Learning What's Real: Disentangling Signals from Measurement Artifacts in Multi-Sensor Data, with Applications to Astrophysics

Pablo Mercader-Perez[1,*] **Carolina Cuesta-Lazaro**[2,3] **Daniel Muthukrishna**[1,4]
Jeroen Audenaert[1] **V. Ashley Villar**[4,5] **David W. Hogg**[2,3] **Marc Huertas-Company**[6]
William T. Freeman[1]

[1]Massachusetts Institute of Technology [2]New York University
[3]Flatiron Institute, Simons Foundation [4]Center for Astrophysics | Harvard & Smithsonian
[5]NSF AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI)
[6]Instituto de Astrofísica de Canarias

Abstract

Data collected from the physical world is always a combination of multiple sources: an intrinsic signal from the physical process of interest and a signal from measurement-dependent artifacts introduced by the sensor or instrument. This secondary signal acts as a confounding factor, limiting our ability to extract information about the underlying physics. Moreover, it poses significant challenges for combining data in heterogeneous or multi-instrument settings. To disentangle these factors of variation, we propose a dual-encoder architecture with a counterfactual generation objective that leverages overlapping observations. The resulting representations explicitly separate intrinsic signals from sensor-specific distortions and noise, and can be used for counterfactual view generation, parameter inference, and instrument-independent similarity search—all unconfounded by measurement artifacts. We demonstrate the effectiveness of our approach in a multi-instrument setting on astrophysical galaxy images from the DESI Legacy Imaging Survey and the Hyper Suprime-Cam Survey. This framework provides a general recipe for scientific self-supervised pretraining: construct training pairs from overlapping observations of the same physical system, treat sensor- or modality-specific effects as augmentations, and learn invariant representations through counterfactual generation.

1 Introduction

Observational data collected from the physical world can be viewed as the result of a causal process. First, an event of interest occurs, generating a physical signal. For example, a distant pulsating star emits electromagnetic waves, a person speaking creates pressure waves in the air, and a beating heart produces pressure waves that propagate through blood vessels. This signal cannot be directly observed. Instead, it must pass through a sensor or instrument that records it. Any instrument used introduces measurement artifacts, often referred to as *instrument systematics* or bias. These effects include transformations like the nonlinearity of Charge-Coupled Devices (CCDs) in a camera or a microphone's frequency response. These act as *confounding* factors. However, with sufficient understanding of the measurement system, they can often be partially modeled and disentangled from the underlying signal. Additionally, stochastic noise imposes a limit on measurement precision.

In other words, $\textit{Observation} = f(\textit{Signal}, \textit{Instrument}) + \textit{Noise}$. In some cases, f is relatively simple, such as a point-spread function (PSF) convolution that only requires accurate characterization of

*Correspondence: pablomer@mit.edu. Code: https://github.com/pablomerc/galaxy-counter.

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

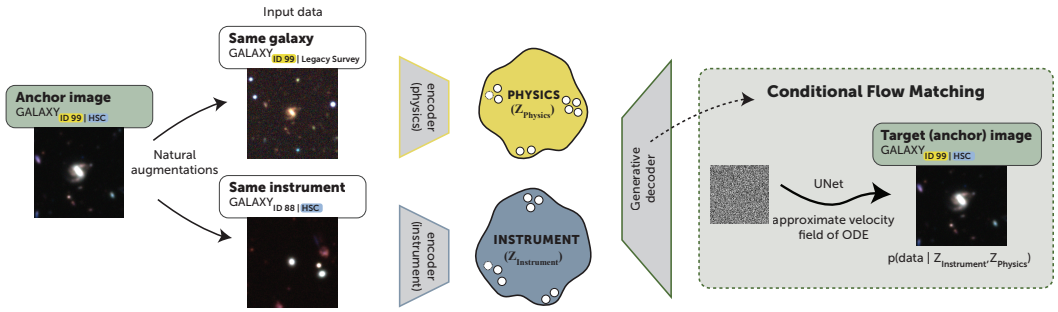


Figure 1: **Counterfactual Reconstruction.** The model learns to disentangle intrinsic galaxy properties from instrument systematics by reconstructing an anchor image. Training uses **data triplets**: an **anchor observation** (signal source s from instrument i), an instrument-augmented observation (**same source** s, a different instrument i'), and a physics-augmented observation (a different source s', **same instrument** i). The *physics* encoder extracts latents from the instrument-augmented pair to capture invariant physical features, while the *instrument* encoder processes the physics-augmented observation to capture instrument-specific effects. These latent variables jointly condition the flow-matching decoder. Crucially, the anchor image is never fed into the encoders; it is used only as a target for the flow-matching loss, forcing the model to generate the reconstruction counterfactually.

the optics. However, it often cannot be modeled fully analytically. The true signal is explained by variables that are independent of the measurement and contain the scientifically relevant information; we refer to these as **physics variables**. The instrumental effects arise from measurement distortions and act as confounders; we refer to these as **instrument variables**.

We present a general approach to disentangle physics and instrument variables when the same system is observed by multiple instruments. The key insight is that cross-matched observations form natural pairs of *same physics, different instrument*, providing a training signal for separating the two sets of factors without requiring an explicit forward model of either instrument.

We demonstrate our approach on astrophysical galaxy images, a domain where modeling instrument systematics is critical. Photons emitted by a distant source must traverse cosmic dust and the Earth's atmosphere before being recorded by ground-based telescopes, each with its own optics, detector characteristics, and observing conditions. As a result, the same galaxy can look remarkably different across two telescopes due to differences in resolution, noise, and calibration. Historically, obtaining observations of the same phenomena with multiple instruments has been essential for distinguishing real physical signals from instrumental artifacts. In this work, we train our model on $\sim$100,000 cross-matched galaxy images from two different ground-based telescopes and show that we learn disentangled representations separating the intrinsic properties of galaxies from the instrument-specific effects, enabling robust inference of galaxy properties, outlier detection, counterfactual generation across instruments, and instrument-independent similarity search.

2 Related work

Disentangled representation learning. Learning representations that separate independent factors of variation is a longstanding goal in representation learning. Tenenbaum and Freeman [1] explored bilinear representations to separate style and content. Early work on variational autoencoders (VAEs), such as β-VAE [2], encouraged disentanglement by regularizing the latent space via the loss function. A natural source of structure is multi-view observations, where the same underlying content is observed under varying conditions or at different times. From a causal perspective, Schölkopf et al. [3] framed disentanglement as the recovery of independent causal mechanisms, connecting representation learning and causal inference. Our framework draws on these ideas but enforces structural disentanglement through a specific architecture-driven information bottleneck and a counterfactual generation objective, rather than relying on hand-engineered loss functions.

Foundation models in the sciences. Pre-training self-supervised representations on large scientific datasets has recently seen success in different areas of science like biology [e.g., 4, 5] and satellite imagery [e.g., 6]. In astrophysics, Walmsley et al. [7] demonstrated the value of self-supervised

representations for downstream tasks. More recently, Parker et al. [8] developed AION-1, a large-scale foundation model trained via masked modeling across modalities such as imaging and spectra from the Multimodal Universe dataset [9]. However, a limitation common to these approaches is that instrument-specific effects are either ignored or treated as independent modalities without explicit disentanglement. This has practical consequences for downstream tasks. For instance, a model developed for ESA's Euclid mission [10] found that instrument systematics, rather than physically interesting sources, dominated the outliers identified in the latent space. Our work addresses this by building the separation of physics and instrument factors into the architecture itself.

Incorporating multi-instrument observations in foundation models. Recently in astrophysics, Audenaert et al. [11] leveraged the causal representation ideas from Schölkopf et al. [12, 3], Hattori et al. [13], to causally separate astrophysical observations into a physics and instrument component by using a dual-encoder architecture. They used a dataset of simulated astronomical time series captured by multiple instruments to learn a *physics latent space* that encodes information about the target star and an *instrument latent space* that encodes information about the measurement configuration. This is achieved by leveraging triplets of overlapping observations and using contrastive learning. While effective, contrastive objectives have inherent limitations for our setting. By aligning representations across views of the same source, they tend to retain only the information shared across all views, effectively collapsing toward the lowest-resolution observation in the set and discarding fine-grained features that are visible in higher-quality instruments. Also importantly, contrastive embeddings are not optimized for pixel-level reconstruction and producing counterfactual views would require training a separate generative model on the learned embeddings, with no guarantee that the necessary information is preserved. Our generative objective addresses these limitations.

3 Methods

We present a counterfactual generation objective designed to learn representations that disentangle physical properties from measurement systematics. Our goal is to capture robust physical information into a physics latent space, while learning a data-driven noise model for each of the instruments.

Counterfactual generative objective. Our framework leverages large-scale scientific datasets with overlapping observations from multiple instruments. In this setting, natural augmentations can be used to construct triplets: an anchor, an observation of the same source (but different instrument), and an observation on the same instrument (but different source), as shown in Fig. 1.

A *physics encoder* receives a sample from the target source but from a different instrument, so it is encouraged to reduce the loss by capturing the underlying physical properties of the source and becoming invariant to the instrumental distortions. Similarly, an *instrument encoder* receives a sample from the target instrument but a different source, so it is encouraged to capture instrument-specific distortions while ignoring the content specific to the source. The decoder combines these two latent embeddings to reconstruct an unseen target or anchor observation.

Formally, consider a set of N sources observed across M instruments, where $x_{j,k}$ denotes the observation of source j as captured by instrument k. We aim to learn the conditional distribution $p\big(x_{j,k} \,\big|\, \{z_{\text{phy}}(x_{j,k'})\}_{k'\neq k},\, \{z_{\text{ins}}(x_{j',k})\}_{j'\neq j}\big)$, for all source–instrument pairs (j,k). Here $\{z_{\text{phy}}(x_{j,k'})\}_{k'\neq k}$ are the physics embeddings extracted from observations of the *same source* j by *other instruments* $k' \neq k$, which provide information about the source's intrinsic physical properties. The set $\{z_{\text{ins}}(x_{j',k})\}_{j'\neq j}$ are the instrument embeddings extracted from observations of *other sources* j' by the *same instrument* k. This formulation allows the model to leverage all available multi-instrument data: the physics of the target source is informed by its appearances across different instruments, while the instrument response is informed by how other sources appear under similar observing conditions. We concatenate the physics and instrument embedding tokens into a variable-length sequence for decoder cross-attention, allowing the number of source and instrument neighbors to vary (see the "Model implementation" paragraph in Section 4 and Appendix A.1).

We model this conditional distribution using flow matching [14], learning a velocity field $u_\theta(x, t, z_{\text{phy}}, z_{\text{ins}})$ that transports samples from Gaussian base distribution to the target distribution. The training loss is given by the error on the predicted velocity field,

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ j\sim[N],k\sim[M] \\ x_{j,k}\sim p_{\text{data}} \\ x_0\sim\mathcal{N}(0,I)}} \left[\left\| u_\theta\left(x_t, t, \{z_{\text{phy}}(x_{j,k'})\}_{k'\neq k}, \{z_{\text{ins}}(x_{j',k})\}_{j'\neq j}\right) - (x_{j,k} - x_0) \right\|^2 \right] \tag{1}$$

where $x_t = (1-t)x_0 + tx_{j,k}$. That is, we sample a source j, a target instrument k, a noise sample x_0, and a timestep t, and train the velocity field to reconstruct the observation $x_{j,k}$ from a noisy state x_t conditioned on physics latents from other instruments and instrument latents from different sources.

Once trained, the physics embedding supports downstream tasks such as outlier detection, parameter inference, classification, or embedding-based retrieval, while the instrument space allows to find similar systematics and observing conditions. The generative decoder enables generating counterfactual views. In astrophysics, a particularly promising use case is to inform follow-up observations by predicting high-resolution observations from expensive facilities using low resolution observations from a different instrument.

4 Application to Astrophysics: Galaxy foundation models

Datasets in astrophysics have grown considerably over the last decade. In particular, galaxy surveys such as DESI, have now imaged over a billion galaxies while taking spectra of over 10 million galaxies. This has sparked interest in developing foundation models for galaxies that learn semantic representations of astronomical data, helping us search large databases [15], identify interesting anomalies, and predict galaxy properties of interest. Up to now, these models have treated data from different instruments as independent modalities [8]. Here, we show that counterfactual reconstruction improves robustness to instrument systematics while enabling new foundational capabilities, including cross-instrument generation, instrument-independent retrieval, and data-driven denoising.

We take a cross-matched dataset from the Multimodal Universe that includes ∼100,000 cross-matched galaxy images from the DESI Legacy Imaging Surveys [Legacy, 16] and the Hyper Suprime-Cam [HSC, 17] survey as a representative multi-instrument setting. HSC is a deeper and higher-resolution instrument that can detect significantly fainter features that would otherwise be masked by noise in Legacy. However, by making this compromise, Legacy covers a larger area of the sky, $\approx 20{,}000\,\mathrm{deg}^2$, compared to HSC, $\approx 1{,}200\,\mathrm{deg}^2$, resulting in a trade-off between coverage and image quality (both photometric depth and angular resolution).

Data preprocessing. We have two sets of heterogeneous flux measurements: HSC provides five color filters with discrete wavelength bands (g, r, i, z, y), and the Legacy Survey uses four (g, r, i, z). Despite the shared designations, they differ significantly in their central wavelengths, pixel scales, zero-points, and noise characteristics. Our preprocessing pipeline follows the strategy established by [8], with one key difference: rather than treating all nine bands as distinct channels, we align the overlapping (g, r, i, z) filters across both surveys into a unified four-channel representation and treat both sources as the same modality. For this purpose, we discard the additional HSC y-band channel, but future work could incorporate it.

Following [8], we normalize the zero-points by rescaling HSC measurements to the Legacy standard of 22.5 mag using the relation: $s = 10^{(ZP-22.5)/2.5}$. Then, we apply an arcsinh normalization to handle the images' large dynamic range. Finally, we crop the HSC images into 48×48 pixel cutouts. To account for the differing pixel scales between surveys ($0.168''$ for HSC vs. $0.262''$ for Legacy), we extract the inner 31×31 pixels of the Legacy images and upsample them to 48×48 pixels via linear interpolation; this ensures spatial alignment and a consistent field of view across pairs.

Data triplets. In this dual-survey setting, each data point has a unique galaxy pair, providing the cross-survey view supplied to the physics encoder. For the instrument views, we condition on up to five images of different galaxies observed by the same survey as the anchor. We select these as the five nearest spatial neighbors in the given survey within an angular separation of 3 arcminutes. This choice is motivated by the fact that instrumental systematics - such as PSF size and depth - vary smoothly across the sky, so spatially nearby observations share similar instrument conditions with the anchor.

Model implementation. We use a modified ResNet-18 [18] for both encoders (physics and instrument) as described in Appendix A.1. We map each input image to a sequence of 4 spatial tokens of 16 dimensions, i.e. a latent representation of shape $(B, 4, 16)$. To accommodate a variable number of conditioning examples, we treat these embeddings as a set of tokens. For a given input set, the tokens from all images are concatenated into a single sequence of variable length. Then, we use cross-attention layers within our velocity network, a `UNet2DConditionModel` [19]. This allows the model to leverage an arbitrary number of context image embeddings during the generative process.

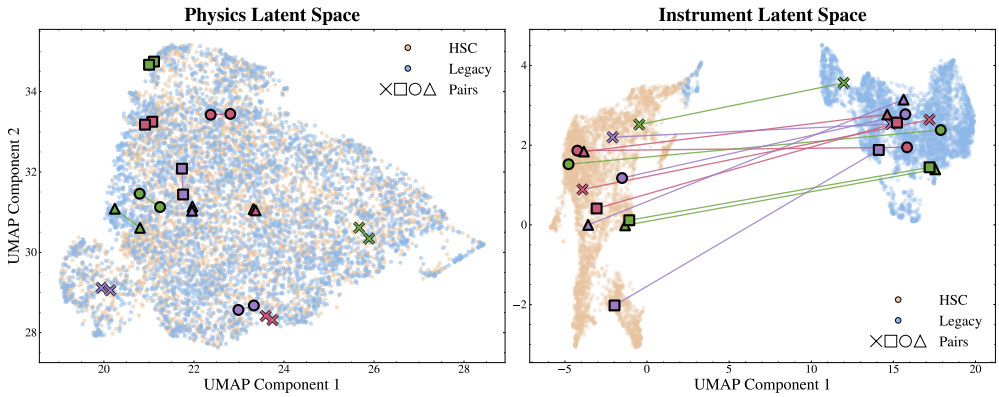


Figure 2: **Disentanglement.** UMAP projections of the physics (left) and instrument (right) latent spaces. Orange and blue points represent HSC and Legacy images, respectively. Matched markers ($\triangle$, $\times$, $\square$, $\circ$) denote cross-survey pairs. In the *physics space*, the encoders produce overlapping distributions where cross-survey pairs are mapped to similar coordinates. In contrast, the *instrument space* shows a clean separation between surveys and pairs, confirming that this space captures survey-specific features rather than intrinsic galaxy properties. A complementary quantitative evaluation of downstream R^2 scores for physical and instrumental properties is shown in Fig. 4.

4.1 Disentangled representations

First, we demonstrate that the physics and instrument latent spaces do show the expected disentanglement. We encode 8,192 image pairs from HSC and Legacy Surveys using both the physics and instrument encoders, and visualize the resulting latent embeddings using a 2D UMAP projection [20] in Fig. 2. The latent space demonstrates a clear separation of information. In the instrument space, HSC and Legacy observations form two distinct clusters, confirming that the encoder successfully captures survey-specific characteristics. In the physics space, the two surveys have largely overlapping distributions, indicating that the encoder has learned an approximate domain-invariant representation of the underlying galaxy properties. The overlap is, however, not expected to be perfect since the two surveys provide different amounts of constraining power on the underlying physical properties. The emergence of alignment of pairs observed in Fig. 2 from the reconstruction objective alone confirms that the architectural bottleneck is sufficient to drive the convergence of physics representations across the two surveys.

Outlier detection unconfounded by instrument systematics. Naively learned representations often bias outlier detection toward rare instrument artifacts rather than physically interesting sources. Our disentangled representation addresses this by separating physics and instrument factors, enabling outlier detection separately in each latent space: rare astrophysical sources in the physics space and rare observing conditions in the instrument space.

We encode all $\sim$100,000 HSC galaxies using the physics encoder and the instrument encoder separately. As a baseline, we encode the same images with AION-1 [8], a foundation model that treats each survey as an independent modality without explicit disentanglement. For each of the three latent spaces, we fit a normalizing flow to the empirical distribution of embeddings and rank all sources by their estimated log-likelihood (Appendix A.7). The lowest-likelihood examples are the candidate outliers.

Fig. 3 shows the top twelve outliers from each latent space. The AION-1 anomalies contain several instrumental pathologies: saturated pixels from bright stars, vertical rainbow bands, and high-noise patterns. Outliers from our instrument space show similar characteristics, confirming that the instrument encoder captures systematics-related variation. In contrast, anomalies from our physics space are visibly free of these artifacts and reveal morphologically unusual sources: galaxy mergers, disturbed systems, and other candidate rare objects. Notably, galaxies sampled near the median anomaly score appear qualitatively similar across the three latent spaces, while their highest-ranked outliers are visibly different. We show cutouts at the full 160×160 pixel size.

Parameter inference robust to instrument systematics. To evaluate the information content and disentanglement of our latent spaces, we train a small MLP (Appendix A.3) to predict physical and

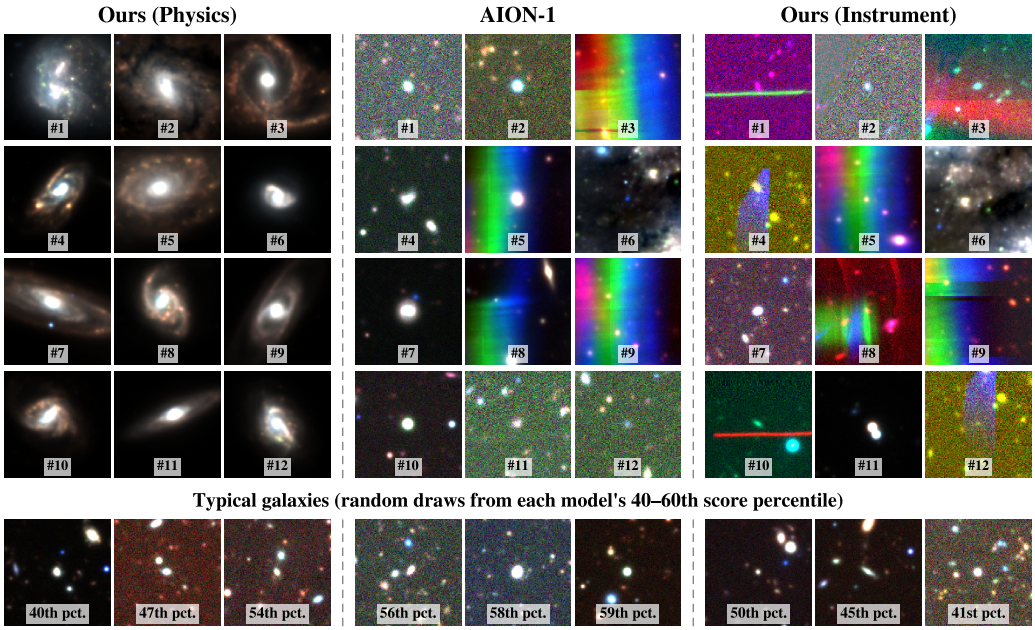


Figure 3: **Outlier Detection.** Top twelve outliers from each of the three latent spaces. **Left:** our physics latent reveals morphologically unusual sources (e.g., galaxy mergers), free of artifacts. **Middle & Right:** anomalies from AION-1 and our instrument latent are dominated by instrumental pathologies (saturated pixels, bands, noise patterns), confirming that our representation separates physical anomalies from instrument systematics. **Bottom row:** for reference, three galaxies drawn uniformly at random from the 40–60th percentile of each model's anomaly score. Typical galaxies look alike in all three spaces.

instrumental properties from embeddings extracted by each encoder on Legacy and HSC images. For the physics latent space we flatten across the token dimension; for the instrument latent space we apply mean pooling. In both cases, we concatenate the HSC and Legacy embeddings before feeding them into the MLP. We compare against AION-1 (Base) embeddings and against a frozen, randomly initialized ResNet-18 with the same architecture as our encoders. Fig. 4 shows the resulting regression R^2 per property.

The physical properties of galaxies were inferred by Hahn et al. [21] from spectroscopic data, not from the images themselves, so they are only partially recoverable from imaging alone. Moreover, we note that these are not "physics only" properties. For instance, the galaxy's redshift determines its apparent brightness, angular size, and observed colors—properties that are also changed by instrument conditions such as depth and PSF. Morphology is also affected by the instrument's PSF. Therefore, the instrument encoder will inevitably retain some of this information. On the other hand, observing conditions, such as PSF size and depth, vary across each survey's footprint as a function of sky position, determined by the tiling strategy and the atmospheric conditions during each field's observation. Therefore, the physics latent space may retain some of this information. Details on the properties and source catalogs are provided in Appendix A.4.

The physics latent space recovers physical properties at a level comparable to AION-1. The randomly initialized ResNet serves as a reference for what an encoder of this architecture captures by default, purely from the inductive biases of the ResNet [22], without any training signal encouraging or discouraging it to encode specific information. Comparing our trained encoders to this baseline isolates what each latent space has actually been pushed to learn. Under this lens, the instrument latent space retains some information about physical properties, in particular redshift, stellar mass, and morphology. As explained above, this is expected to some degree, though we cannot quantify it. In the future, we will study how the capacity of the instrument bottleneck affects these results, especially given that the untrained ResNet exceeds our encoder's predictive power on some properties.

To further disentangle genuine leakage of instrument into physics from shared spatial structure, we train a ResNet to predict each survey's instrument properties directly from the other survey's image

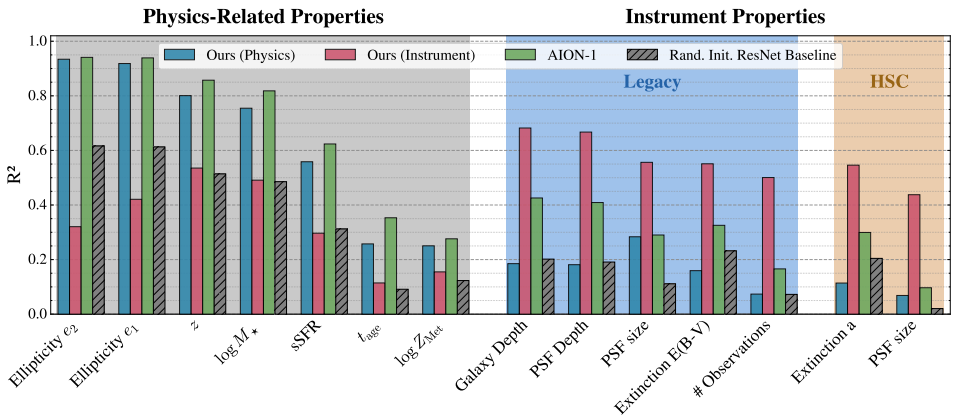


Figure 4: **Probing Latent Disentanglement via Downstream Regression.** We report R^2 scores for the prediction of physics-related properties (left) and instrumental properties (right) on four sets of representations: our physics embeddings (blue), our instrument embeddings (red), AION-1 embeddings which treat each survey as an independent modality (green), and a randomly initialized ResNet-18 of the same architecture as our encoders (hatched black). We distinguish physics-related properties and instrument properties, which are grouped by survey. The asymmetry between the two panels demonstrates that our architecture successfully separates physics properties from instrument-specific information.

("Cross-Predict ResNet"). This cross-prediction baseline measures the shared spatial component: any predictive power it achieves reflects information that is inherently common to both surveys and therefore accessible to either encoder by construction. We find that predictions from physics latents are always worse than the cross-predict baseline, verifying that the physics latent space is not capturing private instrument information. The full results are shown in Appendix A.5.

The key diagnostic is the asymmetry in the split of properties: the physics latents substantially outperform instrument latents at predicting (mostly) physical properties, whereas the pattern reverses for instrument properties. Notably, our latent spaces often contain less information than the baselines when used for cross-prediction (e.g. predicting physics-related properties from the instrument latents), indicating that the model has actively learned to erase instrument-specific information from the physics space and vice versa. On the other hand, we note that AION-1 embeddings retain sensitivity to instrumental properties, but our instrument latent space outperforms both baselines in predicting instrument-related properties. This highlights the primary advantage of our architecture: by explicitly disentangling the latent factors, we achieve a more efficient separation of physics and instrumental properties than monolithic foundation models.

Instrument-independent similarity search and rare-object discovery. We can also use the learned embeddings from our model to search for nearest neighbors (NNs) in the physics and instrument spaces. In astrophysics, we often search large datasets for rare objects, such as galaxy mergers or strong lenses. By performing an L_2 nearest-neighbor search over the encodings generated by our model for our dataset of $\sim$100k cross-matched galaxies, we demonstrate that querying the physics space retrieves physically similar objects from either survey, whereas querying the instrument space isolates objects with similar noise properties. Fig. 5 **(Top)** shows an example where an HSC image is used as the query. A mixture of HSC and Legacy images is found in the top neighbors in the physics space. In fact, the four examples shown could have been found among the top nine NNs, even if we used the Legacy image of that galaxy instead. This demonstrates the ability to perform *instrument-independent similarity search* in physics space.

To further demonstrate the effectiveness of our disentangled physics representation, we also query the physics latent space with objects from a rare class. Fig. 5 **(Bottom)** shows an example where the query is a known strong gravitational lens. The top eight nearest neighbors successfully retrieve multiple HSC images that exhibit strong morphological evidence to be considered candidate lenses. This confirms that the physics latent space effectively isolates and groups complex physical features, streamlining the discovery of rare astrophysical phenomena.

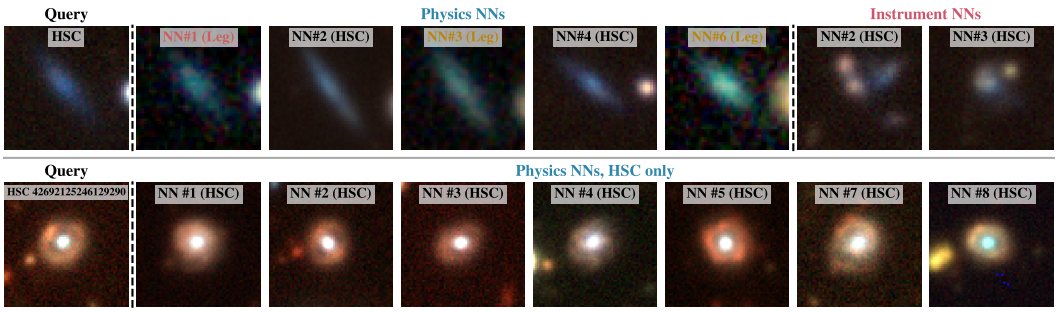


Figure 5: **Instrument-Invariant Search and Rare-Object Discovery. Top (Cross-Survey Retrieval):** We search the latent spaces using an HSC observation as query. *Physics Space:* The pair from the alternate survey is identified as the top-1 neighbor, with additional top-ranked neighbors showing physically similar galaxies from a mixture of both instruments. Notably, the four examples shown could have been found within the top nine nearest neighbors independent of which survey the query came from. *Instrument Space:* Neighbors in the instrument space show similar noise characteristics independent of the underlying galactic signal. **Bottom (Strong Lenses):** Nearest neighbor search within the physics latent space using a known strong lens as query, focusing only on the HSC images. The retrieved top neighbors show features suggestive of strong lensing, demonstrating the model's capacity for rare-object discovery.

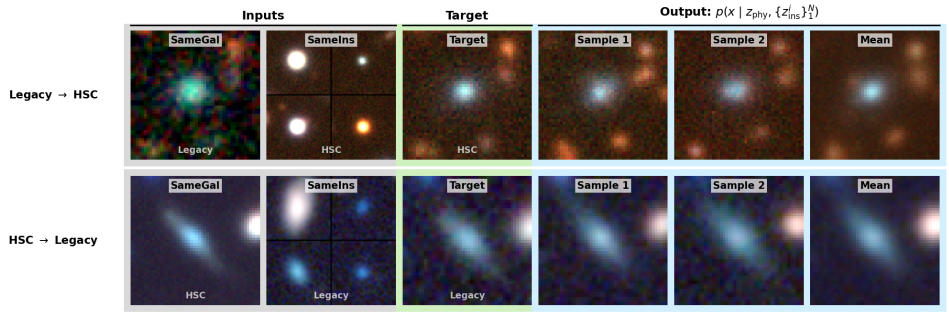


Figure 6: **Cross-instrument Galaxy Reconstructions. Columns 1–2 (Inputs):** The target galaxy observed via an alternate instrument (input to the physics encoder) and a set of up to five different galaxies imaged by the target instrument (input to the instrument encoder). **Column 3 (Ground Truth):** The original target image. **Columns 4–5 (Posterior Samples):** Independent samples generated by the decoder, demonstrating the model's ability to capture noise and uncertainty. **Column 6 (Posterior Mean):** The pixel-wise mean of the generated posterior samples.

4.2 Counterfactual cross-instrument generation.

Fig. 6 shows the model's ability to reconstruct target observations for both surveys and produce diverse posterior samples that capture the inherent uncertainty in the mapping. In addition to being our training objective, we can use counterfactual generation to prioritize follow-up observations. The discovery of rare objects, such as strong gravitational lenses, often depends critically on image quality. HSC's higher-resolution allows to resolve features, such as faint arcs, that are blurred or undetected in Legacy imaging. Our model can generate the expected HSC-like appearance for any galaxy in the Legacy footprint, effectively acting as a learned survey simulator that performs instrument-aware super-resolution and denoising. This could enable more targeted follow-up: one could first generate counterfactuals views to identify the most promising objects, those where the model predicts lensing features that are plausibly present but partly unresolved in the Legacy data. The actual follow-up observation then provides the real photons needed for confirmation.

More broadly, this approach could support population-level studies by extending HSC-like morphological analysis across the full $\sim$20,000 deg^2 Legacy footprint. However, we caution that the model generates the statistically expected appearance given the Legacy input, not a true observation. It can sharpen and denoise features that are marginally detected, but it cannot recover information that is entirely absent from the input. Features well below the Legacy detection threshold will not appear

in the counterfactual view. The generated images are best understood as informed predictions that reduce the search space for follow-up, rather than replacements for real, deeper observations. To evaluate the generated images, we report MSE computed on the generated images. The posterior samples achieve an MSE of 0.081 for HSC-anchored and 0.197 for Legacy-anchored reconstructions, evaluated over 256 held-out galaxies.

Zero-shot pipeline transfer. To further validate that our counterfactual HSC images preserve intrinsic galaxy morphology, we train a ResNet-18 model to predict galaxy ellipticities using 25,000 real HSC images, and apply this frozen model to counterfactual HSC images generated from Legacy observations of held-out galaxies. We found that the predictions match the ground truth with near-identical accuracy for both real ($R^2 = 0.82$) and generated ($R^2 = 0.81$) images. This demonstrates that counterfactually generated images can serve as drop-in replacements for real data, enabling the direct reuse of existing inference pipelines across surveys (Fig. 9, right; Appendix A.6).

Data-driven noise model. The instrument encoder and generative decoder together yield a data-driven noise model: by manipulating the instrument conditioning at inference time, we can simulate observations under different conditions while holding the underlying physics fixed. We demonstrate this in two complementary settings in Fig. 7. First, in a *controlled SNR traversal*, we reconstruct an HSC target galaxy from its Legacy counterpart five times, fixing the initial noise sample x_0 but varying the instrument conditioning across the 5th, 25th, 50th, 75th, and 95th percentiles of the HSC SNR distribution (defined as the mean pixel-wise flux divided by the flux error, averaged across channels). Traversing this axis produces progressively cleaner reconstructions, even surpassing the apparent SNR of the ground truth—indicating an exploitable, manipulable noise model rather than memorization. Second, we use this noise model to *remove strong systematics* from real observations: we pass an HSC image corrupted by severe instrumental distortion to the physics encoder and condition the instrument encoder on randomly selected HSC images to specify a baseline noise profile. Crucially, both encoders receive HSC images here (a combination never seen in training) yet the decoder reconstructs the underlying galaxy morphology while removing the artifacts, confirming that the model separates intrinsic physical signals from superimposed sensor distortions.

A complementary latent-swapping ablation supports this separation: swapping the physics latent makes reconstructions follow the donor galaxy, while swapping the instrument latent makes their pixel-scale noise statistics follow the donor's observing conditions (Appendix A.8, Table 7).

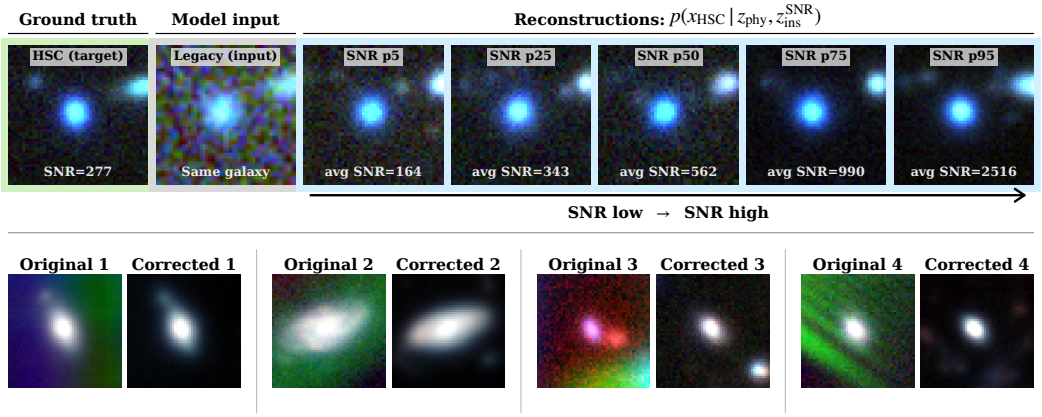


Figure 7: **Data-driven Noise Model. Top (Progressive Denoising):** Reconstructions of a target galaxy from its Legacy counterpart, varying the instrument conditioning across different percentiles of the HSC SNR distribution. The model successfully generates progressively cleaner images; the average SNR shown is that of the images used to condition the reconstruction. **Bottom (Systematic Artifact Removal):** By conditioning the physics encoder on an artifact-corrupted HSC image and the instrument encoder on a random set of HSC images, the model successfully removes severe instrumental distortions while preserving the underlying galaxy morphology.

5 Discussion

Comparison with contrastive learning. Contrastive learning across surveys encourages the encoding of information shared across observations of the same source. However, it does not explicitly encourage, and might even discourage, preserving fine detail visible only in a higher-resolution view. Such detail may nevertheless be scientifically informative. Instead, our method generates a softer pressure to align views and may help preserve more of that information. Moreover, contrastive learning does not natively support pixel-level reconstruction, while our counterfactual generative objective trains the representations to support pixel-level reconstruction.

Appendix A.9 compares reconstruction quality, downstream property prediction performance, and the predictive power difference between lower and higher-resolution observations under the two training objectives. When used for counterfactual reconstruction, a decoder conditioned on frozen contrastive embeddings obtains worse reconstruction quality (25.5% higher MSE) than our model, with a larger relative gap for the higher-resolution HSC Survey than for Legacy. When used for downstream prediction on physics-related properties from the physics latent space, our model outperforms contrastive learning in 6 out of 7 tasks. On the other hand, contrastive-learning produces instrument embeddings that achieve higher predictive power on all five evaluated instrument properties. Notably, the difference in predictive power between embeddings of the higher resolution and lower resolution surveys is greater with our model.

Training data requirements. To evaluate the data efficiency of our method, we vary the number of cross-matched training galaxies while keeping the instrument-neighbor pool fixed. With only 1,000 crossmatches (approximately 1% of the full training set), the model recovers 81% of the full-data improvement in mean R^2 for physics properties, but only 27–34% of the performance for instrument properties (Appendix A.10, Table 11). These results suggest that useful physics representations can be learned with limited cross-survey overlap, while learning an instrument model benefits from access to large scale crossmatched data.

Limitations and future directions. Our approach currently requires overlapping observations across instruments, which limits applicability to regions of shared sky coverage. In future work, we aim to extend the framework to unpaired settings where such overlap is unavailable or limited. Moreover, while we have focused on galaxy imaging, the framework is general and applicable to other domains in astrophysics and beyond. In particular, we plan to apply it to the tens of millions of light curves from NASA's Transiting Exoplanet Survey Satellite [TESS, 23] and Kepler [24, 25] missions.

Another limitation is that our experiments use bands with largely overlapping wavelength coverage across surveys and assume that matched views share information about the source, while measurement artifacts are weakly correlated across different instruments. Extending the framework to substantially different or non-overlapping spectral ranges remains untested. Such observations can probe distinct physical components of a source, so the model would need to distinguish wavelength-dependent source structure from instrumental effects.

6 Conclusion

We have presented a framework for learning to disentangle instrumental effects from physical signals in scenarios with multiple views of the same object, i.e., observations from different instruments, using a simple counterfactual generative objective. Our dual-encoder architecture separates physics-invariant and instrument-specific factors structurally, without requiring hand-engineered contrastive losses or explicit instrument models. Compared to previous contrastive approaches, our method enables counterfactual generation by synthesizing how a source would appear under different observing conditions, and it is designed to scale to settings with more than two instruments. The generative component can also be viewed as a data-driven instrument model, learned implicitly from cross-matched observations. We have shown an application of our model to galaxy data in astrophysics, showing that the physics latent space recovers physical properties comparably to existing foundation models while being robust to instrument systematics, and that counterfactual generation produces realistic cross-survey translations.

Acknowledgments and Disclosure of Funding

This work was supported in part by Advanced Micro Devices, Inc. under the AMD University Program's AI & HPC Cluster. This work was supported by the MIT Generative AI Impact Consortium (MGAIC) and the MIT Office of Research Computing and Data (ORCD). The project that gave rise to these results received the support of a fellowship from "la Caixa" Foundation (ID 100010434). The fellowship code is B006068. We thank the AstroAI center at the Center for Astrophysics | Harvard & Smithsonian, where P M-P was a summer researcher under the supervision of DM, Rafael Martinez-Galarza, and Cecilia Garraffo, for their support and useful discussions on multimodal foundation models for astronomy. We also thank Michael J. Smith for helping access the dataset, and Rocco Di Tella and Rebeka Bottger for insightful technical discussions.

References

[1] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12:1247 – 1283, 2000.

[2] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

[3] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

[4] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

[5] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. URL https://arxiv.org/abs/2106.09553.

[6] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211, 2022. URL https://arxiv.org/abs/2207.08051.

[7] Mike Walmsley, Inigo Val Slijepcevic, Micah Bowles, and Anna M. M. Scaife. Towards galaxy foundation models with hybrid contrastive learning, 2022. URL https://arxiv.org/abs/2206.11927.

[8] Liam Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, Lucas Meyer, Micah Bowles, Sebastian Wagner-Carena, Helen Qu, Siavash Golkar, Alberto Bietti, Hatim Bourfoune, Nathan Casserau, Pierre Cornette, Keiya Hirashima, Geraud Krawezik, Ruben Ohana, Nicholas Lourie, Michael McCabe, Rudy Morel, Payel Mukhopadhyay, Mariel Pettee, Bruno Regaldo-Saint Blancard, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. Aion-1: Omnimodal foundation model for astronomical sciences. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.17960.

[9] The Multimodal Universe Collaboration. The multimodal universe: Enabling large-scale machine learning with 100tb of astronomical scientific data, 2024. URL https://arxiv.org/abs/2412.02527.

[10] Euclid Collaboration, M. Siudek, M. Huertas-Company, M. Smith, G. Martinez-Solaeche, F. Lanusse, S. Ho, E. Angeloudi, P. A. C. Cunha, H. Domínguez Sánchez, M. Dunn, Y. Fu, P. Iglesias-Navarro, J. Junais, J. H. Knapen, B. Laloux, M. Mezcua, W. Roster, G. Stevens, J. Vega-Ferrero, N. Aghanim, B. Altieri, A. Amara, S. Andreon, N. Auricchio, H. Aussel, C. Baccigalupi, M. Baldi, S. Bardelli, P. Battaglia, A. Biviano, A. Bonchi, E. Branchini, M. Brescia, J. Brinchmann, S. Camera, G. Cañas-Herrera, V. Capobianco, C. Carbone, J. Carretero, S. Casas, F. J. Castander, M. Castellano, G. Castignani, S. Cavuoti, K. C. Chambers, A. Cimatti, C. Colodro-Conde, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, F. Courbin, H. M. Courtois, M. Cropper, A. Da Silva, H. Degaudenzi, G. De Lucia, A. M. Di Giorgio, J. Dinis, C. Dolding, H. Dole, F. Dubath, C. A. J. Duncan, X. Dupac, S. Dusini, S. Escoffier, M. Farina, R. Farinelli, F. Faustini, S. Ferriol, F. Finelli, S. Fotopoulou, M. Frailis, E. Franceschi, S. Galeotta, K. George, B. Gillis, C. Giocoli, J. Gracia-Carpio, B. R. Granett, A. Grazian, F. Grupp, S. Gwyn, S. V. H. Haugan, W. Holmes, I. M. Hook, F. Hormuth, A. Hornstrup, K. Jahnke, M. Jhabvala, E. Keihänen, S. Kermiche, A. Kiessling, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, Q. Le Boulc'h, A. M. C. Le Brun, D. Le Mignant, S. Ligori, P. B. Lilje, V. Lindholm, I. Lloro, G. Mainetti, D. Maino, E. Maiorano, O. Mansutti, S. Marcin, O. Marggraf, M. Martinelli, N. Martinet, F. Marulli, R. Massey, S. Maurogordato, H. J. McCracken, E. Medinaceli, S. Mei, M. Melchior, Y. Mellier, M. Meneghetti, E. Merlin, G. Meylan, A. Mora, M. Moresco, L. Moscardini, R. Nakajima, C. Neissner, S. M. Niemi, J. W. Nightingale, C. Padilla, S. Paltani, F. Pasian, K. Pedersen, W. J. Percival, V. Pettorino, S. Pires, G. Polenta, M. Poncet, L. A. Popa, L. Pozzetti, F. Raison, A. Renzi, J. Rhodes, G. Riccio, E. Romelli, M. Roncarelli, R. Saglia, Z. Sakr, A. G. Sánchez, D. Sapone, B. Sartoris, J. A. Schewtschenko, P. Schneider, T. Schrabback, M. Scodeggio, A. Secroun, G. Seidel, M. Seiffert, S. Serrano, P. Simon, C. Sirignano, G. Sirri, L. Stanco, J. Steinwagner, P. Tallada-Crespí, A. N. Taylor, I. Tereno, S. Toft, R. Toledo-Moreo, F. Torradeflot, I. Tutusaus, L. Valenziano, J. Valiviita, T. Vassallo, G. Verdoes Kleijn, A. Veropalumbo, Y. Wang, J. Weller, A. Zacchei, G. Zamorani, F. M. Zerbi, I. A. Zinchenko, E. Zucca, V. Allevato, M. Ballardini, M. Bolzonella, E. Bozzo, C. Burigana, R. Cabanac, A. Cappi, D. Di Ferdinando, J. A. Escartin Vigo, L. Gabarra, J. Martín-Fleitas, S. Matthew, N. Mauri, R. B. Metcalf, and A. Pezzotta. Euclid Quick Data Release (Q1) Exploring galaxy properties with a multi-modal foundation model. *arXiv e-prints*, art. arXiv:2503.15312, March 2025. doi: 10.48550/arXiv.2503.15312.

[11] Jeroen Audenaert, Daniel Muthukrishna, Paul F. Gregory, David W. Hogg, and V. Ashley Villar. Causal foundation models: Disentangling physics from instrument properties, 2025. URL https://arxiv.org/abs/2507.05333.

[12] Bernhard Schölkopf, David W. Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398, 2016. doi: 10.1073/pnas.1511656113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1511656113.

[13] Soichiro Hattori, Daniel Foreman-Mackey, David W. Hogg, Benjamin T. Montet, Ruth Angus, T. A. Pritchard, Jason L. Curtis, and Bernhard Schölkopf. The unpopular Package: A Data-driven Approach to Detrending TESS Full-frame Image Light Curves. *The Astronomical Journal*, 163(6):284, June 2022. doi: 10.3847/1538-3881/ac625a.

[14] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

[15] Nolan Koblischke, Liam Parker, Francois Lanusse, Jo Bovy, Irina Espejo, and Shirley Ho. Semantic search for 100M+ galaxy images using AI-generated captions. *The Astrophysical Journal*, 1005(2):170, July 2026. doi: 10.3847/1538-4357/ae7b22. URL https://doi.org/10.3847/1538-4357/ae7b22.

[16] Arjun Dey, David J. Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, Xiaohui Fan, Joseph R. Findlay, Doug Finkbeiner, David Herrera, Stéphanie Juneau, Martin Landriau, Michael Levi, Ian McGreer, Aaron Meisner, Adam D. Myers, John Moustakas, Peter Nugent, Anna Patej, Edward F. Schlafly, Alistair R. Walker, Francisco Valdes, Benjamin A. Weaver, Christophe Yèche, Hu Zou, Xu Zhou, Behzad Abareshi, T. M. C. Abbott, Bela Abolfathi, C. Aguilera, Shadab Alam, Lori Allen, A. Alvarez, James Annis, Behzad Ansarinejad, Marie Aubert, Jacqueline Beechert, Eric F. Bell, Segev Y. BenZvi, Florian Beutler, Richard M. Bielby,

Adam S. Bolton, César Briceño, Elizabeth J. Buckley-Geer, Karen Butler, Annalisa Calamida, Raymond G. Carlberg, Paul Carter, Ricard Casas, Francisco J. Castander, Yumi Choi, Johan Comparat, Elena Cukanovaite, Timothée Delubac, Kaitlin DeVries, Sharmila Dey, Govinda Dhungana, Mark Dickinson, Zhejie Ding, John B. Donaldson, Yutong Duan, Christopher J. Duckworth, Sarah Eftekharzadeh, Daniel J. Eisenstein, Thomas Etourneau, Parker A. Fagrelius, Jay Farihi, Mike Fitzpatrick, Andreu Font-Ribera, Leah Fulmer, Boris T. Gänsicke, Enrique Gaztanaga, Koshy George, David W. Gerdes, Satya Gontcho A Gontcho, Claudio Gorgoni, Gregory Green, Julien Guy, Diane Harmer, M. Hernandez, Klaus Honscheid, Lijuan (Wendy) Huang, David J. James, Buell T. Jannuzi, Linhua Jiang, Richard Joyce, Armin Karcher, Sonia Karkar, Robert Kehoe, Jean-Paul Kneib, Andrea Kueter-Young, Ting-Wen Lan, Tod R. Lauer, Laurent Le Guillou, Auguste Le Van Suu, Jae Hyeon Lee, Michael Lesser, Laurence Perreault Levasseur, Ting S. Li, Justin L. Mann, Robert Marshall, C. E. Martínez-Vázquez, Paul Martini, Hélion du Mas des Bourboux, Sean McManus, Tobias Gabriel Meier, Brice Ménard, Nigel Metcalfe, Andrea Muñoz-Gutiérrez, Joan Najita, Kevin Napier, Gautham Narayan, Jeffrey A. Newman, Jundan Nie, Brian Nord, Dara J. Norman, Knut A. G. Olsen, Anthony Paat, Nathalie Palanque-Delabrouille, Xiyan Peng, Claire L. Poppett, Megan R. Poremba, Abhishek Prakash, David Rabinowitz, Anand Raichoor, Mehdi Rezaie, A. N. Robertson, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Gregory Rudnick, Sasha Gaines, Abhijit Saha, F. Javier Sánchez, Elodie Savary, Heidi Schweiker, Adam Scott, Hee-Jong Seo, Huanyuan Shan, David R. Silva, Zachary Slepian, Christian Soto, David Sprayberry, Ryan Staten, Coley M. Stillman, Robert J. Stupak, David L. Summers, Suk Sien Tie, H. Tirado, Mariana Vargas-Magaña, A. Katherina Vivas, Risa H. Wechsler, Doug Williams, Jinyi Yang, Qian Yang, Tolga Yapici, Dennis Zaritsky, A. Zenteno, Kai Zhang, Tianmeng Zhang, Rongpu Zhou, and Zhimin Zhou. Overview of the desi legacy imaging surveys. *The Astronomical Journal*, 157(5):168, April 2019. ISSN 1538-3881. doi: 10.3847/1538-3881/ab089d. URL http://dx.doi.org/10.3847/1538-3881/ab089d.

[17] Hiroaki Aihara, Nobuo Arimoto, Robert Armstrong, Stéphane Arnouts, Neta A Bahcall, Steven Bickerton, James Bosch, Kevin Bundy, Peter L Capak, James H H Chan, Masashi Chiba, Jean Coupon, Eiichi Egami, Motohiro Enoki, Francois Finet, Hiroki Fujimori, Seiji Fujimoto, Hisanori Furusawa, Junko Furusawa, Tomotsugu Goto, Andy Goulding, Johnny P Greco, Jenny E Greene, James E Gunn, Takashi Hamana, Yuichi Harikane, Yasuhiro Hashimoto, Takashi Hattori, Masao Hayashi, Yusuke Hayashi, Krzysztof G Hełminiak, Ryo Higuchi, Chiaki Hikage, Paul T P Ho, Bau-Ching Hsieh, Kuiyun Huang, Song Huang, Hiroyuki Ikeda, Masatoshi Imanishi, Akio K Inoue, Kazushi Iwasawa, Ikuru Iwata, Anton T Jaelani, Hung-Yu Jian, Yukiko Kamata, Hiroshi Karoji, Nobunari Kashikawa, Nobuhiko Katayama, Satoshi Kawanomoto, Issha Kayo, Jin Koda, Michitaro Koike, Takashi Kojima, Yutaka Komiyama, Akira Konno, Shintaro Koshida, Yusei Koyama, Haruka Kusakabe, Alexie Leauthaud, Chien-Hsiu Lee, Lihwai Lin, Yen-Ting Lin, Robert H Lupton, Rachel Mandelbaum, Yoshiki Matsuoka, Elinor Medezinski, Sogo Mineo, Shoken Miyama, Hironao Miyatake, Satoshi Miyazaki, Rieko Momose, Anupreeta More, Surhud More, Yuki Moritani, Takashi J Moriya, Tomoki Morokuma, Shiro Mukae, Ryoma Murata, Hitoshi Murayama, Tohru Nagao, Fumiaki Nakata, Mana Niida, Hiroko Niikura, Atsushi J Nishizawa, Yoshiyuki Obuchi, Masamune Oguri, Yukie Oishi, Nobuhiro Okabe, Sakurako Okamoto, Yuki Okura, Yoshiaki Ono, Masato Onodera, Masafusa Onoue, Ken Osato, Masami Ouchi, Paul A Price, Tae-Soo Pyo, Masao Sako, Marcin Sawicki, Takatoshi Shibuya, Kazuhiro Shimasaku, Atsushi Shimono, Masato Shirasaki, John D Silverman, Melanie Simet, Joshua Speagle, David N Spergel, Michael A Strauss, Yuma Sugahara, Naoshi Sugiyama, Yasushi Suto, Sherry H Suyu, Nao Suzuki, Philip J Tait, Masahiro Takada, Tadafumi Takata, Naoyuki Tamura, Manobu M Tanaka, Masaomi Tanaka, Masayuki Tanaka, Yoko Tanaka, Tsuyoshi Terai, Yuichi Terashima, Yoshiki Toba, Nozomu Tominaga, Jun Toshikawa, Edwin L Turner, Tomohisa Uchida, Hisakazu Uchiyama, Keiichi Umetsu, Fumihiro Uraguchi, Yuji Urata, Tomonori Usuda, Yousuke Utsumi, Shiang-Yu Wang, Wei-Hao Wang, Kenneth C Wong, Kiyoto Yabe, Yoshihiko Yamada, Hitomi Yamanoi, Naoki Yasuda, Sherry Yeh, Atsunori Yonehara, and Suraphong Yuma. The hyper suprime-cam ssp survey: Overview and survey design. *Publications of the Astronomical Society of Japan*, 70(SP1), September 2017. ISSN 2053-051X. doi: 10.1093/pasj/psx066. URL http://dx.doi.org/10.1093/pasj/psx066.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[19] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[20] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

[21] ChangHoon Hahn, K. J. Kwon, Rita Tojeiro, Malgorzata Siudek, Rebecca E. A. Canning, Mar Mezcua, Jeremy L. Tinker, David Brooks, Peter Doel, Kevin Fanning, Enrique Gaztañaga, Robert Kehoe, Martin Landriau, Aaron Meisner, John Moustakas, Claire Poppett, Gregory Tarle, Benjamin Weiner, and Hu Zou. The DESI PRObabilistic Value-added Bright Galaxy Survey (PROVABGS) Mock Challenge. *The Astrophysical Journal*, 945(1):16, March 2023. doi: 10.3847/1538-4357/ac8983.

[22] Andrew M. Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y. Ng. On random weights and unsupervised feature learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1089–1096, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

[23] George R. Ricker, Joshua N. Winn, Roland Vanderspek, David W. Latham, Gáspár Á. Bakos, Jacob L. Bean, Zachory K. Berta-Thompson, Timothy M. Brown, Lars Buchhave, Nathaniel R. Butler, R. Paul Butler, William J. Chaplin, David Charbonneau, Jørgen Christensen-Dalsgaard, Mark Clampin, Drake Deming, John Doty, Nathan De Lee, Courtney Dressing, Edward W. Dunham, Michael Endl, Francois Fressin, Jian Ge, Thomas Henning, Matthew J. Holman, Andrew W. Howard, Shigeru Ida, Jon M. Jenkins, Garrett Jernigan, John Asher Johnson, Lisa Kaltenegger, Nobuyuki Kawai, Hans Kjeldsen, Gregory Laughlin, Alan M. Levine, Douglas Lin, Jack J. Lissauer, Phillip MacQueen, Geoffrey Marcy, Peter R. McCullough, Timothy D. Morton, Norio Narita, Martin Paegert, Enric Palle, Francesco Pepe, Joshua Pepper, Andreas Quirrenbach, Stephen A. Rinehart, Dimitar Sasselov, Bun'ei Sato, Sara Seager, Alessandro Sozzetti, Keivan G. Stassun, Peter Sullivan, Andrew Szentgyorgyi, Guillermo Torres, Stephane Udry, and Joel Villasenor. Transiting Exoplanet Survey Satellite (TESS). *Journal of Astronomical Telescopes, Instruments, and Systems*, 1:014003, January 2015. doi: 10.1117/1.JATIS.1.1.014003.

[24] William J. Borucki, David Koch, Gibor Basri, Natalie Batalha, Timothy Brown, Douglas Caldwell, John Caldwell, Jørgen Christensen-Dalsgaard, William D. Cochran, Edna DeVore, Edward W. Dunham, Andrea K. Dupree, Thomas N. Gautier, John C. Geary, Ronald Gilliland, Alan Gould, Steve B. Howell, Jon M. Jenkins, Yoji Kondo, David W. Latham, Geoffrey W. Marcy, Søren Meibom, Hans Kjeldsen, Jack J. Lissauer, David G. Monet, David Morrison, Dimitar Sasselov, Jill Tarter, Alan Boss, Don Brownlee, Toby Owen, Derek Buzasi, David Charbonneau, Laurance Doyle, Jonathan Fortney, Eric B. Ford, Matthew J. Holman, Sara Seager, Jason H. Steffen, William F. Welsh, Jason Rowe, Howard Anderson, Lars Buchhave, David Ciardi, Lucianne Walkowicz, William Sherry, Elliott Horch, Howard Isaacson, Mark E. Everett, Debra Fischer, Guillermo Torres, John Asher Johnson, Michael Endl, Phillip MacQueen, Stephen T. Bryson, Jessie Dotson, Michael Haas, Jeffrey Kolodziejczak, Jeffrey Van Cleve, Hema Chandrasekaran, Joseph D. Twicken, Elisa V. Quintana, Bruce D. Clarke, Christopher Allen, Jie Li, Haley Wu, Peter Tenenbaum, Ekaterina Verner, Frederick Bruhweiler, Jason Barnes, and Andrej Prsa. Kepler Planet-Detection Mission: Introduction and First Results. *Science*, 327(5968):977, February 2010. doi: 10.1126/science.1185402.

[25] David G. Koch, William J. Borucki, Gibor Basri, Natalie M. Batalha, Timothy M. Brown, Douglas Caldwell, Jørgen Christensen-Dalsgaard, William D. Cochran, Edna DeVore, Edward W. Dunham, III Gautier, Thomas N., John C. Geary, Ronald L. Gilliland, Alan Gould, Jon Jenkins, Yoji Kondo, David W. Latham, Jack J. Lissauer, Geoffrey Marcy, David Monet, Dimitar Sasselov, Alan Boss, Donald Brownlee, John Caldwell, Andrea K. Dupree, Steve B. Howell, Hans Kjeldsen, Søren Meibom, David Morrison, Tobias Owen, Harold Reitsema, Jill Tarter, Stephen T. Bryson, Jessie L. Dotson, Paul Gazis, Michael R. Haas, Jeffrey Kolodziejczak, Jason F. Rowe, Jeffrey E. Van Cleve, Christopher Allen, Hema Chandrasekaran, Bruce D. Clarke, Jie Li, Elisa V. Quintana, Peter Tenenbaum, Joseph D. Twicken, and Hayley Wu. Kepler Mission Design, Realized Photometric Performance, and Early Science. *The Astrophysical Journal Letters*, 713(2):L79–L86, April 2010. doi: 10.1088/2041-8205/713/2/L79.

[26] Hiroaki Aihara, Nobuo Arimoto, Robert Armstrong, Stéphane Arnouts, Neta A. Bahcall, Steven Bickerton, James Bosch, Kevin Bundy, Peter L. Capak, James H. H. Chan, Masashi Chiba, Jean Coupon, Eiichi Egami, Motohiro Enoki, Francois Finet, Hiroki Fujimori, Seiji Fujimoto, Hisanori Furusawa, Junko Furusawa, Tomotsugu Goto, Andy Goulding, Johnny P. Greco, Jenny E. Greene, James E. Gunn, Takashi Hamana, Yuichi Harikane, Yasuhiro Hashimoto, Takashi Hattori, Masao Hayashi, Yusuke Hayashi, Krzysztof G. Hełminiak, Ryo Higuchi, Chiaki Hikage, Paul T. P. Ho, Bau-Ching Hsieh, Kuiyun Huang, Song Huang, Hiroyuki Ikeda, Masatoshi Imanishi, Akio K. Inoue, Kazushi Iwasawa, Ikuru Iwata, Anton T. Jaelani, Hung-Yu Jian, Yukiko Kamata, Hiroshi Karoji, Nobunari Kashikawa, Nobuhiko Katayama, Satoshi Kawanomoto, Issha Kayo, Jin Koda, Michitaro Koike, Takashi Kojima, Yutaka Komiyama, Akira Konno, Shintaro Koshida, Yusei Koyama, Haruka Kusakabe, Alexie Leauthaud, Chien-Hsiu Lee, Lihwai Lin, Yen-Ting Lin, Robert H. Lupton, Rachel Mandelbaum, Yoshiki Matsuoka, Elinor Medezinski, Sogo Mineo, Shoken Miyama, Hironao Miyatake, Satoshi Miyazaki, Rieko Momose, Anupreeta More, Surhud More, Yuki Moritani, Takashi J. Moriya, Tomoki Morokuma, Shiro Mukae, Ryoma Murata, Hitoshi Murayama, Tohru Nagao, Fumiaki Nakata, Mana Niida, Hiroko Niikura, Atsushi J. Nishizawa, Yoshiyuki Obuchi, Masamune Oguri, Yukie Oishi, Nobuhiro Okabe, Sakurako Okamoto, Yuki Okura, Yoshiaki Ono, Masato Onodera, Masafusa Onoue, Ken Osato, Masami Ouchi, Paul A. Price, Tae-Soo Pyo, Masao Sako, Marcin Sawicki, Takatoshi Shibuya, Kazuhiro Shimasaku, Atsushi Shimono, Masato Shirasaki, John D. Silverman, Melanie Simet, Joshua Speagle, David N. Spergel, Michael A. Strauss, Yuma Sugahara, Naoshi Sugiyama, Yasushi Suto, Sherry H. Suyu, Nao Suzuki, Philip J. Tait, Masahiro Takada, Tadafumi Takata, Naoyuki Tamura, Manobu M. Tanaka, Masaomi Tanaka, Masayuki Tanaka, Yoko Tanaka, Tsuyoshi Terai, Yuichi Terashima, Yoshiki Toba, Nozomu Tominaga, Jun Toshikawa, Edwin L. Turner, Tomohisa Uchida, Hisakazu Uchiyama, Keiichi Umetsu, Fumihiro Uraguchi, Yuji Urata, Tomonori Usuda, Yousuke Utsumi, Shiang-Yu Wang, Wei-Hao Wang, Kenneth C. Wong, Kiyoto Yabe, Yoshihiko Yamada, Hitomi Yamanoi, Naoki Yasuda, Sherry Yeh, Atsunori Yonehara, and Suraphong Yuma. The Hyper Suprime-Cam SSP Survey: Overview and survey design. *Publications of the Astronomical Society of Japan*, 70:S4, January 2018. doi: 10.1093/pasj/psx066.

[27] Arjun Dey, David J. Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, Xiaohui Fan, Joseph R. Findlay, Doug Finkbeiner, David Herrera, Stéphanie Juneau, Martin Landriau, Michael Levi, Ian McGreer, Aaron Meisner, Adam D. Myers, John Moustakas, Peter Nugent, Anna Patej, Edward F. Schlafly, Alistair R. Walker, Francisco Valdes, Benjamin A. Weaver, Christophe Yèche, Hu Zou, Xu Zhou, Behzad Abareshi, T. M. C. Abbott, Bela Abolfathi, C. Aguilera, Shadab Alam, Lori Allen, A. Alvarez, James Annis, Behzad Ansarinejad, Marie Aubert, Jacqueline Beechert, Eric F. Bell, Segev Y. BenZvi, Florian Beutler, Richard M. Bielby, Adam S. Bolton, César Briceño, Elizabeth J. Buckley-Geer, Karen Butler, Annalisa Calamida, Raymond G. Carlberg, Paul Carter, Ricard Casas, Francisco J. Castander, Yumi Choi, Johan Comparat, Elena Cukanovaite, Timothée Delubac, Kaitlin DeVries, Sharmila Dey, Govinda Dhungana, Mark Dickinson, Zhejie Ding, John B. Donaldson, Yutong Duan, Christopher J. Duckworth, Sarah Eftekharzadeh, Daniel J. Eisenstein, Thomas Etourneau, Parker A. Fagrelius, Jay Farihi, Mike Fitzpatrick, Andreu Font-Ribera, Leah Fulmer, Boris T. Gänsicke, Enrique Gaztanaga, Koshy George, David W. Gerdes, Satya Gontcho A. Gontcho, Claudio Gorgoni, Gregory Green, Julien Guy, Diane Harmer, M. Hernandez, Klaus Honscheid, Lijuan Wendy Huang, David J. James, Buell T. Jannuzi, Linhua Jiang, Richard Joyce, Armin Karcher, Sonia Karkar, Robert Kehoe, Jean-Paul Kneib, Andrea Kueter-Young, Ting-Wen Lan, Tod R. Lauer, Laurent Le Guillou, Auguste Le Van Suu, Jae Hyeon Lee, Michael Lesser, Laurence Perreault Levasseur, Ting S. Li, Justin L. Mann, Robert Marshall, C. E. Martínez-Vázquez, Paul Martini, Hélion du Mas des Bourboux, Sean McManus, Tobias Gabriel Meier, Brice Ménard, Nigel Metcalfe, Andrea Muñoz-Gutiérrez, Joan Najita, Kevin Napier, Gautham Narayan, Jeffrey A. Newman, Jundan Nie, Brian Nord, Dara J. Norman, Knut A. G. Olsen, Anthony Paat, Nathalie Palanque-Delabrouille, Xiyan Peng, Claire L. Poppett, Megan R. Poremba, Abhishek Prakash, David Rabinowitz, Anand Raichoor, Mehdi Rezaie, A. N. Robertson, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Gregory Rudnick, Sasha Safonova, Abhijit Saha, F. Javier Sánchez, Elodie Savary, Heidi Schweiker, Adam Scott, Hee-Jong Seo, Huanyuan Shan, David R. Silva, Zachary Slepian, Christian Soto, David Sprayberry, Ryan Staten, Coley M. Stillman, Robert J. Stupak, David L. Summers, Suk Sien Tie, H. Tirado, Mariana Vargas-Magaña, A. Katherina Vivas, Risa H. Wechsler, Doug Williams, Jinyi Yang, Qian Yang, Tolga Yapici, Dennis Zaritsky,

A. Zenteno, Kai Zhang, Tianmeng Zhang, Rongpu Zhou, and Zhimin Zhou. Overview of the DESI Legacy Imaging Surveys. *The Astronomical Journal*, 157(5):168, May 2019. doi: 10.3847/1538-3881/ab089d.

[28] François Lanusse, Liam Parker, Siavash Golkar, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Ruben Ohana, Mariel Pettee, Bruno Régaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. AstroCLIP: Cross-modal pre-training for astronomical foundation models. *arXiv preprint arXiv:2310.03024*, 2023.

[29] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[30] François Rozet, Felix Andry, François Lanusse, and Gilles Louppe. Zuko: Normalizing flows in PyTorch, 2023. URL https://github.com/probabilists/zuko.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

A Appendix

A.1 Model Architecture

We use a ResNet-18 [18] architecture for the encoders, plus a `UNet2DConditionModel` [19] for the flow-matching generative decoder. The final pooling and projection layers of the ResNet are replaced with a convolution that produces a 2×2 feature map with 16 channels, which is treated as a sequence of 4 tokens of size 16 and passed to the UNet via attention conditioning (matching the cross-attention dimension of 16). Table 1 shows the architecture parameters used in the final training run.

In our two-survey setting, the physics encoder provides 4 tokens from the same source observed by the other instrument, and the instrument encoder provides $4K$ tokens from K neighboring sources observed by the target instrument. Concatenating these tokens along the sequence axis gives $4(1+K)$ conditioning tokens of dimension 16. For our training run we use five instrument neighbors so $K = 5$ and we have a sequence of 24 tokens. This sequence is used as the key and value tokens for the cross-attention layers of the velocity network's cross-attention layers, while the UNet's spatial activations are the queries. Cross-attention is applied at every stage of the network, in all four downsampling blocks and all four upsampling blocks. Because cross-attention can intake a variable-length token sequence, the same decoder parameters can accommodate different numbers of conditioning views.

Table 1: Model architecture parameters for the conditional flow matching model with dual ResNet-18 encoders.

Parameter	**Value**
Data Dimensions	
Image size	48×48
Input channels	4
Conditioning channels	4
UNet Architecture	
Base model channels	128
Channel multiplier	$(1, 2, 4, 4)$
Layers per block	2
Attention head dimension	8
Cross-attention dimension	16
Block types (down/up)	CrossAttnDownBlock2D / CrossAttnUpBlock2D
Encoder Architecture	
Backbone	ResNet-18 (timm)
Pretrained weights	False
Projection layer	Replaced by Conv2d(512 $\rightarrow$ 16, kernel=1)

A.2 Training Configuration

The model was trained using distributed data parallel across 4 NVIDIA H100 GPUs with *bfloat16* mixed precision, completing 75,000 training steps in approximately 8 hours of wall-clock time. Table 2 shows the training configuration and hyperparameters used in the final training run.

A.3 Model Used to Evaluate Downstream Performance

To evaluate the information content of our disentangled latent spaces, we perform supervised regression on a variety of physical and instrumental properties. We use a three-layer Multi-Layer Perceptron (MLP) with hidden dimensions of $[512, 256, 128]$, employing LayerNorm, GELU activations, and 20% dropout. The models are optimized using a Smooth L1 loss to ensure robustness against outliers in the catalog labels.

Table 2: Training configuration and hyperparameters.

Parameter	Value
Optimization	
Optimizer	AdamW
Learning rate	1×10^{-4}
LR scheduler	CosineAnnealingLR
Loss function	Velocity MSE
Flow Matching	
Integration steps (sampling)	250
Interpolation	$x_t = (1-t)x_0 + tx_1$
Target velocity	$v(x_t, t, c) = x_1 - x_0$
Training Details	
Batch size	64
Total training steps	75,000
Number of devices	4 (DDP)
Precision	bf16-mixed
Validation check interval	Every 1,000 steps
Train/val split ratio	95% / 5%

A.4 Galaxy Properties Used for Downstream Task

Dataset composition. All downstream-task evaluations are conducted on a single cross-matched sample of $n = 5{,}469$ paired galaxies. The MLP probe is trained on a random 90% split of this sample and evaluated on the held-out 10%.

For the subset of physical properties that require DESI BGS spectroscopy (redshift z, $\log M_\star$, sSFR, mass-weighted stellar age $t_{\rm age}$ and mass-weighted metallicity $\log Z_{\rm Met}$, all taken from the PROVABGS posterior catalogue), we drop the ~ 800 rows where the spectroscopic label is missing, leaving $n_{\rm spec} = 4{,}666$ galaxies. For every other target — ellipticity components e_1, e_2 (from both the Legacy and HSC shape catalogues), Galactic extinction $E(B{-}V)$ on the Legacy side and the HSC extinction a (cross-matched in from the HSC catalogue `a_*` columns), as well as the instrumental metadata (PSF size, PSF depth, galaxy depth and observation count) — the full $n_{\rm tot} = 5{,}469$ sample is used. See below for a more in-depth description of the individual properties.

Physical properties. Table 3 summarizes the physical parameters used to evaluate the physics latent space. These span a range of galaxy properties: morphological quantities (ellipticity and half-light radius) measured directly from HSC imaging, spectroscopic redshift from DESI, and stellar population properties (stellar mass, specific star formation rate, metallicity, and stellar age) derived from Bayesian SED fitting by PROVABGS [21]. The SED-derived quantities are inferred by jointly fitting stellar population synthesis models to DESI spectra and Legacy Survey photometry, and thus represent indirect measurements that integrate information beyond what is available in broadband imaging alone. Redshift is the most precisely determined of these quantities, as it is measured directly from spectral line positions.

Instrumental properties. Table 4 summarizes the instrumental and observational parameters used to evaluate the instrument latent space. These quantities describe the local observing conditions at each galaxy's position and vary spatially across each survey's footprint. For the Legacy Survey, the catalog provides direct measurements of depth (both for point sources and extended galaxies), PSF size, number of contributing exposures, and foreground dust extinction. HSC provides analogous quantities for PSF size and dust extinction, though computed differently as noted below. We evaluate the instrument latent space on properties from both surveys to assess whether the encoder captures instrument-specific conditions for the survey it is given, without encoding information from the other survey.

Table 3: Physical parameters used for downstream regression evaluation of the physics latent space.

Variable	Source	Description
Ellipticity (e_1, e_2)	HSC catalog	Galaxy shape ellipticity components, measured via the SDSS-style second moments of the surface brightness distribution.
Redshift (z)	DESI spectra	Spectroscopic redshift.
Stellar mass $(M_\star)$	PROVABGS	Total mass in stars,, derived from Bayesian spectral energy distribution (SED) fitting to DESI spectra [21].
Half-light radius (R_{eff})	HSC catalog	Circularised radius enclosing half of the galaxy's total light. A measure of galaxy size.
sSFR	PROVABGS	Specific star formation rate $\log(\text{SFR}/M_\star)$: the rate at which new stars are forming normalised by the galaxy's stellar mass.
$\log Z_{\text{Met}}$	PROVABGS	Gas-phase metallicity: the abundance of elements heavier than helium in the galaxy's gas. Traces the galaxy's chemical enrichment history.
t_{age}	PROVABGS	Mass-weighted stellar population age: the average age of the stars in the galaxy.

Table 4: Instrumental and observational parameters used for downstream regression evaluation of the instrument latent space. All quantities vary spatially across each survey's footprint.

Variable	Bands	Description
HSC (Hyper Suprime-Cam)		
`psf_fwhm`	g, i, r, z	Effective size of the point spread function (arcsec): how much the instrument and atmosphere spread the light from a point source. Analogous to Legacy `PSFSIZE`.†
`a`	g, i, r, z	Milky Way dust extinction A_λ per band (mag). Equivalent to Legacy `EBV` scaled by a filter-dependent coefficient: $A_\lambda = R_\lambda\, E(B-V)$.‡
Legacy Survey (DESI Legacy Imaging Surveys)		
`NOBS`	g, i, r, z	Number of exposures contributing to the coadd.
`GALDEPTH`	g, i, r, z	5σ detection depth for a round exponential galaxy with $r_e = 0.45''$ (mag). Shallower than `PSFDEPTH` because galaxy flux is spread over more pixels.
`PSFDEPTH`	g, i, r, z	5σ detection depth for a point source (mag).
`PSFSIZE`	g, i, r, z	Effective PSF size (arcsec). See HSC `psf_fwhm` above.†
`EBV`	—	Milky Way dust extinction as colour excess $E(B-V)$ (mag). See HSC `a` above.‡

† Both quantities measure how much the instrument and atmosphere spread the light from a point source, but are computed differently: HSC derives it from the second moments of the coadd PSF model at each source position [26]; the Legacy Survey reports the exposure-weighted mean of the per-CCD PSF FWHM [27].

‡ Both are derived from the same Milky Way dust maps. Galactic extinction is a foreground astrophysical effect that depends on sky position, not on the instrument; it is included here because it modifies the observed flux analogously to an instrument systematic.

A.5 Downstream Task Results

Table 5 reports the full set of MLP-probe R^2 scores on the held-out *overlap*-MMU sample, expanding on the summary in Fig. 4. We probe four representations: our physics latents (flattened across tokens), our instrument latents (mean-pooled across tokens), AION-1 (Base) embeddings [8], and a frozen randomly initialized ResNet-18 with the same architecture as our encoders. For instrument properties only, we additionally include a *cross-predict* baseline: a ResNet-18 trained from scratch to predict

Table 5: MLP-probe R^2 on the held-out *overlap*-MMU sample. **Bold** = best in row among the five probe methods (the two right-most columns are literature reference values and are not compared against). 'Ours (Physics)' and 'Ours (Instrument)' use our flow-matching encoder's e_1 (flat) and e_2 (mean-pooled) latents respectively. 'AION-1 (Base)' is the AION-1 Base model [8] probed identically. 'Rand. Init. ResNet' is a frozen randomly-initialised ResNet-18 of the same architecture as our encoders. 'Cross-predict' is a fresh ResNet-18 trained from scratch to predict each survey's instrument properties from the *other* survey's raw image (only defined for instrument targets). Reference values: AION-1-B (Ph+Im) from [8] Table 1; AstroCLIP image-only on Legacy $\{g, r, z\}$ from [28]. Band-averaged where applicable.

Property	Probe on cross-matched Legacy-HSC sample (Im)					Literature Reported Values	
	Ours (Phys)	Ours (Instr)	AION-1-B	Rnd In. ResNet	X-Predict	AION-1-B (Ph+Im)	AstroCLIP (Im)
Phys.-related properties							
Ellipticity e_2	0.934	0.321	**0.941**	0.616	—	—	—
Ellipticity e_1	0.918	0.421	**0.939**	0.613	—	—	—
Redshift z	0.801	0.535	**0.857**	0.514	—	0.930	0.780
$\log M_\star$	0.755	0.491	**0.818**	0.485	—	0.890	0.730
sSFR	0.558	0.297	**0.624**	0.312	—	0.640	0.420
$t_{\rm age}$	0.257	0.114	**0.353**	0.091	—	0.450	0.290
$\log Z_{\rm Met}$	0.250	0.155	**0.276**	0.123	—	0.490	0.360
Legacy Ins. Properties							
Galaxy Depth	0.185	**0.682**	0.426	0.202	0.214	—	—
PSF Depth	0.181	**0.667**	0.409	0.191	0.210	—	—
PSF size	0.284	**0.556**	0.290	0.111	0.313	—	—
Extinction $E(B-V)$	0.159	0.551	0.326	0.232	**0.648**	—	—
# Observations	0.074	**0.501**	0.166	0.073	0.197	—	—
HSC Ins. Properties							
Extinction a	0.114	0.546	0.299	0.204	**0.669**	—	—
PSF size	0.069	**0.438**	0.097	0.021	0.368	—	—

each survey's instrument properties from the *other* survey's raw image. The two right-most columns are external reference values from the literature and are not compared against the other entries.

The physics latent space matches AION-1 within a few percentage points on most physics-related targets, despite using a much smaller latent dimension. The instrument latent space dominates on instrument-related properties, while the physics latents collapse near the random-baseline level on those same properties — consistent with the disentanglement reported in the main text. The cross-predict baseline outperforms our instrument latents only on extinction-like quantities, where information about the other survey's footprint is sufficient to recover the target.

The role of spatial-neighbor pairing. Recall that the instrument encoder is conditioned on the five nearest spatial neighbors of the anchor within the same survey, motivated by the fact that nearby observations share similar instrument conditions. To validate this choice, we train an ablation variant in which the five conditioning galaxies are drawn at random instead. The instrument latent's recovery of instrument properties collapses (e.g. R^2 on Galaxy Depth drops from 0.68 to 0.20, and on $\#$ Observations from 0.50 to nearly 0), while physics-target performance changes only modestly. This confirms that spatially-close pairing supplies the training signal driving instrument-side disentanglement: random pairs do not present consistent enough instrument conditions for the model to separate them from physics.

A.6 Generative Model Evaluation

We assess the counterfactual generative decoder along three complementary axes: (1) preservation of spatial structure across scales, (2) calibration of the pixel-wise posterior, and (3) preservation of intrinsic galaxy morphology under pipeline transfer (the last is presented in the main text; see also Fig. 9, right). All evaluations use a held-out test set, with 32 posterior samples generated per anchor.

Spatial structure preservation. To test that the generated images preserve realistic spatial structure and correlation patterns, we compute the power spectral density (PSD) and pixel-lag autocorrelation for each of the four imaging bands, in both reconstruction directions (Legacy→HSC and HSC→Legacy). Fig. 8 shows strong agreement between ground truth and generated samples across spatial scales for both directions, indicating that the model reproduces realistic galaxy morphologies and instrument noise characteristics rather than pixel-level artifacts (which would manifest as excess

high-frequency power) or overly-smoothed outputs (which would appear as a deficit). The posterior uncertainty bands (shaded regions) appropriately expand at higher spatial frequencies, where noise dominates.

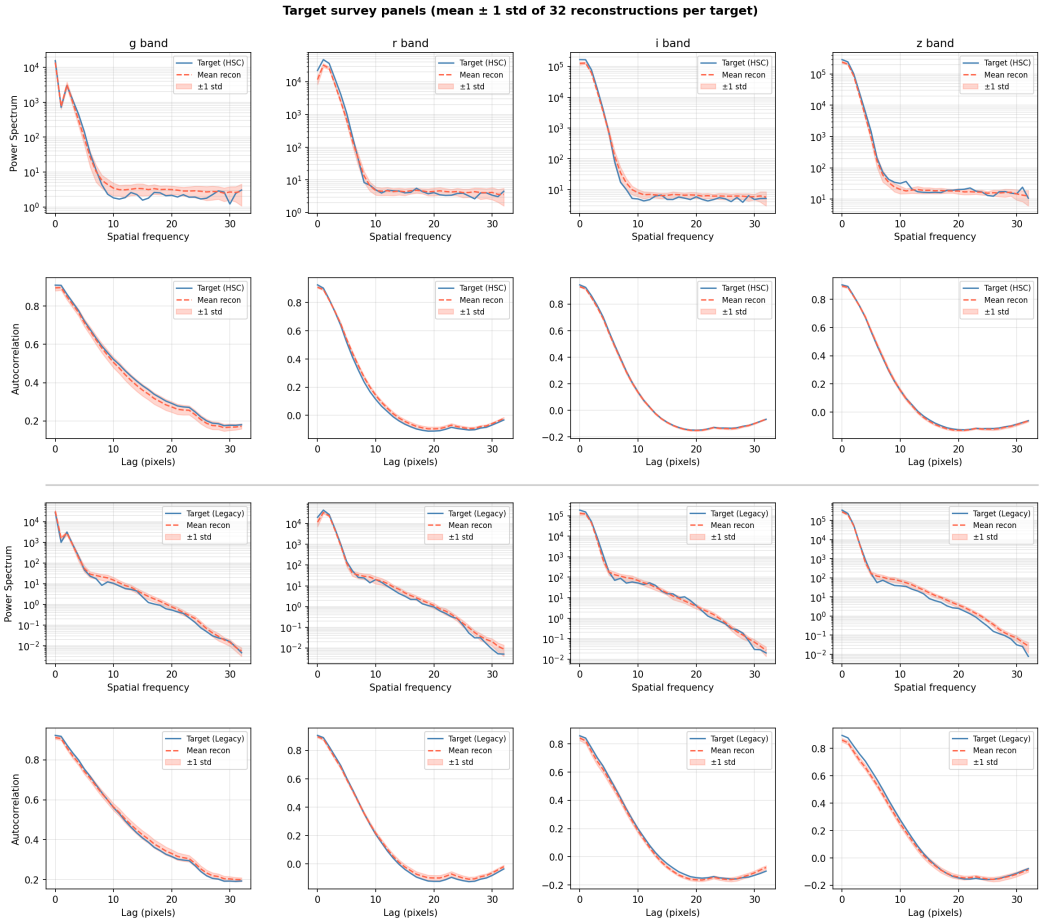


Figure 8: **Spatial Structure Preservation.** Power spectrum (left) and autocorrelation (right) per band for ground-truth and generated images. Each plot shows the true value (blue), the posterior mean from 32 generated samples (orange dashed line), and the 1σ posterior standard deviation (orange shaded region). Strong agreement across spatial frequencies and pixel separation (lag) distances indicates that the model captures realistic morphological structure and noise properties.

Pixel-wise posterior calibration. We further validate the uncertainty calibration of the flow-matching posterior using the pixel-wise Z-score,

$$Z = \frac{x - \mathbb{E}[\hat{x}]}{\mathrm{std}(\hat{x})},$$

where x is the ground-truth pixel value and $\mathbb{E}[\hat{x}]$, $\mathrm{std}(\hat{x})$ are the posterior mean and standard deviation estimated from the generated samples. Fig. 9 (left) shows that Z closely follows a standard normal for both survey directions, indicating that the posterior mean is approximately unbiased and the residual uncertainty is approximately Gaussian. The standard deviations slightly above unity indicate mild overconfidence (the model underestimates the true pixel-wise variance by $\sim 15\%$) and the effect is more pronounced for HSC-anchored reconstructions, consistent with the greater ambiguity in generating higher-resolution images from lower-resolution conditioning.

A.7 Normalizing Flow for Outlier Detection

To rank candidate outliers, we fit a normalizing flow to the empirical density of the embeddings in each latent space and use the negative log-likelihood $-\log p(z)$ as the anomaly score. We use a

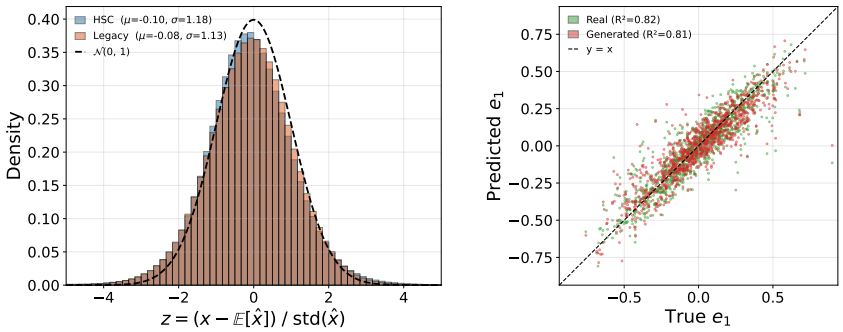


Figure 9: **Left:** Pixel-wise Z-score distribution of generated posterior samples relative to ground-truth target images, across all pixels and held-out galaxies. Both HSC-anchored and Legacy-anchored reconstructions closely approximate a standard normal distribution, indicating that the posterior mean is approximately unbiased, the posterior variance captures the true prediction error, and that the residual uncertainty is approximately Gaussian. The standard deviations slightly above unity indicate mild overconfidence, more pronounced for HSC-anchored reconstructions, consistent with the greater ambiguity in generating higher-resolution images from lower-resolution conditioning. **Right:** R^2 for galaxy ellipticity from a ResNet trained on real HSC images, evaluated on real versus counterfactual HSC images. Near-identical performance shows that existing HSC pipelines can be readily applied to Legacy images through our counterfactual model, extending HSC-quality analysis across the full Legacy footprint.

Neural Spline Flow (NSF; 29) implemented in `zuko` [30], with 6 coupling transforms and a 2-layer MLP of width 64 in each transform. The flow is unconditional (`context=0`) and is trained by maximum likelihood with Adam.

We use the embeddings from just the HSC images for this experiment. We fit a separate flow for each representation we score: (i) our physics latent flattened across the 4 spatial tokens (64D), (ii) our instrument latent mean-pooled across tokens (16D), and (iii) AION-1 (Base) 768D mean-pooled embedding. For AION-1 we also tried fitting the flow to embeddings reduced to 64D via PCA fit on the training split and obtained qualitatively similar outliers. For each space we use 80% of the $\sim$100,000 HSC galaxies to fit the flow and score the full sample under the trained model; the lowest-likelihood examples are the candidate outliers shown in Fig. 3. Table 6 lists the hyperparameters used.

Table 6: Normalizing flow architecture and training hyperparameters used for outlier detection.

Parameter	**Value**
Flow Architecture	
Family	Neural Spline Flow (NSF, `zuko`)
Number of transforms	6
Hidden features per transform	$[64, 64]$
Context dimension	0 (unconditional)
Input dimension	64 / 16 (Ours) / 768 (AION-1)
Optimization	
Loss	Negative log-likelihood
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	512
Epochs	50
Model selection	Lowest mean training NLL (best-state checkpoint)
Data Split	
Train fraction	80%
Scoring set	All $\sim$100,000 HSC galaxies
Random seed	42

A.8 Decoder Reliance on Physics and Instrument Conditioning

We perform an experiment to further test whether the decoder actually uses the two conditioning latents. For 256 held-out galaxies, we generate samples from the same initial noise realization x_0 under three conditions: "intact" conditioning (with the physics latent z_{phy} obtained from the same galaxy observed by the other survey, and the instrument latent z_{ins} obtained from the anchor's spatial neighbors in the target survey), z_{phy} replaced by that of a different held-out galaxy (which we will refer to as the "donor"), and z_{ins} replaced by the donor's instrument conditioning. Donors are drawn within the target survey, so the intervention is not a survey swap. Crucially, each donor also has a ground truth pair in the target survey, so we can test not only how the swap affects agreement with the target but also how the generation follows the donor. We average over 4 posterior samples per galaxy, test both surveys as targets, and report the results in Table 7.

To assess the dependence on physics conditioning, we use the reconstruction MSE as a proxy for source identity. To assess the dependence on instrument conditioning, we define σ_{ap} as a robust, MAD-based estimate of the standard deviation of adjacent-pixel differences, i.e. a proxy for the pixel-scale noise of an image. For each galaxy and conditioning setup, we have three values: $\sigma_{\mathrm{ap,gen}}$ for the generated image, $\sigma_{\mathrm{ap,own}}$ for the real anchor observation, and $\sigma_{\mathrm{ap,donor}}$ for the donor's observation in the target survey.

We evaluate the correlation ρ between these quantities across all the evaluated galaxies, separately for each target survey and condition, to test whose observing conditions the generation inherits. Here, $\rho(\sigma_{\mathrm{ap,gen}}, \sigma_{\mathrm{ap,own}}) \approx 1$ means that the generated noise level tracks its own target's as expected when running the model normally, $\rho(\sigma_{\mathrm{ap,gen}}, \sigma_{\mathrm{ap,donor}}) \approx 1$ means that it tracks the donor's, and $\rho \approx 0$ means little correlation across galaxies.

Table 7: Decoder use of physics and instrument conditioning. We compare intact conditioning with replacing either latent by that of a donor galaxy within the same target survey. MSE measures agreement with the real anchor (own) or donor observation. ρ measures how the generated image's pixel-scale noise correlates with the anchor's or donor's across galaxies. HSC and Legacy denote the target survey. Bold entries highlight agreement with the donor after the corresponding swap.

	Intact		Physics swapped		Instrument swapped	
Metric	HSC	Legacy	HSC	Legacy	HSC	Legacy
MSE vs. own target	0.058	0.186	0.970	1.490	0.060	0.186
MSE vs. donor target	0.970	1.480	**0.058**	**0.188**	0.970	1.480
$\rho(\sigma_{\mathrm{ap,gen}}, \sigma_{\mathrm{ap,own}})$	$+0.99$	$+0.94$	$+0.90$	$+0.72$	-0.03	$+0.17$
$\rho(\sigma_{\mathrm{ap,gen}}, \sigma_{\mathrm{ap,donor}})$	-0.11	-0.10	-0.01	$+0.10$	$\mathbf{+0.89}$	$\mathbf{+0.69}$

Shuffling z_{phy} changes the image to track the donor galaxy, as shown by the MSE against the donor's ground truth image closely matching the intact model's MSE against its own target (0.058/0.188 vs 0.058/0.186), while the noise statistics stay consistent with those of the original target (correlation goes from 0.99 / 0.94 to 0.90 / 0.72). This means our reconstruction matches the large-scale structure (mostly driven by source identity) of the donor galaxy, while the small scale noise structure continues to match that of the original target.

Shuffling z_{ins} has the opposite effect: source identity is untouched (the MSE difference is $0.002 \pm 0.002/0.000 \pm 0.005$, a null result), while the generated noise is different from the original target's instrument properties and tracks that of the donor ($\rho = -0.03/+0.17$ for the original target, and $\rho = +0.89/+0.69$ on the donor).

A.9 Ablation: Counterfactual Generative Objective vs. Contrastive Learning

To evaluate the role of the training objective, we train the same physics and instrument encoders with an InfoNCE contrastive loss [31] instead of our counterfactual generative objective.

For the physics encoder, positives pairs consist of an anchor galaxy image and its cross-survey pair, while negative pairs combine the anchor with images of other galaxies coming from a mixture of both surveys (all the other anchor images in the batch).

For the instrument encoder, positive pairs consist of an anchor and its sky neighbours observed by the same instrument, while negative pairs combine the anchor with neighbors of other anchors in the same batch. Following Audenaert et al. [11], we attach SimCLR projection heads [32] to both encoders and compute the contrastive losses on the projected embeddings.

Both models use the same encoder architectures and training dataset. A notable difference is that the contrastive baseline must pass the anchor image through the encoders to construct the training pairs, while in our model the anchor image is never an input to the encoders. All 5,469 galaxies used for downstream evaluation are excluded from the self-supervised training stage for both models.

We compare reconstruction quality and downstream predictive performance, and examine how the predictive power of models trained with each training objective respond to the addition of higher-resolution observations.

Reconstruction. We train a flow-matching decoder conditioned on frozen contrastive embeddings, using the same counterfactual triplet reconstruction logic as our model. This proves how well the learned representations support counterfactual reconstruction. Table 8 reports reconstruction MSE on 512 held-out validation galaxies.

Table 8: Counterfactual reconstruction MSE for our model and a decoder conditioned on frozen contrastive embeddings. Relative increases are measured against our model.

Conditioning style	**MSE HSC**	**MSE Legacy**	**MSE all**
Ours	0.0474 ± 0.0031	0.1610 ± 0.0047	0.1042
Frozen contrastive embeddings	0.0679 ± 0.0045	0.1936 ± 0.0071	0.1308
Relative MSE increase	**+43.1%**	**+20.3%**	**+25.5%**

Frozen contrastive embeddings obtain a reconstruction MSE 25.5% greater than our model. The relative increase is larger for HSC (43.1%) than for Legacy (20.3%). These results suggest that our generative objective learns representations better suited for counterfactual reconstruction, and that the largest relative benefit is for the higher-resolution targets, indicating that it does not suffer as much from the information asymmetry between the two surveys.

Downstream $\mathbf{R^2}$. We freeze both models and fit an MLP probe on each latent, using five-fold cross-validation over the 5,469 evaluation galaxies. Our physics latent achieves higher R^2 than the contrastive one on 6 of the 7 physics and morphology properties. For log-metallicity $logZ$, the two agree within the fold-to-fold scatter. Nonetheless, the contrastive instrument latents predict all five instrument properties better than ours, yet its physics latents retain even less information about instrument properties than ours does (Table 9).

We attribute this to what each objective requires. The contrastive loss punishes anything the anchor shares with its negatives. Dust extinction-induced color excess E(B-V) is a good example: nearby galaxies share it, so it gets pushed out of the physics latent even though both physics views have it, and pulled into the instrument latent, whose negatives come from other parts of the sky. Our proposed objective has no such pressure, so this information can be partially encoded in the physics latent space and these factors settle wherever the decoder needs them. Our physics latent space holds shared information, but not instrument-private information. The contrastive objective latent space discards shared information, and that costs it at generation time (Table 8).

Table 9: Downstream R^2 ($\times 100$) for embeddings trained using our counterfactual generative objective versus those trained using a contrastive learning baseline. (a) Each property is predicted from its matching latent: physics-related properties (left) from the physics latent and instrument properties (right) from the instrument latent. (b) Each property is predicted from the *other* latent, measuring how much of it is carried across the split.

Physics-related properties				**Instrument properties**			
Property	Ours	Contrastive	Δ	Property	Ours	Contrastive	Δ
(a) from the physics latent				*(a) from the instrument latent*			
z	$\mathbf{77.0 \pm 1.3}$	75.4 ± 1.5	$+1.5 \pm 0.3$	Galaxy depth	75.1 ± 1.1	$\mathbf{85.2 \pm 1.6}$	-10.0 ± 0.2
$\log M_\star$	$\mathbf{73.0 \pm 0.8}$	69.6 ± 1.2	$+3.4 \pm 0.6$	PSF depth	74.2 ± 1.3	$\mathbf{84.6 \pm 2.0}$	-10.4 ± 0.3
sSFR	$\mathbf{51.0 \pm 1.4}$	48.4 ± 1.9	$+2.5 \pm 0.7$	$N_{\rm obs}$	53.5 ± 1.4	$\mathbf{69.9 \pm 0.9}$	-16.4 ± 0.7
$t_{\rm age}$	$\mathbf{22.3 \pm 2.7}$	16.5 ± 2.0	$+5.8 \pm 0.9$	$E(B-V)$	59.1 ± 2.5	$\mathbf{77.0 \pm 1.7}$	-17.9 ± 0.9
$\log Z$	18.5 ± 3.3	19.1 ± 1.0	-0.6 ± 1.4	HSC PSF FWHM	50.0 ± 1.7	$\mathbf{68.4 \pm 1.8}$	-18.4 ± 0.6
e_1	$\mathbf{94.3 \pm 0.7}$	89.9 ± 1.1	$+4.4 \pm 0.4$				
e_2	$\mathbf{94.9 \pm 0.7}$	90.9 ± 1.3	$+4.0 \pm 0.4$				
(b) from the instrument latent				*(b) from the physics latent*			
z	63.4 ± 2.7	50.4 ± 1.9	$+13.0 \pm 1.6$	Galaxy depth	17.0 ± 1.0	-0.8 ± 0.6	$+17.8 \pm 0.3$
$\log M_\star$	61.0 ± 1.3	49.8 ± 2.5	$+11.2 \pm 1.4$	PSF depth	16.6 ± 0.9	-0.6 ± 0.7	$+17.2 \pm 0.2$
sSFR	41.7 ± 2.5	32.3 ± 1.4	$+9.4 \pm 0.9$	$N_{\rm obs}$	5.9 ± 1.3	-2.5 ± 0.7	$+8.4 \pm 0.3$
$t_{\rm age}$	11.3 ± 1.0	5.2 ± 1.9	$+6.1 \pm 0.7$	$E(B-V)$	17.6 ± 2.0	6.6 ± 2.0	$+11.0 \pm 1.0$
$\log Z$	17.7 ± 2.0	6.1 ± 1.9	$+11.6 \pm 1.1$	HSC PSF FWHM	7.0 ± 1.1	2.4 ± 0.8	$+4.6 \pm 0.3$
e_1	59.2 ± 1.2	39.4 ± 3.1	$+19.8 \pm 1.7$				
e_2	70.7 ± 3.1	49.8 ± 4.1	$+20.9 \pm 1.4$				

Predictive advantage of higher-resolution observations. To examine whether the learned representations retain fine detail visible only in the higher-resolution HSC survey, we compare downstream property prediction using physics latents extracted separately from HSC and Legacy images of the same galaxies. For each model and property, we compute $\Delta_{\rm survey} = R^2(z_{\rm phy,HSC}) - R^2(z_{\rm phy,Legacy})$, where $z_{\rm phy,HSC}$ and $z_{\rm phy,Legacy}$ are obtained by encoding only the HSC or Legacy image, respectively, and $R^2(z)$ denotes downstream prediction performance using features z. Thus, $\Delta_{\rm survey}$ measures how much more information about a property is recoverable from the sharper survey's view than from the shallower one's. An objective that keeps only what the two views share should thus give $\Delta_{\rm survey} \approx 0$.

Table 10: Survey differences in downstream property-prediction R^2. For each property, $\Delta_{\rm survey}$ is the HSC-only score minus the Legacy-only score using physics latents. The final column reports the difference between models (ours minus contrastive).

Property	**Ours**	**Contrastive**	**Difference**
Redshift z	$+0.015 \pm 0.002$	$+0.005 \pm 0.003$	$+0.009 \pm 0.004$
$\log M_\star$	$+0.028 \pm 0.009$	$+0.005 \pm 0.006$	$+0.023 \pm 0.010$
sSFR	$+0.013 \pm 0.004$	-0.002 ± 0.003	$+0.015 \pm 0.005$
Stellar age	$+0.005 \pm 0.014$	-0.007 ± 0.003	$+0.011 \pm 0.014$
$\log Z$	$+0.010 \pm 0.003$	$+0.007 \pm 0.004$	$+0.003 \pm 0.005$
Ellipticity e_1	$+0.026 \pm 0.003$	$+0.007 \pm 0.002$	$+0.019 \pm 0.004$
Ellipticity e_2	$+0.019 \pm 0.004$	$+0.010 \pm 0.002$	$+0.009 \pm 0.003$

Table 10 shows a larger HSC advantage for our model than for the contrastive baseline on all seven evaluated properties. The contrastive baseline is close to $\Delta_{\rm survey} \approx 0$, meaning its HSC and Legacy latents are close to interchangeable for the purposes of inference, which is what a collapse toward the lower-resolution view looks like, while ours retains $\Delta_{\rm survey} > 0$. We expect this advantage to grow as the resolution gap widens: for example, when combining ground- and space-based observations, where forcing agreement between views risks discarding precisely the fine-scale information of greatest scientific value.

A.10 Data Efficiency: Scaling with Cross-Matched Training Set Size

When applying our method to new surveys, the availability of cross-matched observations is the main constraint, not the amount of single survey data. Training requires observations of the same galaxy from both instruments to form the anchor and physics pair views. In contrast, the same-instrument conditioning neighbors need not have a counterpart in the other survey. For example, the model in the main text trains on $\sim$100,000 crossmatched anchors but draws its neighbours from a pool of $\sim$468,000 sources, most of which have no crossmatch.

To assess how performance depends on the available overlap, we train models on nested subsets of 1,000, 3,162, 10,000, and 31,622 cross-matched galaxies, while keeping the instrument-neighbor pool fixed at $\sim$468,000 sources (Table 11). This is the most realistic deployment scenario for a new survey, which might have a large pool of data but only limited overlap with previous surveys.

Because an untrained encoder already provides nonzero predictive performance, we measure the percentage of the full-data improvement above this baseline recovered at each training set size. For each property group, we define the fraction of full-data improvement recovered as

$$\eta(N) = \frac{\bar{R}^2_N - \bar{R}^2_{\text{untrained}}}{\bar{R}^2_{N_{\max}} - \bar{R}^2_{\text{untrained}}}$$

where $\bar{R}^2_N$ is the mean downstream score using N training crossmatches and $\bar{R}^2_{N_{\max}}$ is the full-data reference. This assigns 0% to the untrained encoder and 100% to the checkpoint reported in the main text.

Table 11: Downstream performance as a function of number of crossmatches in the training set. Columns show the mean R^2 across seven physics, five Legacy instrument, and two HSC instrument properties, with the percentage of improvement recovered in parentheses. The full-data model is the model used in the main text, its training split contained approximately 99,000 crossmatches. The instrument-neighbor pool is kept fixed across training subsets.

Training crossmatches	**Physics** (7 properties)		**Legacy instrument** (5 properties)		**HSC instrument** (2 properties)	
1,000	0.593	(81.2%)	0.289	(33.8%)	0.219	(26.8%)
3,162	0.612	(89.0%)	0.390	(55.9%)	0.282	(43.7%)
10,000	0.625	(94.3%)	0.471	(73.7%)	0.353	(62.7%)
31,622	0.634	(98.0%)	0.499	(79.8%)	0.375	(68.6%)
Full-data model	0.639	(100.0%)	0.591	(100.0%)	0.492	(100.0%)
Untrained encoder	0.394	(0.0%)	0.135	(0.0%)	0.119	(0.0%)

Table 11 shows that with just 1,000 crossmatches (approximately 1% of the full training set) the model recovers 81.2% of the full-data improvement in mean R^2 for physics properties, but only 27-34% for instrument properties. At 31,622 crossmatches, these fractions reach 98.0% and 69-80%, respectively. These results suggest that most of the performance in physics-related tasks can be recovered with as little as 10^3 cross-survey matches, while prediction of instrument properties continues to improve across the studied range all the way to 10^5.

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 abstract and introduction state our three main contributions: a counterfactual generative objective for disentangling physics from instrument factors, and the resulting downstream capabilities (parameter inference, outlier detection, counterfactual generation, instrument-independent retrieval, and zero-shot pipeline transfer). Each is supported by the experiments described in Section 4 and the appendices.

   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: A dedicated "Limitations and future directions" paragraph appears at the end of the main text, noting that our framework requires overlapping observations across instruments. We also acknowledge that the instrument latent retains residual physics information that we cannot quantify, and in Appendix A.6 we mention that the posterior pixel-wise variance is mildly underestimated ($\sim 15\%$).

   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: [N/A]

   Justification: The paper does not present formal theoretical results; the contributions are methodological and empirical.

   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: The methods section describes the dual-encoder architecture, counterfactual flow-matching objective, triplet construction, and spatial-neighbor pairing strategy. Appendix A.1 details the modified ResNet-18 encoder design and the UNet velocity network, and Appendix (Training Configuration) lists hyperparameters, optimizer, schedule, and number of training steps. The cross-matched Legacy–HSC dataset is constructed from publicly available surveys, and the preprocessing pipeline is fully described.

   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: We attach a simplified version of our codebase as supplemental material. The accompanying `README.md` and `new_dataset_guide.md` provides step-by-step instructions for downloading the cross-matched Legacy–HSC sample from its public source (the Multimodal Universe dataset [9]), preprocessing it, and training and evaluating the model. Both underlying galaxy imaging surveys (DESI Legacy Imaging Surveys and Hyper Suprime-Cam) are publicly accessible.

   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: Architecture, optimizer, hyperparameters, training schedule, and data splits are described in the methods section and the appendices on Model Architecture, Training Configuration, and the MLP probe used for downstream evaluation.

   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: Statistical significance is reported where it is naturally available: 1σ posterior uncertainty bands estimated from 32 different samples per anchor are shown across spatial frequencies and lag distances in Fig. 8, and pixel-wise Z-score distributions across all held-out galaxies are shown in Fig. 9 (left). The downstream regression R^2 values in Table 5 are point estimates from a single MLP probe trained per representation, with early stopping on a held-out validation set to control run-to-run variability. Given the size of the held-out evaluation set ($\sim$thousands of galaxies), the gaps between methods are substantially larger than the sampling-variability noise floor expected from a single R^2 estimate, and we do not expect multi-seed averaging to alter the qualitative ordering.

   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: As reported in the Training Configuration appendix, the model was trained using distributed data parallel across 4 NVIDIA H100 GPUs with *bfloat16* mixed precision, completing 75,000 training steps in approximately 8 hours of wall-clock time.

   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 research uses publicly available astronomical survey data and does not involve human subjects, sensitive data, or applications with foreseeable misuse potential.

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: [N/A]

Justification: Our work targets representation learning for astrophysical galaxy imaging from public surveys. There is no foreseeable direct path to negative societal impacts from disentangling instrument systematics in scientific imaging.

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 released code and any model checkpoints operate on astrophysical galaxy imaging from public surveys and pose no foreseeable risk of misuse.

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: We cite the original sources for all existing assets used: the DESI Legacy Imaging Surveys [16], Hyper Suprime-Cam [17], the Multimodal Universe dataset [9], AION-1 [8], the ResNet architecture [18], the `diffusers` `UNet2DConditionModel` implementation [19], and the flow-matching framework [14]. All datasets used are publicly available under their respective survey data-release policies.

    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: A simplified version of the codebase is submitted as supplemental material, with a `README` describing how to setup the dataset, train, and evaluate. The public code repository is linked in the first-page footnote.

    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: [N/A]

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: [N/A]

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 a component of the core methodology. Any use of LLMs was limited to writing, editing, or formatting and did not impact the scientific contributions of this work.

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.