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

Pablo Mercader-Perez
Massachusetts Institute of Technology
pablomer@mit.edu

Carolina Cuesta-Lazaro
Flatiron Institute, Simons Foundation
Institute for Advanced Studies
cuestalz@mit.edu

Daniel Muthukrishna
Massachusetts Institute of Technology
AstroAI, CfA | Harvard & Smithsonian
danmuth@mit.edu

Jeroen Audenaert
Massachusetts Institute of Technology
jeroena@mit.edu

V. Ashley Villar
Harvard University
The NSF IAIFI
ashleyvillar@cfa.harvard.edu

David W. Hogg
New York University
Flatiron Institute, Simons Foundation
david.hogg@nyu.edu

Marc Huertas-Company
Instituto de Astrofísica de Canarias
mhuertas@iac.es

William T. Freeman
Massachusetts Institute of Technology
billf@mit.edu

Abstract

Data collected from the physical world is always a combination of multiple sources: an underlying signal from the physical process of interest and a signal from measurement-dependent artifacts from the sensor or instrument. This secondary signal acts as a confounding factor, limiting our ability to extract information about the physics underlying the phenomena we observe and complicating heterogeneous or multi-instrument observations. We propose a deep learning framework that leverages overlapping observations, a dual-encoder architecture, and a counterfactual generation objective to disentangle these factors of variation. The resulting representations explicitly separate intrinsic signals from sensor-specific distortions and noise, and can be used for counterfactual view generation and more robust parameter inference under changing measurement conditions. We demonstrate the effectiveness of our approach on $\sim 100{,}000$ cross-matched galaxy images from the DESI Legacy Imaging Survey and the Hyper Suprime-Cam Survey. More broadly, the framework suggests a recipe for scientific multi-instrument pretraining: construct overlapping pairs, treat measurement 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. This signal cannot be directly

Conference on Physics and AI at Stanford University (PAI 2026).

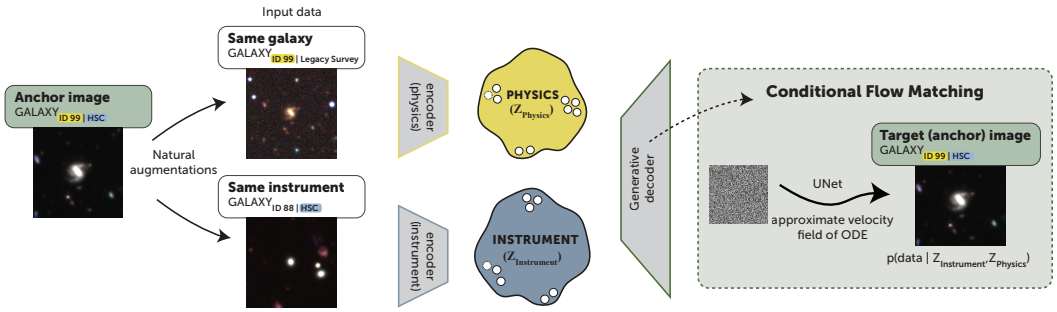

Figure 1: **Counterfactual reconstruction framework.** Training uses triplets consisting of an anchor observation, a same-source observation from a different instrument, and a same-instrument observation of a different source. The physics encoder extracts invariant physical structure, the instrument encoder extracts measurement artifacts, and the decoder reconstructs the withheld anchor counterfactually.

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 act as confounding factors, and are accompanied by stochastic noise and uncertainty. In other words, $\text{Observation} = f(\text{Signal}, \text{Instrument}) + \text{Noise}$. 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 learning to disentangle physics and instrument variables in settings where overlapping observations of the same physical system are available from different instruments. The key insight is that cross-matched observations form natural pairs: the same underlying physics observed under different instrument conditions. This provides a training signal for separating the two factors without requiring an explicit forward model of either instrument. We demonstrate the approach on astrophysical galaxy images, where the same galaxy observed by two different surveys can appear remarkably different because of resolution, noise, and calibration differences. 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.

2 Related Work

Recent work on disentangled representation learning has ranged from general latent-variable formulations such as β-VAE [1] to structured multi-view and causal perspectives in computer vision [2–4]. In astrophysics, recent foundation models have mostly treated different surveys as independent modalities [5], which is useful for downstream prediction but does not explicitly separate invariant physical content from survey-specific systematics. A recent multi-instrument time-series model moved in this direction by using contrastive learning to separate physics from instrument effects by leveraging overlapping observations [6]. Our approach uses the same overlap signal, but differs by making counterfactual generation the training objective that drives the properties of the learned representations. It is also related to data-driven noise modeling and likelihood-based treatments of non-Gaussian measurement noise in astrophysics [7, 8], but here the instrument is treated as a black box and its effects are learned implicitly from cross-matched observations.

3 Method

We present a counterfactual generation objective designed to disentangle physical properties from measurement systematics within learned representations. Our framework leverages large-scale scientific datasets that contain 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 from a different instrument, and an observation from the same instrument but of a 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 capture the underlying physical properties of the source and become invariant to 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 source content. The decoder combines these two latent embeddings to reconstruct an unseen target or anchor observation.

$$p(x_{j,k} \mid \{z_{\mathrm{phy}}(x_{j,k'})\}_{k' \neq k}, \{z_{\mathrm{ins}}(x_{j',k})\}_{j' \neq j}) \tag{1}$$

Here $x_{j,k}$ denotes source j observed by instrument k. The conditional distribution is modeled with flow matching [9]. This framework accommodates a variable number of source and instrument samples during conditioning. By aggregating these conditioning embeddings through an attention mechanism, the model naturally handles varying input counts. Crucially, the anchor image is never fed to the encoders; it is used only as the target for the reconstruction loss, forcing the model to generate the anchor counterfactually. Rather than learning embeddings first and attaching a separate generative model later, the counterfactual reconstruction objective learns the embeddings and the generative model jointly. This forces the latent spaces to retain invariant physical structure in the physics latents and survey-specific distortions in the instrument latents.

Compared to contrastive approaches, this generative objective has three practical advantages. First, it does not require hand-designed positive and negative pairs for each latent space, and it is less sensitive to the composition of a training batch. Second, counterfactual generation is the training objective itself, so the learned representations are optimized to support high-fidelity synthesis rather than requiring a separate generative model on top of the embeddings. Finally, the decoder and instrument encoder together act as a learned data-driven instrument model that maps latent observing conditions into pixel-level effects.

4 Galaxy Application and Results

We apply the model to a cross-matched galaxy dataset from the Multimodal Universe [10] containing $\sim 100{,}000$ matched images from the DESI Legacy Imaging Surveys [11] and the Hyper Suprime-Cam Survey [12]. HSC is a higher-resolution survey that detects significantly fainter features, while Legacy covers a much larger sky area, creating a trade-off between image quality and footprint. We align the shared $griz$ bands into a common four-channel representation, rescale zero-points and pixel scales, and crop both surveys to a common field of view. In this dual-survey setting, each galaxy pair provides the physics view. Likewise, up to five nearby galaxies (selected for their small angular separation from the anchor survey) provide the instrument views, exploiting the high spatial correlation of point spread functions (PSF), depth, and other instrumental noise properties. Both encoders use a ResNet-18 backbone, while the velocity network, operating as a flow-matching generative decoder, is implemented as a `UNet2DConditionModel` [13] that leverages cross-attention layers to flexibly integrate a variable number of conditioning inputs.

Counterfactual generation. Fig. 2 shows the model's ability to reconstruct target observations for both surveys and produce diverse posterior samples that capture uncertainty in the mapping. HSC is a deeper, higher-resolution instrument that resolves features that are blurred or undetected in Legacy imaging. Our model can therefore generate the expected HSC-like appearance for galaxies in the Legacy footprint, acting as a learned survey simulator that performs instrument-aware super-resolution and denoising.

Quantitatively, the posterior samples achieve an MSE of 0.081 for HSC-anchored and 0.197 for Legacy-anchored reconstructions over 256 held-out galaxies. To assess the quality of our generated images, we trained a ResNet on real HSC images to predict galaxy ellipticities. Then, we tested it on both real HSC images and synthetic HSC images made with our method. We find nearly identical accuracy on real and counterfactual HSC images ($R^2 = 0.82$ versus 0.81). Importantly, this shows that inference pipelines designed for HSC data transfer well through the counterfactual model, allowing to apply them to data from the Legacy Survey.

Disentanglement. We evaluate the separation of the latent spaces both qualitatively and quantitatively. Qualitatively, inspection of UMAP projections of the two latent spaces (not shown) confirms this separation: cross-survey pairs of the same galaxy map to adjacent locations in the shared physics

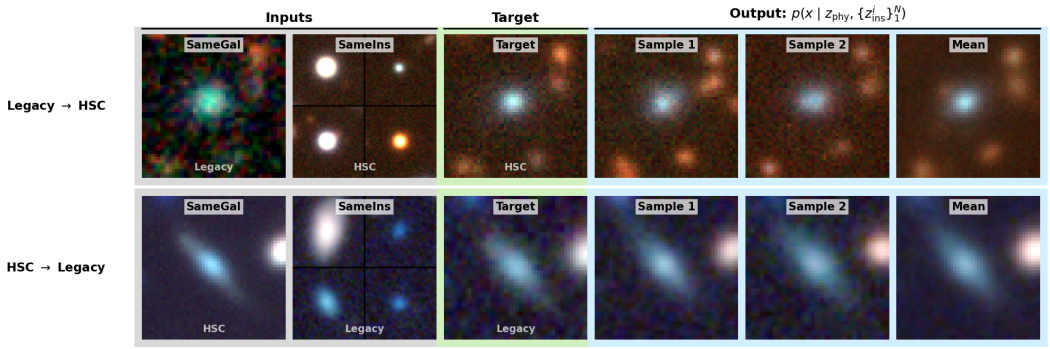

Figure 2: **Counterfactual multi-instrument galaxy reconstructions.** The model generates the withheld target observation by combining the intrinsic physical representation of the source (from a same-galaxy observation in the alternate survey) with the instrumental representation of the target survey (from images of a variable number of neighboring galaxies in the target survey). The posterior samples preserve the galaxy's main structure, while their stochasticity captures uncertainty.

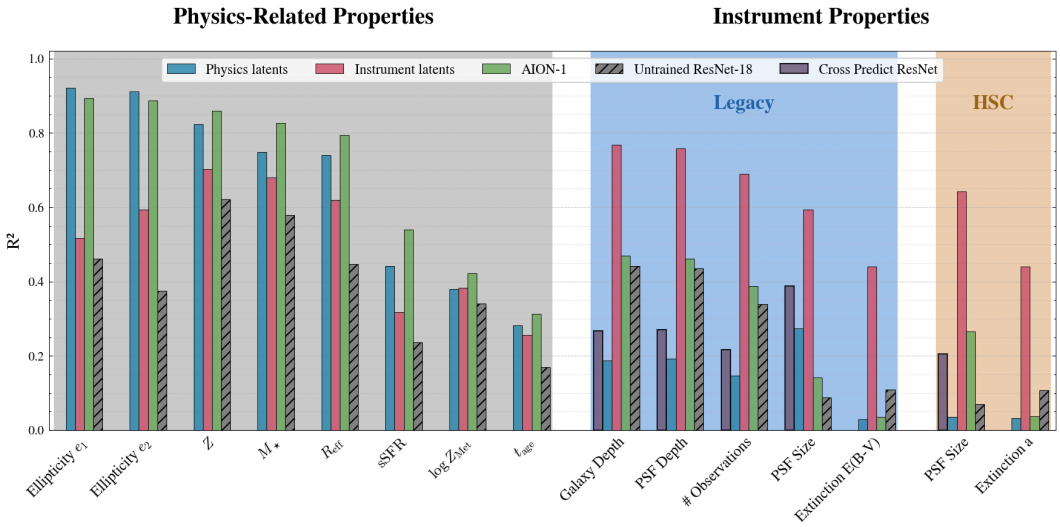

Figure 3: **Downstream regression on the disentangled latents.** Physics latents are strongest on physics-related properties, while instrument latents are strongest on survey-dependent observing conditions. The asymmetric split is the key evidence that the model separates intrinsic galaxy information from instrument systematics.

space, whereas the instrument space cleanly segregates the data by survey. To evaluate the information content quantitatively, we train a small MLP to predict physical and instrumental properties from the frozen embeddings extracted by each encoder. We compare against AION-1 embeddings, which treat the two surveys as independent modalities, and against a randomly initialized frozen ResNet-18 baseline. Fig. 3 shows the resulting regression R^2 per property. These physical properties are not perfectly "physics only": redshift changes apparent brightness, angular size, and observed colors, and morphology is also affected by the PSF. To separate genuine leakage from shared spatial structure, we additionally train a ResNet to predict each survey's instrument properties directly from the other survey's image; this cross-prediction baseline measures information that is inherently shared by the two surveys. 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. The physics latent space contains less information about instrument properties than the cross-prediction baseline and the randomly initialized ResNet baseline, indicating that it has actively erased instrument-related information. On the other hand,

the instrument latent space outperforms all baselines on instrument-related properties, highlighting the strength of our learned noise model.

5 Conclusion

We presented a counterfactual generative framework to disentangle intrinsic physical signals from instrument-specific artifacts. By framing measurement systematics as augmentations and employing a dual-encoder architecture, our model isolates domain-invariant physics without requiring explicit forward models or hand-engineered contrastive pairing. Applied to the DESI Legacy and HSC surveys, the resulting physics latents support robust downstream parameter inference, while the decoder acts as a data-driven instrument model capable of high-fidelity cross-survey translation. Beyond this two-instrument demonstration, our approach scales naturally to general multi-sensor scientific pretraining: leverage observational overlap, treat sensor effects as learned augmentations, and optimize for counterfactual generation.

References

[1] 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.

[2] Remi Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video, 2017. URL https://arxiv.org/abs/1705.10915.

[3] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1283–1292, 2017. doi: 10.1109/CVPR.2017.141. URL https://doi.org/10.1109/CVPR.2017.141.

[4] 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. URL https://doi.org/10.1109/JPROC.2021.3058954.

[5] Liam Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, et al. 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.

[6] 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.

[7] Ronan Legin, Alexandre Adam, Yashar Hezaveh, and Laurence Perreault-Levasseur. Beyond gaussian noise: A generalized approach to likelihood analysis with non-gaussian noise. *The Astrophysical Journal Letters*, 949(2):L41, 2023. doi: 10.3847/2041-8213/acd645. URL https://doi.org/10.3847/2041-8213/acd645.

[8] Ronan Legin, Maximiliano Isi, Kaze W. K. Wong, Yashar Hezaveh, and Laurence Perreault-Levasseur. Gravitational-wave parameter estimation in non-gaussian noise using score-based likelihood characterization, 2024. URL https://arxiv.org/abs/2410.19956.

[9] 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.

[10] 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.

[11] Arjun Dey et al. Overview of the desi legacy imaging surveys. *The Astronomical Journal*, 157(5):168, 2019. doi: 10.3847/1538-3881/ab089d. URL https://doi.org/10.3847/1538-3881/ab089d.

[12] Hiroaki Aihara et al. The hyper suprime-cam ssp survey: Overview and survey design. *Publications of the Astronomical Society of Japan*, 70(SP1), 2017. doi: 10.1093/pasj/psx066. URL https://doi.org/10.1093/pasj/psx066.

[13] 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.