use a LoRA-adapted ViT (Hu et al., 2022) only as a frozen measurement head, and a separate frozen encoder for FAISS retrieval (Johnson et al., 2021). The two roles are kept apart so that retrieval is not optimized against the same loss used to train the generator.

3 METHOD

Problem Formulation. Let $\mathbf{x}_{\mathrm{HR}} \in \mathbb{R}^{3\times224\times224}$ denote a high-resolution survey image from the DESI Legacy Imaging Surveys (Dey et al., 2019) and let

$$\mathbf{c} = \left(z,\, F_g,\, F_r,\, F_z,\, e_1,\, e_2,\, R_e,\, n_{\mathrm{Sérsic}}\right) \in \mathbb{R}^8 \tag{1}$$

denote its catalog-derived measurement vector (redshift, three-band fluxes, ellipticity components, effective radius, and Sérsic index (Sérsic, 1963)). We seek to learn the conditional image distribution $p(\mathbf{x} \mid \mathbf{c})$ while exposing a scientifically interpretable interface between model behavior and survey discovery.

Framework Overview. The framework comprises three components (Fig. 1): (i) a conditional diffusion generator G_θ that synthesizes a 64×64 galaxy image from Gaussian noise and a target measurement vector; (ii) a super-resolution network S_ϕ that maps the low-resolution output to the 224×224 resolution required for scientific measurement; (iii) an independently trained and frozen measurement proxy $\mathcal{M}$ that estimates catalog-derived quantities from the generated image and supplies a differentiable measurement-alignment signal.

Training proceeds in two stages. Each component is first independently pre-trained on real observations. We then jointly fine-tune G_θ and S_ϕ while keeping $\mathcal{M}$ fixed, so that the proxy cannot adapt to generated artifacts and remains a non-adaptive training constraint.

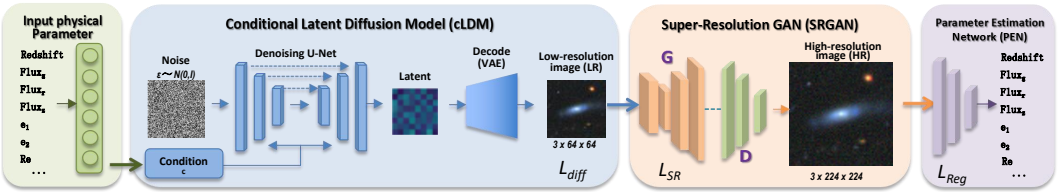


Figure 1: Overview of the measurement-aligned generative framework. Given an eight-dimensional catalog-derived measurement vector $\mathbf{c}$ comprising redshift, $g/r/z$-band fluxes, ellipticity components (e_1, e_2), effective radius R_e, and Sérsic index, a conditional DDPM first synthesizes a $3 \times 64 \times 64$ galaxy image. A SwinIR network then upsamples the image to 256×256, followed by a center crop to $3 \times 224 \times 224$. During joint fine-tuning, a fixed measurement proxy provides a differentiable consistency loss between the requested and recovered measurements. Solid arrows denote forward computation; dashed arrows denote gradient propagation during joint fine-tuning.

3.1 CONDITIONAL DIFFUSION GENERATOR

We represent $p_\theta(\mathbf{x}_{64} \mid \mathbf{c})$ with a conditional denoising diffusion probabilistic model (Ho et al., 2020; Nichol & Dhariwal, 2021). The denoising backbone is a U-Net ($\sim$13.4M parameters) with base channel width 64, channel multipliers $(1, 2, 4)$, Mish activations, and self-attention at the lower-resolution feature levels. Detailed architecture specifications and training hyperparameters are provided in Table 6 (Appendix A).

All condition statistics are computed exclusively from the training catalog. Each component is standardized as

$$\widetilde{c}_d = \frac{c_d - \mu_d}{\sigma_d}, \qquad d = 1, \ldots, 8, \tag{2}$$

where μ_d and σ_d denote the training-set mean and standard deviation. The standardized vector $\widetilde{\mathbf{c}}$ is projected by a multilayer perceptron and injected into the residual blocks of the U-Net.

Given a clean low-resolution image $\mathbf{x}_0$, the forward process is

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

with $T = 1000$ diffusion steps and a cosine noise schedule (Nichol & Dhariwal, 2021). The network predicts the injected noise via $\epsilon_\theta(\mathbf{x}_t, t, \widetilde{\mathbf{c}})$.

Classifier-free guidance is enabled by replacing the condition with a null embedding with probability 0.1 during training. At inference, conditional and unconditional predictions are combined as

$$\widehat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, \mathbf{c}) = \epsilon_\theta(\mathbf{x}_t, t, \emptyset) + s\big[\epsilon_\theta(\mathbf{x}_t, t, \widetilde{\mathbf{c}}) - \epsilon_\theta(\mathbf{x}_t, t, \emptyset)\big], \tag{4}$$

where $s = 6.0$ is the guidance scale. An exponential moving average of the diffusion parameters with decay 0.995 is used during sampling. We employ stochastic DDIM sampling (Song et al., 2021) ($\eta_{\mathrm{DDIM}} = 1.0$) for both de novo generation and img2img restoration; see Table 7 in Appendix A for full inference settings.

The independent diffusion pre-training objective is

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{c}, t, \boldsymbol{\epsilon}}\Big[\big\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t, \widetilde{\mathbf{c}})\big\|_2^2\Big]. \tag{5}$$

3.2 Super-Resolution Module

The module S_ϕ is implemented with SwinIR (Liang et al., 2021). It maps a 64×64 RGB input to a 256×256 output via $4\times$ upsampling; a symmetric center crop then yields a 224×224 image,

$$\mathbf{x}_{224} = \mathrm{Crop}_{224}\big(S_\phi(\mathbf{x}_{64})\big). \tag{6}$$

The network uses embedding dimension 180, six residual Swin Transformer stages (depth 6 each), and window size 8; full architectural details are listed in Table 8 (Appendix A).

During independent pre-training, S_ϕ is optimized on paired real low- and high-resolution observations:

$$\mathcal{L}_{\mathrm{SR}} = \mathbb{E}_{(\mathbf{x}_{\mathrm{LR}}, \mathbf{x}_{\mathrm{HR}})}\Big[\big\|\mathrm{Crop}_{224}\big(S_\phi(\mathbf{x}_{\mathrm{LR}})\big) - \mathbf{x}_{\mathrm{HR}}\big\|_1\Big]. \tag{7}$$

Horizontal and vertical flips are applied during pre-training. Unlike a fixed post-processing upsampler, S_ϕ remains trainable during joint fine-tuning because super-resolution artifacts can alter photometric and morphological measurements recovered from the final image.

3.3 Fixed Measurement Proxy

The measurement proxy is $\mathcal{M} = R \circ E$, where E is a DINOv3 ViT-L/16 visual encoder (Siméoni et al., 2025) and R is a two-layer regression head. The encoder is adapted to galaxy images with LoRA modules (Hu et al., 2022) ($r = 8, \alpha = 8$) in the final six transformer blocks; additional LayerNorm parameters are also fine-tuned. The regression head is $\mathrm{LayerNorm} \to \mathrm{Linear}(1024, 512) \to \mathrm{GELU} \to \mathrm{Dropout}(0.1) \to \mathrm{Linear}(512, 8)$. Complete training hyperparameters and adaptation details are given in Table 9 (Appendix A).

Spatial tokens are aggregated by excluding the CLS token, ranking the remaining patch tokens by ℓ_2 norm, and averaging the top $k = 64$ tokens to form the global representation for R. Input images are resized to 224×224 and normalized with the same preprocessing throughout proxy pre-training, joint fine-tuning, and evaluation.

Let $\widetilde{\mathcal{M}}(\mathbf{x})$ denote the proxy prediction in standardized target space. The proxy is trained independently on real high-resolution images by minimizing

$$\mathcal{L}_{\mathrm{proxy}} = \frac{1}{\sum_{d=1}^{8} w_d} \sum_{d=1}^{8} w_d\, \mathrm{SmoothL1}\big(\widetilde{\mathcal{M}}_d(\mathbf{x}_{\mathrm{HR}}),\, \widetilde{c}_d\big), \tag{8}$$

where $w_z = 10$ for redshift and $w_d = 1$ otherwise. Physical units are recovered only for reporting:

$$\widehat{c}_d = \mu_d + \sigma_d \widetilde{\mathcal{M}}_d(\mathbf{x}). \tag{9}$$

After pre-training, all parameters of $\mathcal{M}$ are frozen. During joint fine-tuning, $\mathcal{M}$ is placed in evaluation mode with dropout disabled to provide a stable gradient signal. During evaluation, MC Dropout (Gal & Ghahramani, 2016) is enabled only in the regression head with $K_{\mathrm{MC}} = 20$ stochastic forward passes. Given these predictions $\widehat{\mathbf{c}}^{(k)}$, we compute

$$\overline{\mathbf{c}} = \frac{1}{K_{\mathrm{MC}}} \sum_{k=1}^{K_{\mathrm{MC}}} \widehat{\mathbf{c}}^{(k)} \tag{10}$$

and use their sample variance as an approximate predictive uncertainty.

Freezing $\mathcal{M}$ prevents direct co-adaptation with the generator. However, $\mathcal{M}$ remains a learned approximation of the catalog-level mapping rather than an unbiased physical measurement operator. Generated images are therefore evaluated with both proxy-mediated consistency and independent catalog-based measurements.

3.4 Measurement-Aligned Joint Fine-Tuning

After independent pre-training, G_θ and S_ϕ are jointly optimized while $\mathcal{M}$ remains fixed:

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{SR}} \mathcal{L}_{\mathrm{SR}} + \lambda_{\mathrm{align}}(n) \mathcal{L}_{\mathrm{align}}, \tag{11}$$

where $\lambda_{\mathrm{SR}} = 1.0$, n is the optimization step, and $\lambda_{\mathrm{align}}(n)$ increases linearly from 0 to 0.1 over the first 500 joint-training steps.

At each joint-training iteration, we sample a low-noise timestep

$$t' \sim \mathcal{U}\{0, \ldots, \tau_{\max} - 1\}, \qquad \tau_{\max} = 0.02\,T. \tag{12}$$

A real low-resolution image is perturbed to $\mathbf{x}_{t'}$, and the diffusion model predicts the corresponding clean image

$$\widehat{\mathbf{x}}_0 = \frac{\mathbf{x}_{t'} - \sqrt{1 - \bar{\alpha}_{t'}}\, \epsilon_\theta(\mathbf{x}_{t'}, t', \widetilde{\mathbf{c}})}{\sqrt{\bar{\alpha}_{t'}}}. \tag{13}$$

The predicted clean image is clamped to the valid range, super-resolved, and center-cropped:

$$\widehat{\mathbf{x}}_{224} = \mathrm{Crop}_{224}\Big[S_\phi\big(\mathrm{Clamp}(\widehat{\mathbf{x}}_0)\big) \Big]. \tag{14}$$

The measurement-alignment loss is computed in standardized target space:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{8} \sum_{d=1}^{8} \mathrm{SmoothL1}\big(\widetilde{\mathcal{M}}_d(\widehat{\mathbf{x}}_{224}),\, \widetilde{c}_d\big). \tag{15}$$

Although $\mathcal{M}$ is frozen, its computational graph remains differentiable, so gradients from $\mathcal{L}_{\mathrm{align}}$ propagate through $\mathcal{M} \to S_\phi \to \widehat{\mathbf{x}}_0 \to G_\theta$, with no parameter updates to $\mathcal{M}$. Restricting alignment to low-noise timesteps yields predicted clean images with sufficient morphological structure for meaningful measurement and avoids unstable gradients from heavily corrupted samples.

4 Experiments

4.1 Setup

Data. We use DESI Legacy Imaging Surveys RGB cutouts (Dey et al., 2019) paired with eight catalog measurements $\mathbf{c} = \{z, f_g, f_r, f_z, e_1, e_2, r_e, n\}$, cross-matched with Galaxy Zoo DESI (Walmsley et al., 2023) ($21{,}445$ pairs for generative/proxy training, $80/10/10$, seed 42). Retrieval experiments additionally use a $46{,}130$-image labeled catalog and a $3{,}128{,}494$-cutout survey gallery. Morphology catalogues provide observational context (Lintott et al., 2008; Willett et al., 2013; Dieleman et al., 2015; Domínguez Sánchez et al., 2018).

Protocols. **Baseline**: independently trained diffusion + SwinIR. **Joint**/MA: jointly fine-tuned under frozen $\mathcal{M}$. Closed-loop comparisons use $N{=}100$ redshift-stratified conditions (de novo and img2img, noise strength 0.5). Control tests use 32 mid-density anchors (330 generations). Known rares are the top 1% by kNN tail score s_{tail}. For each query we generate M samples, reject those failing proxy discrepancy/uncertainty thresholds, and retrieve with fixed encoder $\mathcal{F}$ (DINOv3 ViT-L/16). Ranked neighbors are evidence, not automatic labels. Sampling uses classifier-free guidance ($s{=}6$; train cond-drop 0.1) and EMA (0.995).

4.2 Proxy Reliability

On $N{=}2{,}145$ held-out real images, $\mathcal{M}$ reaches mean $R^2{=}0.925$ (Table 1; per-parameter 0.777–0.980). MC Dropout barely changes estimates and correlates weakly with error (Pearson/Spearman $0.092/0.120$), so uncertainty is a filter, not calibrated coverage. Performance is stable in redshift/flux/r_e/SNR tails; the extreme Sérsic axis fails (R^2: $0.919 \to -2.56$ from central 90% to outer 1%). We treat extreme-n hypotheses as unsupported by the proxy.

Table 1: Frozen proxy on real HR test images ($N{=}2{,}145$). Full per-parameter and MC columns in appendix.

	MAE	RMSE	R^2
z / f_g / f_r / f_z	0.011 / 7.11 / 15.0 / 28.3	0.014 / 11.0 / 22.6 / 41.9	0.777 / 0.975 / 0.974 / 0.972
e_1 / e_2 / r_e / n	0.020 / 0.022 / 0.377 / 0.484	0.029 / 0.032 / 0.554 / 0.651	0.980 / 0.975 / 0.885 / 0.859
Mean	6.41	9.60	0.925

4.3 Specificity and Counterfactual Faithfulness of the Control Space

We next ask whether the eight-dimensional condition vector provides a faithful semantic interface to the generator. All images in this subsection are produced by the Measurement-Aligned model and re-measured by frozen $\mathcal{M}$.

Leave-group necessity and specificity. For each physical group $G \in \{\mathrm{brightness}, \mathrm{color}, \mathrm{shape}, \mathrm{scale}\}$, we compare the full-condition baseline against two ablations that neutralize G: replacing the corresponding coordinates by catalog medians, or shuffling them across the batch. The full baseline achieves a mean standardized error of $\mathrm{MAE}_z{=}0.156$. As shown in Table 2, ablating a group primarily increases error on that specific group, while leaving the held-out axes largely unaffected. The strongest effects appear for shape ($\Delta\mathrm{MAE}_z{=}+0.770$ median / $+1.166$ shuffle) and scale ($+0.668$ / $+0.656$), whereas changes on held-out axes remain negligible ($\lesssim 0.04$). This strict decoupling argues against shortcut leakage: if the model were using brightness as a proxy to imitate shape changes, we would expect to see correlated error increases on held-out groups, which is not observed. Color exhibits the weakest leave-group response ($+0.026$–0.053), consistent with the entanglement of band ratios and overall flux in RGB cutouts.

Table 2: Leave-group ablation on mid-density anchors ($N{=}32$ per cell). Δ_{group} is the increase in group-wise MAE_z relative to the full condition; Δ_{held} is the corresponding change on complementary axes.

Group	Ablation	$\mathrm{MAE}_z^{\mathrm{group}}$	Δ_{group}	Δ_{held}	N
Full baseline (mean $\mathrm{MAE}_z{=}0.156$*)*		—	—	—	32
brightness	median	0.471	**+0.397**	+0.102	32
brightness	shuffle	0.547	**+0.473**	+0.120	32
color	median	0.105	+0.026	+0.013	32
color	shuffle	0.133	+0.053	+0.020	32
shape	median	0.931	**+0.770**	+0.035	32
shape	shuffle	1.326	**+1.166**	+0.030	32
scale	median	0.838	**+0.668**	+0.011	32
scale	shuffle	0.827	**+0.656**	+0.006	32

Counterfactual sufficiency. Holding non-target axes fixed, we sweep brightness (f_r), color ratios (f_g/f_r, f_z/f_r), effective radius r_e, Sérsic index n, and ellipticity $|e|$ along seven-step paths in well-supported regions of condition space. Proxy-recovered measurements track the requested controls with Spearman correlations of at least 0.96 on all six paths (Pearson 0.941–0.993; Table 3). Qualitative grids further show continuous and physically readable changes in photometry and morphology (Fig. 2). Together, these results support treating the condition vector as a behaviorally interpretable control space, while also exposing color as a partially entangled axis of the RGB interface.

Table 3: Counterfactual sweeps (7 steps per path; non-swept axes held fixed). Spearman and Pearson correlate the swept input with $\hat{c} = \mathcal{M}(x)$.

Path	Sweep	N	Spearman	Pearson	MAE_z
brightness	f_r	7	**1.000**	0.993	0.177
color g/r	f_g/f_r	7	0.964	0.952	0.258
color z/r	f_z/f_r	7	**1.000**	0.946	0.223
scale	r_e	7	**1.000**	0.990	0.161
Sérsic	n	7	0.964	0.941	0.187
ellipticity	$\|e\|$	7	0.964	0.967	0.162

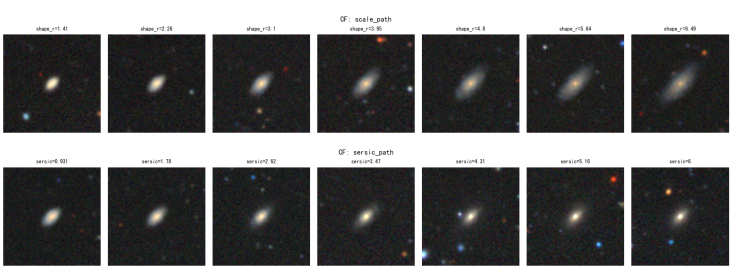


Figure 2: Qualitative counterfactual grids under Measurement-Aligned generation.

4.4 Measurement Alignment Improves Closed-Loop Faithfulness

A generated prototype is scientifically useful only insofar as it instantiates the condition used to define the hypothesis. We therefore compare Measurement Alignment with the independent Baseline under the same frozen proxy, using $N{=}100$ redshift-stratified catalog queries in both de novo and img2img modes.

Joint training reduces overall closed-loop MAE from 18.249 to 7.962 under noise-conditioned sampling (56.4% relative improvement) and from 17.839 to 7.415 under img2img (58.4%; Table 4). Standardized errors and valid rates ($\mathrm{MAE}_z \leq 1$) improve in the same direction, indicating that the gains are not artifacts of raw flux scale. Improvements in both generation modes further suggest that the alignment objective is not merely a trick for pure noise sampling.

These closed-loop scores remain proxy-mediated: they measure consistency with $\mathcal{M}$, which is also used in the alignment loss. A generator could in principle overfit the proxy's errors. We therefore treat the retrieval and expert studies below as external tests on real survey pixels, scored with a separate encoder $\mathcal{F}$.

To check whether the advantage persists outside the bulk of the catalog distribution, we stratify queries by a kNN density score and evaluate matched Baseline and Joint generations (40 queries per slice). On the extreme stratum, closed-loop MAE_z decreases from 0.488 to 0.269 ($\Delta{=}-0.218$); the same improvement appears at the ge_q99 cut, and valid rates remain 1.0 throughout. These results confirm that Measurement Alignment improves faithfulness consistently, both in-distribution and in objectively defined low-density regions.

Table 4: Closed-loop comparison of Baseline and Measurement-Aligned Joint models on matched catalog queries (N=100). Negative Δ favors Joint.

Mode	$\mathrm{MAE_B}$	$\mathrm{MAE_J}$	ΔMAE	Imp.	$\mathrm{MAE}_{z,\mathrm{B}}$	$\mathrm{MAE}_{z,\mathrm{J}}$	$\Delta\mathrm{MAE}_z$	$\mathrm{Valid_{B/J}}$
noise	18.249	7.962	-10.287	56.4%	0.353	0.186	-0.167	0.995/1.000
img2img	17.839	7.415	-10.424	58.4%	0.356	0.182	-0.174	0.995/1.000

4.5 From Interpretable Hypotheses to Survey Retrieval

We finally test whether an interpreted low-density condition can be converted into an empirically useful survey query.

Leave-one-out retrieval of known rares. Known rares are defined automatically as the top 1% of catalog objects by s_{tail} (N=462). For each rare condition $c^\star$ we generate M=4 Measurement-Aligned prototypes, retain those with measurement discrepancy $d_{\mathrm{meas}} \leq 1.0$ (accept rate 99.7%; 461 rares keep at least one accepted sample), and encode them with a frozen DINOv3 ViT-L/16 encoder. The query embedding is the L2-normalized mean of accepted prototype features. Under a leave-one-out protocol over the $46{,}130$-image catalog, success requires recovering the exact left-out rare.

Measurement-Aligned prototypes achieve Recall@10/50/100/500 of $0.121/0.245/0.323/0.588$ with median rank 283, substantially outperforming a condition-space Catalog-NN baseline (Recall@500 0.052, median rank 5101; Table 5, Fig. 3). Catalog-NN is intentionally a condition-space reference rather than an image foundation-model competitor: it asks whether visual prototypes contribute information beyond nearest neighbors in measurement space. Even when exact identity recovery fails, Top-10 visual neighbors remain physically closer to $c^\star$ than random catalog images (mean absolute z-scored distance 0.877 vs. 1.778), and the Top-1 neighbor has a median 8-D condition cosine of 0.880 (random mean -0.061).

Table 5: Leave-one-out rare retrieval and survey-scale expert review.

Method	R@50	R@500	Med. rank	N
Catalog-NN	0.000	0.052	5101	462
MA prototype + DINOv3	0.245	0.588	283	461

Survey-scale screening and expert validation. We next ask whether the same prototypes can screen a million-object gallery and yield neighbors that human experts judge as visually similar to the model's explanation. Rarity itself is catalog- and threshold-dependent; we therefore evaluate visual agreement with the prototype rather than an ambiguous rarity label. Accepted prototypes are re-encoded with the survey retrieval encoder (d=512) and queried against a FAISS IVFFlat index over $3{,}128{,}494$ cutouts. For Q=461 queries and Top-K=50, approximate search completes in 0.73 s (≈ 631 queries per second) after a one-time index load of 7.5 s; 99.3% of Top-50 neighbors exceed cosine similarity 0.55, with mean Top-1 similarity 0.754 (Table 5).

Two experts then rate 100 randomly sampled Top-50 prototype–neighbor pairs as `similar`, `partial`, `dissimilar`, or `artifact`, answering whether a similar source was recovered. Pooling 200 ratings, 79.5% are labeled similar and 88.5% fall in $\{\texttt{similar}, \texttt{partial}\}$, while 11.5% are dissimilar and none are artifacts. Exact agreement is 81% (Cohen's $\kappa \approx 0.46$); soft positive/negative agreement rises to 91%. These results support retrieval consistency at survey scale, but they do not establish the astrophysical novelty of every returned candidate.

5 Conclusion

We presented a measurement-aligned generative framework that converts interpretable physical conditions into testable queries for galaxy discovery. Catalog measurements provide semantic grounding; leave-group ablations and counterfactual paths test whether those controls are used

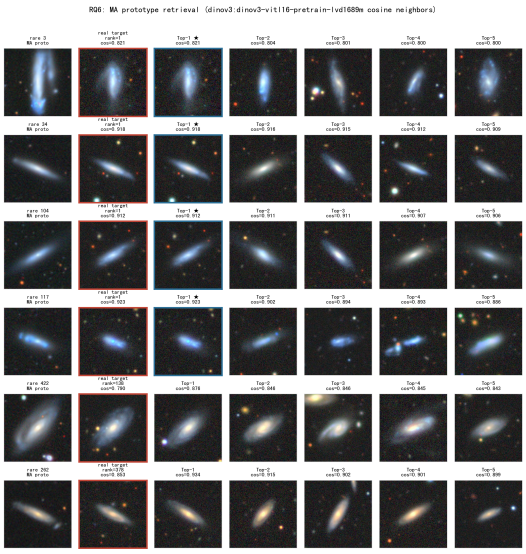


Figure 3: Prototype-based retrieval examples. Each row shows a Measurement-Aligned prototype, the held-out real target, and retrieved neighbors. Both successful recoveries and failure cases are retained.

specifically and faithfully; and measurement-consistent prototypes turn the interpreted condition into an external retrieval experiment on real survey data.

Empirically, the frozen proxy reaches a mean R^2 of 0.925 on held-out real images while exposing a clear limitation in extreme Sérsic tails. Shape and scale controls are highly specific, and six counterfactual paths are tracked with Spearman correlations of at least 0.96.Measurement Alignment reduces closed-loop error by 56.4% in noise-conditioned generation and 58.4% in img2img generation. In leave-one-out retrieval, Measurement-Aligned prototypes achieve Recall@500 of 0.588, compared with 0.052 for a condition-space baseline. When scaled to more than three million survey cutouts, blinded experts label 79.5% of sampled prototype–neighbor pairs as similar and 88.5% as similar or partially similar.

Our claim is intentionally narrower than complete mechanistic interpretability of the diffusion network. Semantic inputs do not by themselves prove faithful internal use, and proxy agreement does not establish physical truth. The main limitations arise in low-support regions, where the proxy degrades, color is entangled with flux, and visual similarity does not imply astrophysical novelty. Future work should complement the learned proxy with official survey measurement pipelines and pursue object-level follow-up for coherent candidate populations.

6 AI USE STATEMENT

Generative AI tools assisted with language polishing and code development. All AI-generated content was manually reviewed and validated by the authors, who take full responsibility for the final manuscript. No AI was used to fabricate data or results.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Arjun Dey, David J. Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, et al. Overview of the desi legacy imaging surveys. *The Astronomical Journal*, 157(5):168, 2019. doi: 10.3847/1538-3881/ab089d.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Sander Dieleman, Kyle W. Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015. doi: 10.1093/mnras/stv632.

Helena Domínguez Sánchez, Marc Huertas-Company, Mariangela Bernardi, Domenico Tuccillo, and Joseph L. Fischer. Improving galaxy morphologies for sdss with deep learning. *Monthly Notices of the Royal Astronomical Society*, 476(3):3661–3676, 2018. doi: 10.1093/mnras/sty338.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2376–2384, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020.

François Lanusse, Rachel Mandelbaum, Siamak Ravanbakhsh, Chun-Liang Li, Peter Freeman, and Barnabás Póczos. Deep generative models for galaxy image simulations. *Monthly Notices of the Royal Astronomical Society*, 504(4):5543–5555, 2021. doi: 10.1093/mnras/stab1214.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1833–1844, 2021.

Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, et al. Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 2008. doi: 10.1111/j.1365-2966.2008.13689.x.

Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018. doi: 10.1145/3236386.3241340.

Chenrui Ma, Zechang Sun, Tao Jing, Zheng Cai, Yuan-Sen Ting, Song Huang, and Mingyu Li. Can AI dream of unseen galaxies? conditional diffusion model for galaxy morphology augmentation. *The Astrophysical Journal Supplement Series*, 2025. GalaxySD; arXiv:2506.16233.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020. doi: 10.1109/ACCESS.2020.2976199.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

José Luis Sérsic. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletín de la Asociación Argentina de Astronomía*, 6:41–43, 1963.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Michael J. Smith, James E. Geach, Ryan A. Jackson, Nikhil Arora, Connor Stone, and Stéphane Courteau. Realistic galaxy image simulation via score-based generative models. *Monthly Notices of the Royal Astronomical Society*, 511(2):1808–1818, 2022. doi: 10.1093/mnras/stac130.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018.

Mike Walmsley, Tobias Géron, Sandor Kruk, Anna M. M. Scaife, Chris Lintott, Karen L. Masters, James M. Dawson, Hugh Dickinson, Lucy Fortson, Izzy L. Garland, et al. Galaxy zoo desi: Detailed morphology measurements for 8.7m galaxies in the desi legacy imaging surveys. *Monthly Notices of the Royal Astronomical Society*, 526(3):4768–4786, 2023. doi: 10.1093/mnras/stad2919.

Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, Karen L. Masters, Brooke D. Simmons, Kevin R. V. Casteels, Edward M. Edmondson, Lucy F. Fortson, Sugata Kaviraj, William C. Keel, et al. Galaxy zoo 2: Detailed morphological classifications for 304,122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013. doi: 10.1093/mnras/stt1458.

Mikaeel Yunus, John Wu, Timothy Heckman, and Benne Holwerda. Conditional diffusion models for generating images of SDSS-like galaxies. In *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2024.