DESIGNING CONTINUOUS CONDITIONING FOR GANS FROM WAE LATENT STRUCTURE

Pavlo Potapenko[1,*], **Sébastien Bompas**[1,*] **& Stefan Sandfeld**[1,2]
[1]Institute for Advanced Simulations: Materials Data Science and Informatics (IAS-9)
Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
[2]Faculty of Georesources and Materials Engineering
RWTH Aachen University, 52056 Aachen, Germany
{p.potapenko,s.bompas,s.sandfeld}@fz-juelich.de

ABSTRACT

Fast conditional generative models are used as surrogates in scientific workflows (e.g., parameter sweeps and inner-loop inference), but conditioning GANs on *continuous* scalar labels remains challenging: the label space is infinite, quantitative fidelity matters, and many conditioning mechanisms rely on conditional normalization incompatible with normalization-free backbones like R3GAN. We ask which conditioning mechanism is *structurally* appropriate for continuous labels under one-pass sampling constraints in a normalization-free GAN. Our approach uses Wasserstein autoencoders (WAEs) as a diagnostic tool: by analyzing the label-conditional aggregated posterior across eight datasets spanning scientific domains, we find that in many continuous-label settings, label variation aligns with simple, low-complexity directions in latent space well captured by feature-wise affine shifts and scales. Guided by this empirical structure and by R3GAN's constraints, we propose a lightweight FiLM-style conditioning module that maps a normalized scalar label to per-channel scale and shift parameters and injects them inside R3GAN bottleneck blocks, preserving normalization-free design while retaining single-network-evaluation sampling at inference. Across datasets, we link the latent label structure to practical conditioning choices and evaluate not only sample quality but also conditional fidelity and generalization to missing-label intervals. We show that FiLM-based modulation improves controllability and label interpolation compared with input concatenation, and that the same diagnostic predicts when simple affine conditioning is sufficient (e.g., label-interval hold-out) and when additional embedding capacity is required under non-monotone label–attribute links.

1 INTRODUCTION

Generative Adversarial Networks (GANs) provide a flexible framework for learning complex data distributions and synthesizing new samples, and have become a widely used tool in generative modeling (Goodfellow et al., 2014). In scientific settings, where data arise from experiments or from parameterized numerical simulations (e.g., PDE solution fields), generative models can serve as *surrogates* that enable rapid sampling of high-dimensional states and support downstream tasks such as uncertainty quantification and fast parameter sweeps (Zhu et al., 2019). A central design axis in such surrogates is *conditioning*, i.e., learning a distribution $p(x \mid y)$ over features x that generates samples corresponding to user-specified attributes or physical parameters, the condition y.

Continuous conditioning with scalar regression labels is significantly more complex than class conditioning, since the target space is infinite and label fidelity requires quantitative accuracy rather than category correctness. In scientific settings, the required accuracy is often high because small errors in y may correspond to meaningfully different physical regimes (Callen, 1985; Faye, 2011). At the same time, the label coverage is often sparse or uneven across the range, which can destabilize

[*]Equal contribution.

naive conditional GAN training and reduce controllability (Ding et al., 2020; Heyrani Nobari et al., 2021).

A strong candidate for this regime is $R_1 + R_2+$ RpGAN (R3GAN), a modern normalization-free GAN with one-step sampling and competitive image quality (Huang et al., 2024). In contrast, diffusion models typically sample iteratively and therefore require multiple network evaluations per sample (Ho et al., 2020; Dhariwal & Nichol, 2021a; Preechakul et al., 2022; Rombach et al., 2022; **?**). R3GAN's high image quality, one-step sampling, and fast inference are attractive for scientific surrogates, since such a model must generate high-quality samples faster than the original simulation pipeline.

Most continuous-conditioning GAN baselines inject label information through conditional normalization layers, including Conditional Batch Normalization (CBN) (Brock et al., 2019; Ding et al., 2020; Heyrani Nobari et al., 2021). R3GAN removes normalization layers for stability and simplicity (Huang et al., 2024), thereby constraining conditioning choices. In this work, we study continuous scalar conditioning under one-pass sampling constraints and a normalization-free GAN backbone, and show how Wasserstein autoencoder (WAE)-based latent diagnostics predict which conditioning pathway generalizes, especially under label gaps.

Before choosing a conditioning mechanism, we ask what label–data relationship the dataset exhibits. Here, $z \sim P_Z$ denotes latent noise and $y \in \mathcal{Y} \subset \mathbb{R}$ a continuous label (e.g., a parameter or age); the conditional generator produces samples via $x = G_\theta(z, y)$. We answer this diagnostic question using a WAE, because it provides an encoder–decoder decomposition and a latent space we can inspect. By analyzing the label-conditional aggregated posterior $Q_{Z|Y=y}$ across datasets, we obtain simple empirical models of how y changes latent codes. These observations motivate Feature-wise Linear Modulation (FiLM) (Perez et al., 2017) as a lightweight, normalization-free conditioning primitive compatible with R3GAN.

We first diagnose label-dependent latent structure with WAE (Tolstikhin et al., 2018), then design a matching conditioning mechanism for normalization-free R3GAN, and test controllability and generalization under label gaps and non-monotone labels.

Our contributions:

- We clarify how WAE- and GAN-type objectives relate in terms of induced convergence, and connect WAE and relativistic GAN divergences through a convergence hierarchy (Section 4, Appendix C).
- We use WAEs to diagnose label-aligned latent structure across eight datasets spanning scientific problems and natural images (Section 4.1).
- We introduce a normalization-free FiLM conditioning mechanism for R3GAN and show that it outperforms input concatenation in both sample quality and label fidelity (Sections 5 and 6).
- We study label-encoder capacity under FiLM, showing that simple (often effectively linear) embeddings can match or outperform stronger alternatives without auxiliary training, and we support this with a Procrustes trajectory analysis (Appendix I).
- We evaluate robustness beyond the standard setting by testing label-interval interpolation and a controlled non-monotonic failure mode (mirrored Ising) (Section 6).

2 Related Work

In the following we only highlight the utmost relevant contributions; a more in-depth overview can be found in Appendix A.

Conditioning and continuous labels. Conditioning can be applied at both the generator and discriminator levels, e.g., as in conditional GANs (cGANs) (Mirza & Osindero, 2014), with variants such as ACGAN (Odena et al., 2017) and projection discriminators (Miyato & Koyama, 2018). Generator conditioning is often implemented via conditional normalization or related feature-wise modulation (Dumoulin et al., 2017; Brock et al., 2019; Michalski et al., 2019); FiLM provides a normalization-free modulation primitive Perez et al. (2017). For continuous scalar labels under sparse coverage, Continuous Conditional GANs (CcGANs) use vicinal objectives (Ding et al., 2020),

while Performance Conditioned Diverse GAN (PcDGAN) targets inverse design with additional diversity/coverage losses (Heyrani Nobari et al., 2021). Continuous conditioning has also been explored for diffusion models (Zhao et al., 2024; Ding et al., 2025; Xie et al., 2026), but sampling is iterative (Ho et al., 2020; Ding et al., 2025) and can often be computationally expensive. A broader overview is given in Bourou et al. (2024).

Normalization-free backbone. We build on the normalization-free R3GAN baseline (Huang et al., 2024). Since many continuous-conditioning methods assume conditional normalization and/or modify the loss (e.g., CcGAN, PcDGAN, physics-informed GAN (PI-GAN)), we keep the R3GAN loss unchanged and instead design a normalization-free conditioning mechanism guided by a WAE-based diagnostic.

3 PRELIMINARIES

Notations. We write $x \sim P_X$ for real samples, $z \sim P_Z$ for latent noise, and $y \in \mathcal{Y} \subset \mathbb{R}$ for the continuous conditioning label. The conditional generator produces $x = G_\theta(z, y)$, and the critic/discriminator $D_\psi(x, y)$ scores how compatible an input x is with label y. For the latent-structure analysis, we train a Wasserstein autoencoder with encoder $Q_\phi(z \mid x)$; the label-conditional aggregated posterior $Q_{Z|Y=y}$ summarizes the distribution of encoded latents at label value y and is analyzed in Section 4.2. The complete measure-theoretic notation and conventions are given in Appendix B.

Conditioning and FiLM. The condition is a scalar label y or learned embedding $c(y)$. FiLM denotes feature-wise affine modulation, where a feature tensor h is mapped to $\gamma(y) \odot h + \beta(y)$ with label-dependent gain $\gamma(y)$ and bias $\beta(y)$ (Perez et al., 2017); $\odot$ is the element-wise product.

4 MOTIVATION: LINK BETWEEN WAE AND GAN

Our goal is to understand how a continuous semantic label y relates to systematic variation in the data distribution. A standard way to make such structure visible is to study a lower-dimensional representation, motivating the use of autoencoder (AE) models. The geometry and interpretability of an AE latent space depend strongly on both the training objective and architectural choices (Kächele et al., 2025). Since we use the learned latent structure to guide a conditioning mechanism for GANs, we aim to align the AE setup with the downstream GAN study along two axes: (i) the objective should induce a notion of convergence that is compatible with GAN training, and (ii) the model components should share inductive biases so that latent-space observations transfer to the generator.

Objective alignment. We choose WAE objective because its OT-relaxation induces a weak notion of convergence aligned with GAN training in a sequence sense; formal hierarchy in Appendix C.

Model-component alignment. Architecturally, a WAE decoder and a GAN generator both implement a map $z \mapsto x$ and can therefore share the same parameterization up to minor details. Likewise, a WAE encoder and a GAN discriminator share the direction $x \mapsto$ features; in practice, the discriminator backbone can be reused as an encoder by replacing the final scalar head with a d_Z-dimensional output (or distribution parameters in the stochastic case).

In realistic finite-capacity settings, the aggregated posterior learned by a WAE may not perfectly match the chosen prior; deterministic encoders can leave "holes" in latent space, while stochastic encoders can yield smoother latent structure (Rubenstein et al., 2018). We therefore evaluate both encoder variants in our latent diagnostics (details in Appendix E).

To connect label structure to the learned latent geometry, we study the label-conditional aggregated posterior

$$Q_{Z|Y=y}(A) := \int_{\mathcal{X}} Q_\phi(A \mid x)\, dP_{X|Y=y}(x)\,. \tag{8}$$

We then view this label-dependent latent distribution as the pushforward of simple base noise through a label-dependent map u,

$$u(\cdot, y)_{\#}\mathcal{N}(0, I) \approx Q_{Z|Y=y}\,, \tag{9}$$

i.e., $z_{\text{wae}} = u(z, y)$ with $z \sim \mathcal{N}(0, I)$. Studying the structure of u (or its learned approximation) provides a concrete way to characterize how labels organize latent variation and motivates transferring this structure to conditional GAN training.

4.1 WAE LATENT-SPACE DIAGNOSTICS FOR CONTINUOUS LABELS

We train WAE models on eight datasets: four scientific problems (Ising, Cahn–Hilliard, Kolmogorov flow, and Dots) and four natural-image benchmarks (MNIST, CIFAR-10, CIFAR-100, UTKFace). Dataset definitions and preprocessing are summarized in Appendix D, while WAE training details are available in Appendix E. Importantly, the WAE has no direct access to the label value during training, so any label-aligned structure in the latent space is emergent rather than enforced.

To motivate conditioning mechanisms for continuous labels, we quantify how a scalar label y organizes latent representations in a trained WAE. For each dataset, we encode samples into latent vectors $z_i \in \mathbb{R}^{dz}$ and ask whether the induced latent geometry supports an approximately ordered one-dimensional structure. We use three latent coordinates: latent radius r, geodesic coordinate g, and two-anchor coordinate s, together with global and local ordering statistics. Formal definitions are in Appendix E.2.

4.2 EMPIRICAL LABEL-CONDITIONAL STRUCTURE AND IMPLICATIONS FOR CONDITIONING

Full latent-space monotonicity diagnostics across all datasets are reported in Appendix E.3 (Table 7), together with PCA projections (Figs. 9–10). In the main text, we focus on the design implications. Across the scientific datasets, the WAE latents exhibit an approximately ordered organization w.r.t. the continuous label (often revealed more clearly after label-preserving latent averaging), whereas CIFAR-10/100 show no meaningful global ordering and UTKFace exhibits only a weak/heterogeneous ordering consistent with age being only one of several factors of variation.

A recurring pattern: affine modulation. Across datasets with strong global ordering, latent representations change with y in a manner that is well captured by label-dependent shifts and scalings. This behaviour is clearest for Dots, where $\rho(g, y)$ and $\rho(s, y)$ are close to 1, and the violation rate is near zero. A similar effect appears for Cahn–Hilliard after latent averaging, where $\rho(g, y)_{\text{avg}}$ and $\rho(s, y)_{\text{avg}}$ increase to values close to 1 while the local violation rate drops substantially. Kolmogorov exhibits the same qualitative trend, with $\rho(g, y)_{\text{avg}}$ and $\rho(s, y)_{\text{avg}}$ approaching 0.9 after averaging, although the violation rate remains moderate, which is consistent with a globally ordered coordinate and residual local mixing.

These trends motivate a simple parametric description in which a label-conditioned latent variable is generated from a base variable through a feature-wise affine transformation,

$$u(z, y) \approx \gamma(y) \odot z + \beta(y), \tag{1}$$

where $z \sim \mathcal{N}(0, I)$, $\gamma(y) \in \mathbb{R}^{dz}$ and $\beta(y) \in \mathbb{R}^{dz}$ vary smoothly with y, and $\odot$ denotes element-wise multiplication. Eq. 1 mirrors the core mechanism of FiLM, namely label-dependent per-channel gains and biases. Empirically, when the monotonicity diagnostics are strongest, as for Dots and averaged Ising, the fitted shift term $\beta(y)$ is close to linear in y (Appendix G).

Dataset-specific aspects. The Ising dataset differs from Dots and Cahn–Hilliard in that the strongest association with the label is captured by the latent radius, with $\rho(r, y)$ close to -1 and remaining large in magnitude after averaging. The geodesic correlations are also substantial, and $\rho(s, y)_{\text{avg}}$ becomes high after averaging, but the violation rate increases, which indicates that the global ordering coexists with stronger local inconsistencies. This pattern is consistent with label variation being expressed primarily through a magnitude-like factor rather than a single Euclidean axis and can be approximated as

$$u(z, y) \approx \Big(\gamma(y)\|z\|_2 + \beta(y)\Big) \frac{z}{\|z\|_2 + \varepsilon}, \tag{2}$$

with $\gamma(y), \beta(y) \in \mathbb{R}$ and a small $\varepsilon > 0$ for numerical stability. This does not change the main implication for conditioning, because a label-dependent shift and scale remain a natural mechanism for controlling either axis-aligned (Eq. 1) or magnitude-aligned (Eq. 2) degrees of freedom.

For natural-image benchmarks, the statistics are weaker. CIFAR-10 and CIFAR-100 show correlations close to zero and relatively high violation rates, which is expected because the class index does not define an ordinal semantic continuum. UTKFace is different because age is ordinal and yields higher correlations after averaging, but the violation rate remains high, which suggests that age is only one of several competing factors in the unsupervised representation. MNIST shows moderate correlations and a lower violation rate that decreases further after averaging, which is consistent with locally coherent class structure even though the digit index itself is not ordinal.

Design implication for conditioning in R3GAN. The main takeaway is that, on the scientific datasets we study, strong global ordering is accompanied by latent behaviour that is well described by feature-wise shifts and scalings, either along an axis-like coordinate or through a magnitude-like factor. We therefore adopt FiLM as the conditioning primitive in R3GAN.

5 PROPOSED GAN ARCHITECTURE

This section designs and evaluates conditioning mechanisms for continuous labels in R3GAN, guided by the latent-space diagnostics from our WAE study (Sec. 4.1). We keep all modifications consistent with the normalization-free R3GAN backbone (Huang et al., 2024), which uses ResNet and ConvNeXt-style blocks, and focus on one-pass conditional generation.

5.1 DESIGN CHOICES INFORMED BY LATENT SPACE DIAGNOSTICS

What the diagnostics suggest. Table 7 shows that several scientific datasets exhibit a measurable monotone organization of unsupervised WAE latents with respect to a continuous label, whereas MNIST and CIFAR-10/100 do not show meaningful global ordering. What matters is not the exact latent geometry, but that the effect of y can be captured by smooth, label-dependent shifts and scalings of intermediate representations.

FiLM as conditioning primitive. We therefore adopt feature-wise affine modulation as the conditioning primitive. Given an intermediate feature tensor $h \in \mathbb{R}^{C\times H\times W}$ and a normalized scalar label $y \in [0, 1]$, a FiLM layer applies

$$\mathrm{FiLM}(h; y) \;=\; \gamma(y) \odot h + \beta(y), \tag{3}$$

where $\gamma(y), \beta(y) \in \mathbb{R}^C$ are broadcast across spatial dimensions and $\odot$ denotes channel-wise multiplication (Perez et al., 2017). This introduces label-dependent gain and bias without using normalization layers.

Where to inject FiLM in R3GAN. We insert FiLM inside the generator bottleneck blocks, *between* the depthwise (spatial) and pointwise (channel) convolutions. Concretely, if h_{dw} denotes the depthwise output, we apply $\tilde{h}_{\mathrm{dw}} = \mathrm{FiLM}(h_{\mathrm{dw}}; y)$ before the pointwise projection and residual connection. This preserves the original R3GAN block structure while enabling label-dependent channel modulation after spatial aggregation. The visual representation is shown in Fig. 1.

How to embed the label. To map the scalar label to FiLM parameters, we first build a label embedding $c(y)$ and then predict $\gamma(y)$ and $\beta(y)$ via linear readouts. We test three choices: (i) `Id`: $c(y) = y$; (ii) `MLP`: $c(y) = \mathrm{MLP}(y)$ using one hidden ReLU layer (Heyrani Nobari et al., 2021); and (iii) Improved Label Input, `ILI`, a pretrained 5-layer ReLU embedding used in CcGAN (Ding et al., 2020). These options span different embedding capacities and reflect the spectrum in Table 7. Dots shows an almost one-dimensional progression in the WAE diagnostics, so we test `Id`. Cahn–Hilliard, Kolmogorov, and Ising show weaker monotone ordering, so we also test a small `MLP` embedding to capture mild nonlinear label–feature relations. UTKFace is a common continuous-conditioning benchmark where `ILI` has been reported to work well, so we include it as a baseline.

Given $c(y)$, each FiLM site predicts per-channel parameters through linear maps

$$\gamma(y) = W_\gamma c(y) + b_\gamma, \qquad \beta(y) = W_\beta c(y) + b_\beta.$$

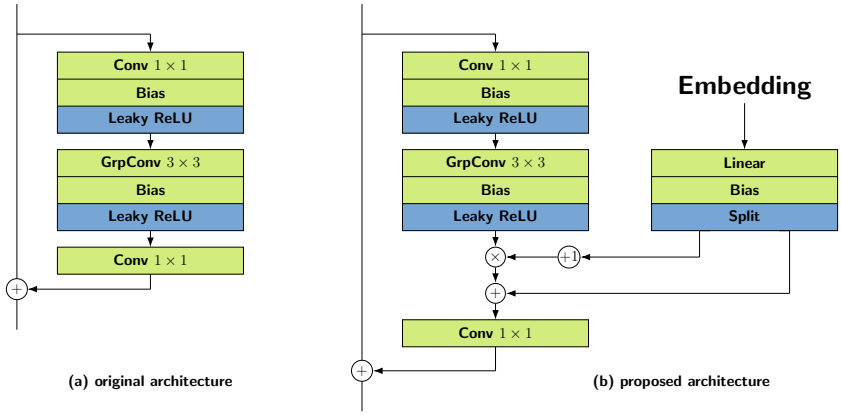

Figure 1: a) Schematical representation of R3GAN's inverted bottleneck. b) Proposed positioning of FiLM within the inverted bottleneck.

5.2 Baselines and comparisons

We compare FiLM against a standard Concat baseline, which injects a label embedding $c(y)$ at the generator input. In all configurations, the discriminator is conditioned through a projection head (Miyato & Koyama, 2018) that uses the same embedding option $c(y)$ to modulate the final score. For MLP, we instantiate separate embedding networks for the generator and discriminator and train them jointly with their respective models. For improved label input (ILI), we follow CcGAN and use a single pretrained embedding that is shared between generator and discriminator and kept fixed during R3GAN training (Ding et al., 2020).

6 GAN results

This section evaluates continuous-label conditioning in R3GAN. We emphasize Ising because it is the most challenging dataset in our suite from a representation-complexity perspective. The estimated intrinsic dimensionality is $d_{\mathrm{ID}} \approx 88$, which increases the difficulty of distribution learning (Chakraborty & Bartlett, 2025). We found Fréchet Inception Distance (FID) to be unreliable for scientific images, so we report label fidelity through label MSE and sample quality through PSD-MMD (Appendix J).

Experiment 1 (Concat vs. FiLM on Ising). We first compare conditioning mechanisms on Ising while keeping label-embedding capacity comparable. We train models with Concat using $c(y) \in \{\mathrm{MLP}, \mathrm{ILI}\}$ and with FiLM using $c(y) \in \{\mathrm{Id}, \mathrm{MLP}, \mathrm{ILI}\}$. Table 1 shows that FiLM substantially reduces PSD-MMD relative to Concat and also improves label MSE compared to the same embedding family. The best overall result is obtained by FiLM with ILI (6.051×10^{-4} MSE), while FiLM with MLP achieves the lowest PSD-MMD (2.008×10^{-3}).

Table 1: Conditioning mechanism comparison on Ising.

Model	Label MSE ↓	PSD-MMD ↓
Concat + MLP	1.243×10^{-3}	8.626×10^{-3}
Concat + ILI	8.026×10^{-4}	6.719×10^{-3}
FiLM + Id	1.031×10^{-3}	2.232×10^{-3}
FiLM + MLP	$\mathbf{6.501 \times 10^{-4}}$	$\mathbf{2.008 \times 10^{-3}}$
FiLM + ILI	$\mathbf{6.051 \times 10^{-4}}$	$\mathbf{2.138 \times 10^{-3}}$

Experiment 2 (choice of $c(y)$ under FiLM across datasets). We next fix FiLM and compare label embeddings $c(y) \in \{\text{ID}, \text{MLP}, \text{ILI}\}$ on Cahn–Hilliard, Dots, Kolmogorov, and Ising. Table 2 and the last row of Table 1 show that MLP achieves the lowest label MSE on all four datasets. PSD-MMD also favors MLP on Cahn–Hilliard and Ising, while ILI attains the lowest PSD-MMD on Dots and Kolmogorov.

Table 2: FiLM across datasets.

Dataset	$c(y)$	**Label MSE** $\downarrow$	**PSD-MMD** $\downarrow$
Cahn–Hilliard	Id	3.094×10^{-1}	3.468×10^{0}
Cahn–Hilliard	MLP	$\mathbf{1.012 \times 10^{-3}}$	$\mathbf{2.530 \times 10^{0}}$
Cahn–Hilliard	ILI	1.292×10^{-3}	3.522×10^{0}
Dots	Id	1.172×10^{-1}	8.215×10^{-1}
Dots	MLP	$\mathbf{2.421 \times 10^{-3}}$	2.275×10^{-1}
Dots	ILI	3.385×10^{-3}	$\mathbf{2.033 \times 10^{-1}}$
Kolmogorov	Id	2.304×10^{-1}	2.040×10^{0}
Kolmogorov	MLP	$\mathbf{5.118 \times 10^{-2}}$	1.928×10^{0}
Kolmogorov	ILI	1.132×10^{-1}	$\mathbf{1.756 \times 10^{0}}$

Experiment 3 (label-interval holdout on Ising). We remove a contiguous interval of Ising labels from training and evaluate only on that held-out interval. Table 3 shows that ILI and MLP achieve substantially lower label MSE than ID. Among the three, MLP yields the lowest PSD-MMD on the held-out interval.

Table 3: Ising label-interval holdout. Metrics are computed on the held-out label interval only. Lower is better. FiLM is the conditioning mechanism

$c(y)$	**Label MSE** $\downarrow$	**PSD-MMD** $\downarrow$
Id	3.859×10^{-3}	3.408×10^{-2}
MLP	$\mathbf{1.089 \times 10^{-4}}$	$\mathbf{1.114 \times 10^{-2}}$
ILI	$\mathbf{1.014 \times 10^{-4}}$	1.358×10^{-2}

Experiment 4 (Mirrored Ising with non-monotonic labels). Mirrored Ising makes the label–attribute relation non-monotonic by construction (Appendix D.5). Table 4 shows a large degradation for ID and a smaller degradation for ILI, while MLP remains accurate. MLP achieves the lowest label MSE and the lowest PSD-MMD on this non-monotonic setting.

Table 4: Mirrored Ising results. FiLM is the conditioning mechanism

$c(y)$	**Label MSE** $\downarrow$	**PSD-MMD** $\downarrow$
Id	1.812×10^{-1}	1.137×10^{0}
MLP	$\mathbf{8.265 \times 10^{-4}}$	$\mathbf{6.325 \times 10^{-3}}$
ILI	2.055×10^{-3}	1.942×10^{-2}

7 DISCUSSION

What the WAE analysis tells us. Across several continuous-label scientific datasets, the WAE diagnostics suggest that y primarily controls a low-dimensional (often near one-dimensional) direction in latent space. This motivates simple affine conditioning (Eqs. 1–2). Consistently, on Ising, replacing input concatenation with FiLM improves both sample quality and label fidelity, supporting conditioning via distributed modulation in normalization-free backbones.

Label encoder capacity under FiLM: linear often suffices. With FiLM fixed, a lightweight label pathway is frequently enough: MLP often matches or outperforms ILI without auxiliary training. A Procrustes trajectory analysis (Appendix I) further suggests that $c_{\mathrm{MLP}}(y)$ is effectively linear when the data–label relationship is monotone, deviating only in intrinsically non-monotonic regimes (mirrored Ising).

Non-monotonic labels and the role of nonlinearity. Mirrored Ising isolates a non-monotonic label–attribute mapping, where minimal embeddings fail (identity degrades sharply in both label MSE and PSD-MMD; Table 4). In contrast, MLP remains accurate and achieves the best PSD-MMD, indicating that nonlinearity is mainly needed when the conditional mapping is multi-valued in y.

Limitations and future work. First, the WAE-based latent analysis is descriptive and does not guarantee that $Q_{Z|Y=y}$ is globally one-dimensional or monotone on arbitrary datasets. Extending these diagnostics to additional scientific domains and to higher-dimensional parameters is a natural next step. Second, our evaluation relies on PSD-MMD and label MSE computed via a ResNet-34 regressor; developing protocols that better reflect physical constraints and downstream utility remains important. Third, we focus on FiLM as a normalization-free conditioning mechanism. Other normalization-free alternatives—such as low-rank hypernetworks generating per-layer modulation weights, gated residual adapters, or attention-based conditioning—may provide further gains while remaining compatible with the R3GAN backbone. Fourth, our empirical study targets moderately complex scientific datasets with low-dimensional conditioning variables. Benchmarking on substantially more complex physical systems (e.g., strongly multi-parameter regimes, sharp bifurcations, or highly nonlocal dependencies) could be a useful next step, as such settings may require richer label embeddings or more expressive conditioning mechanisms. Finally, the relativistic pairing loss in R3GAN could be extended to continuous labels using vicinal objectives or label-neighborhood pairing rules, which may improve label fidelity when labels are sparsely sampled.

8 Conclusion

We investigated continuous conditioning in normalization-free GANs through ablations on conditioning mechanisms and label encoders, complemented by latent-structure diagnostics. A WAE-based analysis indicates that, for several scientific datasets, label variation is often organized along a low-dimensional (frequently near one-dimensional) direction, motivating feature-wise affine conditioning. Consistent with this, replacing input concatenation with FiLM improves both sample quality and label fidelity, suggesting that injecting the conditional signal throughout the network is more effective than providing it only at the input.

A second takeaway concerns the label pathway itself. Across many settings, a lightweight label embedding is sufficient: MLP often matches or outperforms ILI while requiring no auxiliary training, and a Procrustes-based trajectory analysis (Appendix I) indicates that the learned conditioning behaves effectively linearly whenever the label–data relationship is monotone. Additional nonlinearity becomes important primarily in intrinsically non-monotonic regimes, as demonstrated by mirrored Ising, where minimal embeddings fail, and a small nonlinear map is needed for disambiguation.

Practical guideline. For continuous conditioning with normalization-free backbones, we recommend the following workflow: (i) use FiLM (feature-wise affine modulation) rather than input concatenation as the default conditioning mechanism; (ii) start with the simplest label encoder—a linear layer from y to modulation parameters (or, equivalently, a shallow MLP)—since it is cheap, stable, and often sufficient; (iii) increase label-encoder capacity only when behaviour indicates genuine conditional complexity (e.g., poor extrapolation, multimodal/non-monotonic label effects).

Author Contributions

P.P. and S.B. contributed equally to this work, each leading complementary parts of the project. P.P. was responsible for the formal analysis, investigation, visualization, writing of the original draft, and carried out the WAE implementation, analysis, and training. S.B. led the software implementation of the R3GAN models, their training, and contributed to writing through review and editing. S.S.

supervised the project, provided guidance throughout all stages of the research, and reviewed and edited the manuscript.

ACKNOWLEDGMENTS

This work was supported by computational resources provided by the Helmholtz Association within the framework of the Helmholtz Foundation Model Initiative (project SOL-AI).

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. URL https://arxiv.org/abs/1701.07875.

Anis Bourou, Valérie Mezger, and Auguste Genovesio. Gans conditioning methods: A survey, 2024. URL https://arxiv.org/abs/2408.15640.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

John W. Cahn and John E. Hilliard. Free energy of a nonuniform system. i. interfacial free energy. *The Journal of Chemical Physics*, 28(2):258–267, 1958. doi: 10.1063/1.1744102.

Herbert B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. John Wiley & Sons, 2 edition, 1985. ISBN 9780471862567.

Saptarshi Chakraborty and Peter L. Bartlett. On the statistical properties of generative adversarial models for low intrinsic data dimension, 2025. URL https://arxiv.org/abs/2401.15801.

Ting Chen, Mario Lucic, Neil Houlsby, and Sylvain Gelly. On self modulation for generative adversarial networks, 2019. URL https://arxiv.org/abs/1810.01365.

Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6069–6078, 2020. doi: 10.1109/CVPR42600.2020.00611.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b. URL https://arxiv.org/abs/2105.05233.

Xin Ding, Yongwei Wang, Zuheng Xu, William J. Welch, and Z. Jane Wang. Continuous conditional generative adversarial networks: Novel empirical losses and label input mechanisms. *arXiv preprint arXiv:2011.07466*, 2020.

Xin Ding, Yongwei Wang, Kao Zhang, and Z. Jane Wang. Ccdm: Continuous conditional diffusion models for image generation, 2025. URL https://arxiv.org/abs/2405.03546.

Ian L. Dryden and Kanti V. Mardia. *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons, Ltd., Chichester, UK, 2 edition, 2016. ISBN 9781119072492.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJO-BuT1g.

Martin Erdmann, Jonas Glombitza, Gregor Kasieczka, and Uwe Klemradt. *Deep Learning for Physics Research*. WORLD SCIENTIFIC, February 2021.

Guillaume Faye. An introduction to bifurcation theory (lecture notes), 2011. URL https://www.math.univ-toulouse.fr/~gfaye/ENS11/chap_bif.pdf. Definition 1.1.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Amin Heyrani Nobari, Wei Chen, and Faez Ahmed. Pcdgan: A continuous conditional diverse generative adversarial network for inverse design. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 606–616, New York, NY, USA, 2021. Association for Computing Machinery.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Nick Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The GAN is dead; long live the GAN! a modern GAN baseline. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31:253–258, 1925. doi: 10.1007/BF02980577.

Christian Jacobsen, Yilin Zhuang, and Karthik Duraisamy. Cocogen: Physically consistent and conditioned score-based generative models for forward and inverse problems. *SIAM Journal on Scientific Computing*, 47(2):C399–C425, 2025. doi: 10.1137/24M1636071. URL https://doi.org/10.1137/24M1636071.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, 2024. doi: 10.1109/CVPR52733.2024.00889.

Alexia Jolicoeur-Martineau. On relativistic f-divergences. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4931–4939. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/jolicoeur-martineau20a.html.

Fabian Kächele, Maximilian Coblenz, and Oliver Grothe. A comparison of latent space modeling techniques in a plain-vanilla autoencoder setting. *Machine Learning*, 114(7):151, May 2025. ISSN 1573-0565. doi: 10.1007/s10994-025-06784-3. URL https://doi.org/10.1007/s10994-025-06784-3.

T. Kadeethum, D. O'Malley, Y. Choi, H.S. Viswanathan, N. Bouklas, and H. Yoon. Continuous conditional generative adversarial networks for data-driven solutions of poroelasticity with heterogeneous material properties. *Computers & Geosciences*, 167:105212, 2022. ISSN 0098-3004. doi: https://doi.org/10.1016/j.cageo.2022.105212. URL https://www.sciencedirect.com/science/article/pii/S0098300422001613.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. doi: 10.1093/biomet/30.1-2.81.

Hiba Kobeissi and Emma Lejeune. Mechanical mnist - cahn-hilliard, 2022.

Hiba Kobeissi, Saeed Mohammadzadeh, and Emma Lejeune. Enhancing mechanical metamodels with a generative model-based augmented training dataset. *Journal of Biomechanical Engineering*, 144(12), August 2022.

W Lenz. Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern. *Z. Phys.*, 21: 613–615, 1920.

Vincent Michalski, Vikram Voleti, Samira Ebrahimi Kahou, Anthony Ortiz, Pascal Vincent, Chris Pal, and Doina Precup. An empirical study of batch normalization and group normalization in conditional computation, 2019. URL https://arxiv.org/abs/1908.00061.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByS1VpgRZ.

Binh Duong Nguyen, Pavlo Potapenko, Aytekin Demirci, Kishan Govind, Sébastien Bompas, and Stefan Sandfeld. Efficient surrogate models for materials science simulations: Machine learning-based prediction of microstructure properties. *Machine Learning with Applications*, 16:100544, 2024. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2024.100544. URL https://www.sciencedirect.com/science/article/pii/S2666827024000203.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2642–2651. PMLR, 2017. URL https://proceedings.mlr.press/v70/odena17a.html.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019. URL https://arxiv.org/abs/1903.07291.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2212.09748.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10619–10629, June 2022.

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

Paul K. Rubenstein, Bernhard Schoelkopf, and Ilya Tolstikhin. On the latent space of wasserstein auto-encoders, 2018. URL https://arxiv.org/abs/1802.03761.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, Mar 1966. ISSN 1860-0980. doi: 10.1007/BF02289451. URL https://doi.org/10.1007/BF02289451.

Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, José Miguel Hernández-Lobato, Richard E. Turner, and Emile Mathieu. On conditional diffusion models for pde simulations, 2024. URL https://arxiv.org/abs/2410.16415.

Michal Stepien, Carlos A.S. Ferreira, Seyedbehzad Hosseinzadehsadati, Teeratorn Kadeethum, and Hamidreza M. Nick. Continuous conditional generative adversarial networks for data-driven modelling of geologic co2 storage and plume evolution. *Gas Science and Engineering*, 115: 204982, 2023. ISSN 2949-9089. doi: https://doi.org/10.1016/j.jgsce.2023.204982. URL https://www.sciencedirect.com/science/article/pii/S2949908923001103.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018.

Nicholas Walker, Ka-Ming Tam, and Mark Jarrell. Deep learning on the 2-dimensional ising model to extract the crossover region with a variational autoencoder. *Scientific Reports*, 10(1), August 2020.

Yiwen Xie, Yang Peng, An Wang, Shuobei Sun, and Zhongyu Hou. Multivariate flow dynamics-conditioned diffusion for automated structural optimization of semi-filled micro gas chromatography columns. *Journal of Chromatography A*, 1765:466502, 2026. ISSN 0021-9673. doi: https://doi.org/10.1016/j.chroma.2025.466502. URL https://www.sciencedirect.com/science/article/pii/S0021967325008465.

Liu Yang, Dongkun Zhang, and George Em Karniadakis. Physics-informed generative adversarial networks for stochastic differential equations. *SIAM Journal on Scientific Computing*, 42 (1):A292–A317, 2020. doi: 10.1137/18M1225409. URL https://doi.org/10.1137/18M1225409.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization, 2019. URL https://arxiv.org/abs/1901.09321.

Yanxuan Zhao, Peng Zhang, Guopeng Sun, Zhigong Yang, Jianqiang Chen, and Yueqing Wang. Ccdpm: A continuous conditional diffusion probabilistic model for inverse design. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):17033–17041, Mar. 2024. doi: 10.1609/aaai.v38i15.29647. URL https://ojs.aaai.org/index.php/AAAI/article/view/29647.

Yinhao Zhu, Nicholas Zabaras, Phaedon-Stelios Koutsourelakis, and Paris Perdikaris. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *Journal of Computational Physics*, 394:56–81, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2019.05.024. URL https://www.sciencedirect.com/science/article/pii/S0021999119303559.

A Related Work

Conditioning of GANs. Conditioning can be applied at both the generator and discriminator levels. Mirza and Osindero introduce cGANs by providing the condition y to both the generator and the discriminator, often via concatenation or early-layer injection (Mirza & Osindero, 2014). ACGAN adds an auxiliary label-prediction head to the discriminator, which explicitly enforces label consistency (Odena et al., 2017), while projection discriminators condition the discriminator through an inner product between image features and a label embedding (Miyato & Koyama, 2018). On the generator side, many strong models inject class label y throughout the network via feature-wise modulation, often implemented through conditional normalization of various kinds: instance, batch, and group (CIN, CBN, CGN, respectively) (Dumoulin et al., 2017; Brock et al., 2019; Michalski et al., 2019). Arguably, the most commonly used is CBN, where a label embedding predicts per-channel scale and shift, as CBN is used as a part of the common backbone-BigGAN (Brock et al., 2019). FiLM abstracts the same scale-and-shift operation without relying on normalization layers (Perez et al., 2017). Related directions include self-modulation, where the generator predicts its own feature-wise modulation parameters from its internal state (Chen et al., 2019). Semantic image synthesis frameworks such as SPADE condition generation on dense segmentation maps via spatially adaptive normalization, which yields strong spatial control but is less aligned with our setting because it requires pixel-level conditioning signals (Park et al., 2019). For an extensive overview of conditioning strategies, we refer readers to the recent survey (Bourou et al., 2024).

Continuous conditioning with regression labels. CcGANs target continuous scalar regression labels under sparse and uneven label coverage; for this, Ding et al. (2020) propose vicinal objectives and label-input mechanisms. PcDGAN targets continuous conditioning for inverse design and introduces losses that improve diversity and coverage, combined with a simpler embedding architecture and CBN (Heyrani Nobari et al., 2021). Diffusion models have also been adapted for continuous conditioning (Zhao et al., 2024; Ding et al., 2025; Xie et al., 2026). Diffusion sampling remains iterative and therefore requires multiple network evaluations per sample, typically increasing sampling cost compared to one-step GAN generation Ho et al. (2020); Ding et al. (2025).

Generative surrogates for scientific workflows. Conditional generative models are used as surrogates in settings where fast sampling enables parameter sweeps, uncertainty quantification, and inverse design. Examples include GAN-based surrogates for poroelasticity and geologic CO_2 storage (Kadeethum et al., 2022; Stepien et al., 2023), physics-informed GAN variants for stochastic forward and inverse problems (Yang et al., 2020), and conditional score-based models for PDE problems (Shysheya et al., 2024; Jacobsen et al., 2025). This motivates our focus on continuous conditioning under one-pass sampling constraints.

Normalization-free backbone and our focus. Huang et al. propose R3GAN as a modern GAN baseline that removes normalization layers and reports strong FID with one-step sampling (Huang et al., 2024). Many continuous-conditioning GAN methods assume conditional normalization in the generator backbone; additionally, models such as CcGAN, PcDGAN, and PI-GAN modify standard loss terms for continuous conditioning (e.g., hard-, soft-, and singular- vicinal losses) or introduce additional loss terms (e.g., diversity or physics-informed loss terms) which conflicts with the R3GAN design. We therefore keep the R3GAN loss unchanged and instead focus on a normalization-free conditioning mechanism guided by a WAE-based diagnostic that summarizes the observed label-dependent latent structure.

FiLM-style modulation in diffusion models. Many diffusion architectures inject conditioning by predicting per-channel affine parameters and applying them to intermediate features, which is closely related to FiLM (Perez et al., 2017). In the class-conditional diffusion model of Dhariwal and Nichol, conditioning is introduced through adaptive normalization in residual blocks, where embeddings predict feature-wise scale and shift parameters that modulate normalized activations (Dhariwal & Nichol, 2021b). Beyond U-Nets, diffusion transformers use analogous mechanisms based on adaptive LayerNorm, where conditioning vectors produce scale and shift parameters (and often additional gates) applied inside transformer blocks (Peebles & Xie, 2023).

B NOTATION AND BASIC INFORMATION

The following section introduces notation in detail and summarizes the basic objects used throughout the paper.

Spaces and random variables. We write $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ for the data space with $\mathcal{B}_{\mathcal{X}}$ the (Borel) σ-algebra on $\mathcal{X}$, and analogously $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ for the latent space. Random variables are written in uppercase, for example $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$, and realizations are written in lowercase, for example x and z. The label space is $\mathcal{Y} \subseteq \mathbb{R}$ unless stated otherwise, and $Y \in \mathcal{Y}$ denotes a continuous label.

Distributions and conditioning. P_X denotes the data distribution on $\mathcal{X}$, P_Z denotes the latent prior on $\mathcal{Z}$, and $P_{X,Y}$ denotes the joint distribution. Conditional distributions are written as $P_{X|Y=y}$ and $P_{Y|X=x}$. Expectations are written as $\mathbb{E}_{X\sim P_X}[\cdot]$, with subscripts omitted when the distribution is clear.

Pushforward measures. For a measurable map T and a distribution P, the pushforward is written as $T_{\#}P$. For a generator $G_\theta : \mathcal{Z} \to \mathcal{X}$, the induced model distribution is $P_{G_\theta} := (G_\theta)_{\#}P_Z$.

Models. The generator is G_θ, parameterized by θ. The discriminator or critic is $D_\psi : \mathcal{X} \to \mathbb{R}$, parameterized by ψ. The encoder of a WAE is a Markov kernel $Q_\phi(z \mid x)$ on $\mathcal{Z}$, parameterized

by ϕ. The aggregated posterior induced by Q_ϕ and P_X is denoted by $Q_{Z,\phi}$. The label-conditional aggregated posterior is denoted by $Q_{Z|Y=y}$.

Divergences and convergence ordering. W_1 denotes the 1-Wasserstein distance on $\mathcal{P}(\mathcal{X})$ induced by the chosen ground cost. D_f denotes an f-divergence, and D_f^{Sy}, D_f^{Rp}, and D_f^{Ra} denote the symmetric, relativistic pairing, and relativistic average divergences, following Jolicoeur-Martineau (2020). For two divergences D_1 and D_2, the relation $D_1 \preceq D_2$ means that convergence in D_2 implies convergence in D_1 for any limit distribution. The strict relation $D_1 \prec D_2$ means $D_1 \preceq D_2$ and the converse implication fails for at least one sequence. Appendix C states the precise definitions and assumptions.

WAE objective. $\mathcal{L}_{\mathrm{WAE},\lambda}(P_X, P_{G_\theta})$ denotes the penalized WAE objective value with penalty weight $\lambda > 0$ (Tolstikhin et al., 2018). The objective is minimized over encoders and combines a reconstruction term with a prior-matching penalty. When the arguments are clear, we write $\mathcal{L}_{\mathrm{WAE},\lambda}$.

R3GAN objective. R3GAN trains a generator G_θ and a critic D_ψ using a relativistic pairing loss and adds zero-centered gradient penalties on both real and generated samples (Huang et al., 2024). We define the pairing term and penalties as

$$\mathcal{L}_{\mathrm{Rp}}(\theta, \psi) := \mathbb{E}_{z\sim P_Z,\, x\sim P_X}\Big[f\big(D_\psi(G_\theta(z)) - D_\psi(x)\big)\Big], \tag{4}$$

$$R_1(\psi) := \frac{\gamma}{2}\,\mathbb{E}_{x\sim P_X}\big[\|\nabla_x D_\psi(x)\|_2^2\big], \tag{5}$$

$$R_2(\theta, \psi) := \frac{\gamma}{2}\,\mathbb{E}_{z\sim P_Z}\big[\|\nabla_x D_\psi(G_\theta(z))\|_2^2\big], \tag{6}$$

where $\gamma > 0$ and $f(t) = -\log(1 + \exp(-t))$. Training alternates minimization of

$$\mathcal{L}_D(\psi; \theta) := -\mathcal{L}_{\mathrm{Rp}}(\theta, \psi) + R_1(\psi) + R_2(\theta, \psi), \tag{7}$$

$$\mathcal{L}_G(\theta; \psi) := \mathcal{L}_{\mathrm{Rp}}(\theta, \psi). \tag{8}$$

For conditional generation, G_θ and D_ψ take y as an additional input through the chosen conditioning mechanism.

C Divergence-Based Convergence Hierarchy

C.1 Setting and model components

Let $(\mathcal{X}, d)$ be a compact metric space with Borel σ-algebra $\mathcal{B}_\mathcal{X}$. Let $\mathcal{P}(\mathcal{X})$ denote the set of Borel probability measures on $\mathcal{X}$. The data distribution is $P_X \in \mathcal{P}(\mathcal{X})$.

Let $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ be a latent space with a fixed prior $P_Z \in \mathcal{P}(\mathcal{Z})$ (e.g. $P_Z = \mathcal{N}(0, I)$).

Generator. A measurable map $G_\theta : \mathcal{Z} \to \mathcal{X}$ induces the model distribution $P_{G_\theta} := (G_\theta)_\# P_Z \in \mathcal{P}(\mathcal{X})$.

Critic. A measurable function $C : \mathcal{X} \to \mathbb{R}$ (in theory optimized over all measurable C) is used to define adversarial objectives.

Encoder. An encoder is a Markov kernel $Q_\phi(\cdot \mid x)$ on $\mathcal{Z}$ given $x \in \mathcal{X}$. Its aggregated posterior is

$$Q_{Z,\phi}(A) := \int_\mathcal{X} Q_\phi(A \mid x)\, dP_X(x), \qquad A \in \mathcal{B}_\mathcal{Z}.$$

C.2 Divergences and the WAE functional

We use two (standard) notions of discrepancy: (i) optimal transport (OT) / Wasserstein distance on $\mathcal{X}$, (ii) GAN-type variational divergences built from a concave function f. To avoid ambiguity, we reserve φ for classical Csiszár f-divergences and f for the concave GAN loss used below.

C.2.1 WASSERSTEIN DISTANCE

For $P, Q \in \mathcal{P}(\mathcal{X})$ define the 1-Wasserstein distance

$$W_1(P,Q) := \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} d(x,x')\, d\pi(x,x'),$$

where $\Pi(P,Q)$ is the set of couplings with marginals P, Q.

C.2.2 CLASSICAL (CSISZÁR) f-DIVERGENCE

For a convex $\varphi : (0,\infty) \to \mathbb{R}$ with $\varphi(1) = 0$, define

$$D_\varphi(P\|Q) := \int_{\mathcal{X}} \varphi\left(\frac{dP}{dQ}\right) dQ \qquad (P \ll Q).$$

C.2.3 SYGAN / RELATIVISTIC GAN DIVERGENCES (CONCAVE f)

Following Jolicoeur-Martineau (2020), fix a concave function $f : \mathbb{R} \to \mathbb{R}$ satisfying the minimal conditions

$$\begin{aligned} &f(0) = 0, \quad f \text{ differentiable at } 0, \quad f'(0) \neq 0, \\ &\sup_{t\in\mathbb{R}} f(t) =: M > 0, \quad \arg\sup_{t\in\mathbb{R}} f(t) > 0. \end{aligned} \tag{9}$$

For distributions P, Q with common support, define:

$$D_f^{\mathrm{Sy}}(P,Q) := \sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{x\sim P}\big[f(C(x))\big] + \mathbb{E}_{y\sim Q}\big[f(-C(y))\big],$$

$$D_f^{\mathrm{Rp}}(P,Q) := \sup_{C:\mathcal{X}\to\mathbb{R}} 2\,\mathbb{E}_{x\sim P,\, y\sim Q}\big[f(C(x) - C(y))\big],$$

$$D_f^{\mathrm{Ra}}(P,Q) := \sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{x\sim P}\Big[f\big(C(x) - \mathbb{E}_{y\sim Q}C(y)\big)\Big] + \mathbb{E}_{y\sim Q}\Big[f\big(\mathbb{E}_{x\sim P}C(x) - C(y)\big)\Big].$$

Under (9), the relativistic objectives define statistical divergences (i.e. are ≥ 0 and equal to 0 iff $P = Q$); see Theorem 3.1 in Jolicoeur-Martineau (2020).

C.2.4 WAE OBJECTIVE AS A PENALIZED OPTIMAL TRANSPORT (OT) RELAXATION

A key OT identity for deterministic decoders (Theorem 1 in Tolstikhin et al. (2018)) states:

$$W_1(P_X, P_{G_\theta}) = \inf_{Q(\cdot|x):\ Q_Z = P_Z} \mathbb{E}_{x\sim P_X,\, z\sim Q(\cdot|x)}\big[d\big(x, G_\theta(z)\big)\big],$$

where Q_Z is the aggregated posterior induced by Q.

The WAE objective is obtained by relaxing the hard constraint $Q_Z = P_Z$ with a penalty D_Z on $\mathcal{P}(\mathcal{Z})$:

$$\mathcal{L}_{\mathrm{WAE},\lambda}(G_\theta) := \inf_{\phi} \left\{\mathbb{E}_{x\sim P_X,\, z\sim Q_\phi(\cdot|x)}\big[d\big(x, G_\theta(z)\big)\big] + \lambda\, D_Z\big(Q_{Z,\phi}, P_Z\big)\right\},$$

where $D_Z(\cdot,\cdot) \geq 0$ and $D_Z(Q_Z, P_Z) = 0 \iff Q_Z = P_Z$.

C.3 CONVERGENCE INDUCED BY A DIVERGENCE AND WEAKNESS PREORDER

Let D be any nonnegative functional on $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ with $D(P,Q) = 0 \iff P = Q$ on its domain. We write $P_n \xrightarrow{D} P$ if $D(P_n, P) \to 0$.

Weakness (sequence-based). We follow the sequence-based definition in Def. 3.2 Jolicoeur-Martineau (2020): for divergences D_1, D_2 we say D_1 is *weaker* than D_2 if

$$D_2(P_n, P) \to 0 \implies D_1(P_n, P) \to 0,$$

and the converse implication fails in general. We say D_1 is a *weakest* distance if $D_1(P_n, P) \to 0 \iff P_n \Rightarrow P$ (convergence in distribution).

C.4 HIERARCHY: WAE, WASSERSTEIN, SYGAN, RPGAN, RAGAN

Theorem C.1 (Wasserstein is weakest; relativistic GAN divergences are stronger)**.** *Assume $(\mathcal{X}, d)$ is compact and let f satisfy (9). Then:*

1. *W_1 metrizes weak convergence on $\mathcal{P}(\mathcal{X})$; in particular W_1 is a* weakest *distance.*
2. *The GAN divergences satisfy the strict weakness chain*

$$W_1 \preceq D_f^{\mathrm{Sy}} \preceq D_f^{\mathrm{Rp}} \preceq D_f^{\mathrm{Ra}},$$

with converses failing in general.

Justification and references. Item (1) is classical (Kantorovich-Rubinstein / weak-$*$ topology on compact $\mathcal{X}$); see e.g. the discussion in the Wasserstein GAN (WGAN) supplementary material (Arjovsky et al., 2017) and references therein. Item (2) is exactly Theorem 3.2 of Jolicoeur-Martineau (2020). □

Proposition C.2 (WAE is no stronger than Wasserstein)**.** *Assume $D_Z(Q_Z, P_Z) = 0 \iff Q_Z = P_Z$. Then for any fixed G_θ and $\lambda > 0$,*

$$\mathcal{L}_{\mathrm{WAE},\lambda}(G_\theta) \leq W_1(P_X, P_{G_\theta}),$$

hence $\mathcal{L}_{\mathrm{WAE},\lambda}$ is weaker than W_1 as a convergence criterion.

Proof. In the constrained OT identity above, restrict to encoders with $Q_Z = P_Z$. For such encoders, the penalty term vanishes, so the penalized infimum is bounded above by the constrained infimum, which equals $W_1(P_X, P_{G_\theta})$. □

Final convergence hierarchy. Combining Proposition C.2 and Theorem C.1 yields:

$$\begin{aligned} D_f^{\mathrm{Ra}}(P_n, P) \to 0 \quad &\Rightarrow \quad D_f^{\mathrm{Rp}}(P_n, P) \to 0 \\ &\Rightarrow \quad D_f^{\mathrm{Sy}}(P_n, P) \to 0 \\ &\Rightarrow \quad W_1(P_n, P) \to 0 \\ &\Rightarrow \quad \mathcal{L}_{\mathrm{WAE},\lambda}(G_n) \to 0. \end{aligned} \tag{10}$$

whenever the objects are well-defined (common support for the GAN divergences, and fixed decoder family for the WAE objective).

Remark (about classical f-divergences). Many standard GAN losses correspond (via variational representations) to classical Csiszár f-divergences (e.g. Jensen-Shannon in the original GAN). In compact domains, such divergences typically induce topologies stronger than weak convergence, hence are stronger than W_1; the WGAN supplementary material (Arjovsky et al., 2017) provides explicit comparisons for common choices (KL/JS/TV vs. Wasserstein).

D SCIENTIFIC DATASETS DETAILS

For the above experiments and investigations, four datasets were used that cover domains of statistical physics, solid mechanics, fluid dynamics, and biology. In the following, we introduce the theoretical background and a particular application field. All scientific datasets were produced by the authors and are available under an open-access license (CC-BY-NC-SA 2.0).

Label retrieval for evaluation. Several labels are defined through nontrivial simulation-side quantities, so evaluating label fidelity of generated samples requires a consistent image-to-label map. We therefore train a ResNet-34 regressor $\widehat{y} = R_\eta(x)$ on real data for each dataset and use $\widehat{y}$ as the label estimate for both real and generated images when reporting label-consistency metrics (He et al., 2016). This removes dataset-specific hand-crafted estimators and yields a uniform evaluation protocol across all datasets.

D.1 CAHN-HILLIARD DATASET

The Cahn-Hilliard dataset (Fig. 2) is commonly used in the solid mechanics and biomechanical engineering community as a benchmark dataset (see, (Kobeissi & Lejeune, 2022; Kobeissi et al., 2022)).

The microstructure in the Cahn-Hilliard model represents two phases, such as the different chemical elements in an alloy. Its evolution is governed by a set of partial differential equations (Cahn & Hilliard, 1958).

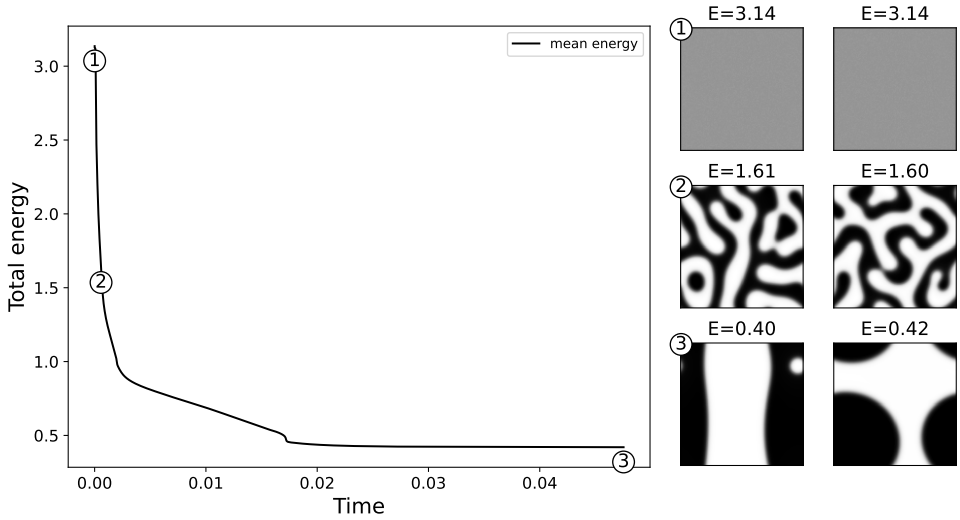

Figure 2: Evolution of the free energy against time. The images on the right show the microstructural evolution in 2 different simulations. The high energy state at a solid solution (1) quickly induces the formation of medium-sized features (2), which require a much longer time to merge into the bigger arrangement.

As opposed to the Ising dataset, the Cahn-Hilliard system is fully deterministic. The Cahn-Hilliard equations are given by:

$$\frac{\partial c}{\partial t} = M_c \nabla^2 \frac{\delta E}{\delta c}, \tag{11}$$

with the free energy E and a mobility coefficient of the interface M_c. The free energy density ψ consists of terms for the potential and gradient energy density:

$$\psi = \psi^{\text{bulk}} + \psi^{\text{grad}} \tag{12}$$

where $\psi^{\text{bulk}} = c_0 c^2 (1-c)^2$ and $\psi^{\text{grad}} = \frac{1}{2} k_c |\nabla c|^2$. The two constants c_0 and k_c are the density scale and the gradient energy density, respectively. The energy functional is then

$$E = \int_\Omega \psi \, \mathrm{d}\Omega = \int_\Omega c_0 c^2 (1-c)^2 + \frac{1}{2} k_c |\nabla c|^2, \mathrm{d}\Omega \tag{13}$$

To solve the above equation finite element simulations were performed. The initial random microstructure resembles a solid solution, where 2 phases are evenly mixed with an average concentration of 0.5 and small random perturbations or 0.02 of the field. The system is then subjected to the process of free energy minimization, which induces the phase separation and formation of big and smooth features. The label, in this case, is the free energy value of the system. To ensure the uniform distribution of the target label, the simulations are performed with a very small time step of 10^{-6}, but the field is only stored if it has a sufficient difference in free energy compared to the previously stored snapshot. Overall, 120 simulations were conducted with a total of 64k images with 128×128 pixels per each.

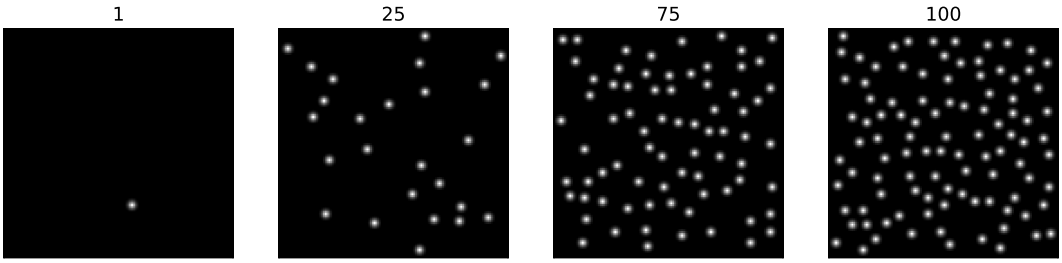

Figure 3: Dots dataset: examples of images at different dots counts.

D.2 DOT DATASET

The Dot dataset is generated by populating a black canvas with randomly positioned Gaussians. The width of a Gaussian is set to 5 pixels. The Gaussians are consequently added one after another. Before sampling a random position, the algorithm estimates the available space based on a condition of no covering, and only up to 50% overlap of Gaussians is allowed.

The number of Gaussians ranges from 1 to 100 on the image. The dataset contains 50k images of 128×128 pixel. The image examples can be seen on Fig. 3.

D.3 KOLMOGOROV DATASET

The so-called Kolmogorov simulation in computational fluid dynamics (CFD) is designed to study turbulence by modeling how energy cascades through different scales of motion within a fluid. The fundamental equations governing these simulations are the Navier-Stokes equations, which describe the motion of fluid substances. In the following, we briefly describe the theory and important aspects of the numerical implementation of Kolmogorov simulations.

The Navier-Stokes equations for an incompressible fluid are given by:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p + \nu \nabla^2 \mathbf{u} + \mathbf{f} \tag{14}$$

$$\nabla \cdot \mathbf{u} = 0 \,, \tag{15}$$

where $\mathbf{u}$ is the velocity field, t is the time, ρ is the fluid density, p is the pressure field, ν is the kinematic viscosity, and $\mathbf{f}$ is an external forcing function. In Kolmogorov flow simulations, the forcing function is typically chosen to induce a specific type of motion. For this dataset, a sinusoidal forcing function was chosen, which drives the flow and helps sustain turbulence. Unlike the Ising model, the Kolmogorov (as well as the Cahn-Hilliard) models are deterministic models. To increase the variance of the data, the sinusoidal forcing function was superimposed with small, random fluctuation sampled from a uniform distribution. The simulation was numerically implemented based on Fourier transformations, which is a convenient way of computing derivatives and, at the same time incorporating periodic boundary conditions.

Based on the velocity field, one can then compute the vorticity ω field (examples in Fig. 4):

$$\omega = \nabla \times \mathbf{u} \,. \tag{16}$$

The total kinetic energy in 2D was used as a continuous label; it is computed by

$$KE = \int_V \frac{1}{2} \left(u^2 + v^2\right) \, dV \,. \tag{17}$$

We ran a total of 250 simulations and selected $63,3$ thousand snippets, ensuring the dataset label distribution was as close as possible to uniform.

D.4 ISING MODEL

The 2D Ising model is a mathematical model (Ising, 1925; Lenz, 1920) used in statistical mechanics and theoretical physics to understand phase transitions, particularly in ferromagnetic materials but

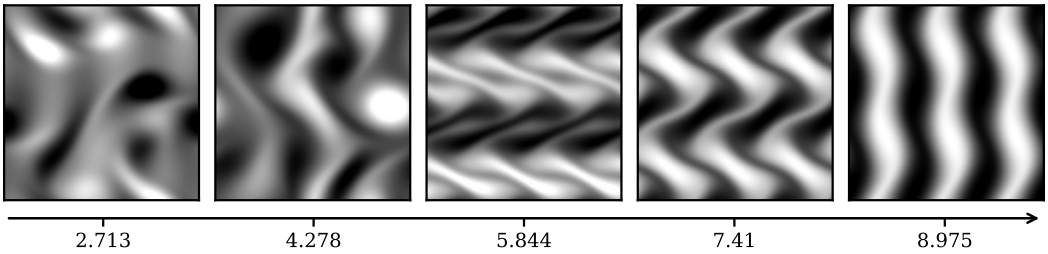

Figure 4: Kolmogorov dataset. Grayscale renderings of the simulated flow field are shown at increasing energy value (labels below each panel), highlighting the emergence and deformation of coherent, wave-like structures.

also in the context of quantum mechanics. The 2D Ising dataset is one of the quasi-standard datasets frequently used in scientific (ML) applications for benchmark purposes (Nguyen et al., 2024; Erdmann et al., 2021; Walker et al., 2020). Therefore, we will explain it in some more detail.

The numerical Ising model in two dimensions consists of a lattice (grid) of spins that can take one of two values: $+1$ (up) or -1 (down). Each spin represents a magnetic moment of an atom in a ferromagnetic material, and neighboring spins interact with each other, giving rise to a rich dynamical behvaior. To describe the system, the Hamiltonian is typically used. There, the energy of the system is given by the interaction between the neighboring dipoles i and j,

$$E = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j \ , \tag{18}$$

and the interaction of those dipoles with an external field applied to the system, h. Then, the Hamiltonian is written as:

$$H = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j - \mu \sum_i h_i \sigma_i \ , \tag{19}$$

where $\langle i, j \rangle$ is the sum over the nearest neighbours, J is the coupling force between the i^{th} and j^{th} magnetic dipole, $\sigma \in \{-1, 1\}$ is the magnetic dipole of a given site, and μ is the magnetic moment. For the current dataset, we simplify Eq. (19) by setting the external field h to 0 and the coupling force J to 1. By fixing $J > 0$, we are in the ferromagnetism regime: the spins in the lattice tend to align in the same direction. With this we obtain:

$$H = -\sum_{\langle i,j \rangle} \sigma_i \sigma_j \ . \tag{20}$$

The probability of the system being in a particular configuration of up and down spins at temperature T is given by a Boltzmann distribution. In a (ML) context, the "configuration" is represented as a black and white image while the temperature is used as the continuous conditioning variable. Fig. 5 shows examples for different temperature-image pairs. One of this model's peculiarities is that the system undergoes a phase transition near the Curie temperature T_C: this is seen as going from a large, ordered structure of spins oriented in the same direction (the large patches in the upper row of the figure) to spins that are randomly oriented without any clear pattern (the lower row images in the figure).

The $N = 128 \times 128$ lattice sites are randomly initialized, and a Metropolis Monte Carlo algorithm with periodic boundary conditions is applied. In each of these Monte Carlo steps, the image may slightly change, also changing the energy according to (20). The simulation terminates When the system reaches equilibrium or if a maximum number of steps is reached. Further details can be found in (Nguyen et al., 2024). For each temperature $T \in [0, 2 \times \mathrm{T}_c]$, several simulations are performed, and the dipole configuration from the simulation is saved as a black and white image (0 for a negative dipole and 255 for a positive one).

The image size is 128×128 pixels, and the dataset consists of altogether 50,000 images, uniformly distributed over the (scaled) temperature range of interest, $[0, \approx 5]$.

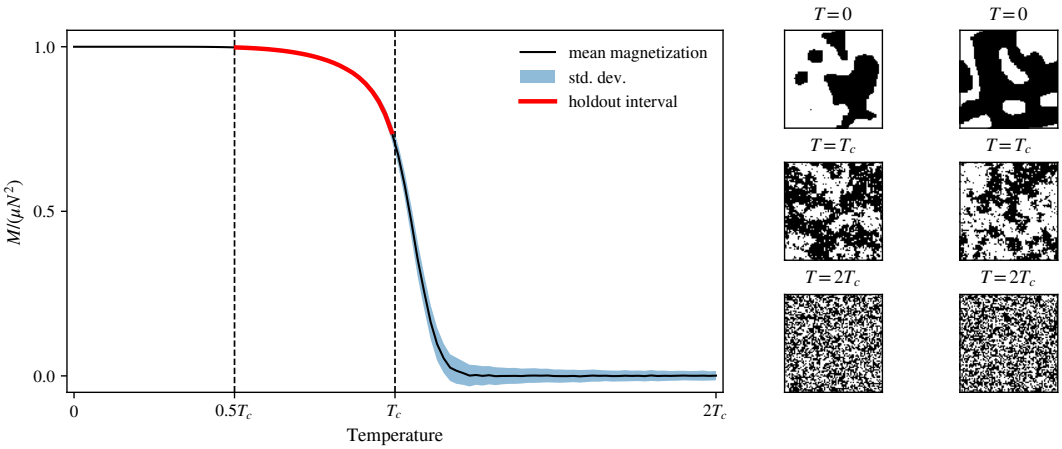

Figure 5: Evolution of the global magnetization of the lattice against the temperature. We can see that the system undertakes a phase transition at the Curie temperature T_c. Some examples of the microstructure obtained at different temperatures are shown on the right. The magnetization M can be used to track the state of the system. We also display the holdout interval for the experiment.

D.5 Mirrored Ising dataset

To study conditioning under non-monotonic label–attribute relations, we additionally construct a mirrored Ising dataset from the base Ising data. Starting from the original Ising dataset with temperature label $T \in [T_{\min}, T_{\max}]$, we create an extended label range by appending a reversed copy of the dataset. Each original sample (x, T) appears twice, once with its original label T and once with a mirrored label

$$T_{\mathrm{mir}} := T_{\max} + (T_{\max} - T), \tag{21}$$

so that the mirrored range spans $T_{\mathrm{mir}} \in [T_{\max}, 2T_{\max} - T_{\min}]$ and traverses the same images in reverse order. The maximum label in the mirrored dataset, therefore, corresponds to images that are identical to those at the minimum temperature in the original dataset, and intermediate labels introduce a fold in the label–attribute mapping. This construction preserves the marginal image distribution but renders the conditional distribution $p(x \mid T)$ non-injective and non-monotonic, providing a controlled stress test for label embeddings and conditioning mechanisms.

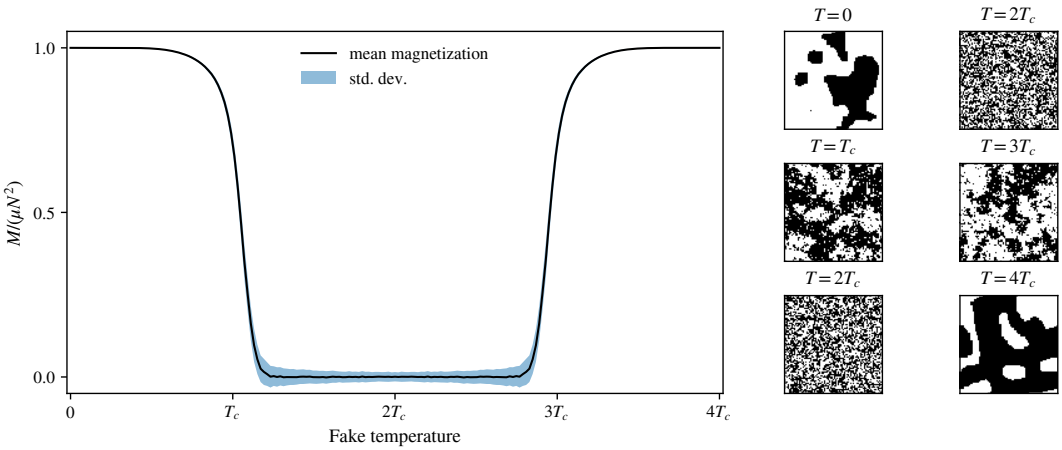

Figure 6: Evolution of the global magnetization of the lattice against the temperature for the mirrored Ising experiment used to investigate datasets with a non-monotonic label-data relationship.

E WAE TRAINING DETAILS

This appendix summarizes the WAE training protocol and model selection used in Sec. 4.1. We train WAE-MMD models with an Maximum Mean Discrepancy (MMD) penalty based on the inverse multiquadric (IMQ) kernel (Tolstikhin et al., 2018). The reconstruction objective depends on the dataset type. For continuous-field datasets (Cahn–Hilliard, Kolmogorov), we optimize a Charbonnier reconstruction loss, while for the remaining datasets we optimize mean squared error (MSE). Table 5 reports the selected hyperparameters and architecture settings. Table 6 reports the corresponding loss decomposition at the selected checkpoint.

Splitting protocol. For datasets with class/integer labels (MNIST, CIFAR-10/100, UTKFace, Dots), we use a stratified split that preserves class/integer proportions between the training and validation sets. For the Ising dataset, samples are selected to be approximately uniform across the continuous label range. For the Cahn–Hilliard and Kolmogorov datasets, which contain multiple trajectories/simulations, we split by simulation ID to avoid leakage between training and validation Nguyen et al. (2024).

Hyperparameter grid and selection rule. We perform a grid search over encoder type (Deterministic vs. Random), learning rate in $\{3 \cdot 10^{-3}, 10^{-3}, 5 \cdot 10^{-4}\}$, and regularization weight $\lambda \in \{10, 5, 1, 10^{-1}, 10^{-2}, 10^{-3}\}$. Each model is trained for 1000 epochs. For each dataset, we select the final configuration by minimizing the validation reconstruction loss under the same reconstruction objective used during training. After selection, we report the total objective together with the reconstruction and MMD terms on the training epoch average and on the validation split.

Encoder variants. The *Deterministic* encoder outputs a single latent code z per input, which is passed directly to the decoder. The *Random* encoder outputs the parameters of a diagonal Gaussian $Q_\phi(z \mid x)$ and uses the reparameterization trick to sample z during training.

Architecture and implementation details. The encoder and decoder reuse the same convolutional residual backbone as the R3GAN implementation to keep inductive biases aligned across the latent analysis and the subsequent conditional GAN study. The architecture is fully convolutional and avoids normalization layers. Resolution changes are implemented by interpolation-based downsampling and upsampling operators that use padding and cropping to suppress boundary artifacts, which is especially important for periodic scientific fields. The residual blocks follow an inverted-bottleneck design with an expansion factor of 2, a grouped 3×3 spatial convolution, and leaky-ReLU activations. The code initializes weights with an MSR-style rule and uses a residual projection initialized with zero gain so that each block starts close to an identity map (Zhang et al., 2019). All models used warmup for 50 epochs until it reaches LR in Table 5, further training is done with cosine annealing.

Model depth depends on image resolution. For 32×32 datasets (MNIST, CIFAR-10, CIFAR-100), we use 3 resolution stages and latent dimension $d_Z = 32$. For 128×128 datasets (Ising, Cahn–Hilliard, Kolmogorov, Dots, UTKFace), we use 5 resolution stages and latent dimension $d_Z = 128$. The decoder mirrors the encoder with the same number of stages. Table 5 additionally reports the interpolation mode used in the upsampling operator, since some continuous-field datasets benefit from bilinear upsampling. The reconstruction on the validation set is available on Fig 7 for the scientific dataset and on Fig 8 for the natural images.

Reporting conventions. $\text{Recon}_{\text{val}}$ is the selection criterion and is reported under the same reconstruction objective used for training. MMD_{val} is the IMQ-kernel MMD penalty at the selected checkpoint. $\text{Total}_{\text{val}}$ is the full WAE-MMD objective evaluated at the same checkpoint and is included for completeness.

E.1 LATENT SPACE VISUALIZATION

Figures 9 and 10 visualize Principal Component Analysis (PCA) projections of WAE latent codes (points colored by the normalized label). Figure 9 shows continuous-label scientific datasets, where averaging label-preserving views reduces encoder noise and reveals structured, label-aligned geometry. In these cases, the averaged latents and the radial summaries $\|\bar{z}\|_2$ versus label highlight simple

Table 5: Selected WAE configurations after grid search (WAE-MMD with IMQ kernel). "Blocks" denotes the number of resolution stages. "Upsampling" denotes the interpolation mode used in the decoder upsampling operator.

Dataset	Res.	Blocks	d_Z	Upsampling	Padding	Encoder	LR	λ
MNIST	32×32	3	32	Nearest	Zeros	Det.	10^{-3}	10^{-1}
CIFAR-10	32×32	3	32	Nearest	Zeros	Rand.	$3 \cdot 10^{-3}$	1
CIFAR-100	32×32	3	32	Nearest	Zeros	Rand.	$3 \cdot 10^{-3}$	1
Ising	128×128	5	128	Nearest	Circular	Rand.	$5 \cdot 10^{-4}$	10^{-2}
Cahn–Hilliard	128×128	5	128	Bilinear	Circular	Det.	10^{-3}	1
Kolmogorov	128×128	5	128	Bilinear	Circular	Det.	10^{-3}	10^{-1}
Dots	128×128	5	128	Nearest	Zeros	Det.	$5 \cdot 10^{-4}$	10^{-3}
UTKFace	128×128	5	128	Nearest	Zeros	Det.	10^{-3}	10

Table 6: Loss decomposition at the selected checkpoint. "Recon" denotes the reconstruction objective used for training, which is Charbonnier for continuous-field datasets (Cahn–Hilliard, Kolmogorov) and MSE otherwise. We report the total objective, reconstruction term, and MMD term on the training epoch average (tr) and validation split (val), shown as tr/val.

Dataset	Recon	$\text{Total}_{\text{tr/val}}$	$\text{Recon}_{\text{tr/val}}$	$\text{MMD}_{\text{tr/val}}$
MNIST	MSE	3.253e−3/3.237e−3	3.229e−3/3.122e−3	2.414e−4/1.148e−3
CIFAR-10	MSE	1.669e−2/1.680e−2	1.665e−2/1.670e−2	3.474e−5/9.881e−5
CIFAR-100	MSE	1.656e−2/1.690e−2	1.652e−2/1.684e−2	3.789e−5/5.972e−5
Ising	MSE	4.440e−1/4.428e−1	4.436e−1/4.424e−1	4.127e−2/3.982e−2
Cahn–Hilliard	Charb.	5.592e−2/5.648e−2	5.567e−2/5.627e−2	2.554e−4/2.121e−4
Kolmogorov	Charb.	7.539e−3/7.634e−3	7.447e−3/7.542e−3	9.184e−4/9.281e−4
Dots	MSE	1.472e−2/1.574e−2	1.469e−2/1.571e−2	3.439e−2/3.492e−2
UTKFace	MSE	1.481e−2/1.392e−2	1.313e−2/1.365e−2	1.686e−4/2.672e−5

empirical modulation patterns, including approximately monotone and in some datasets strongly radial organization, which motivates the conditioning choices studied in the main text. Figure 10 shows natural-image datasets with discrete labels, where the full latent clouds are largely unstructured with respect to label, while the label-wise centroids trace dataset-dependent, often irregular trajectories. The accompanying radial plots $\|z\|_2$ versus label indicate weak or inconsistent global monotonic trends in MNIST and CIFAR, while UTKFace exhibits a clearer centroid-level ordering that reflects the semantic progression of age.

E.2 Latent space coordinate

Latent radius r. We define the latent radius

$$r_i := \|z_i\|_2. \tag{10}$$

We report the Spearman rank correlation $\rho(r, y)$. A large $|\rho(r, y)|$ indicates that the label is strongly associated with a magnitude-like degree of freedom.

Geodesic coordinate g. To capture ordering along a potentially curved latent trajectory, we construct a weighted kNN graph $G_k = (V, E_k, w)$ with $V = \{1, \dots, n\}$. We add a directed edge $(i, j) \in E_k$ if j is among the k nearest neighbors of i in latent space, with weight

$$w_{ij} := \|z_i - z_j\|_2 \,. \tag{11}$$

For any path $\pi = (v_0, \dots, v_m)$, let $\ell(\pi) := \sum_{t=0}^{m-1} w_{v_t v_{t+1}}$ and define the shortest-path distance

$$d_{G_k}(i, j) := \min_{\pi:\, i \rightsquigarrow j} \ell(\pi) \,, \tag{12}$$

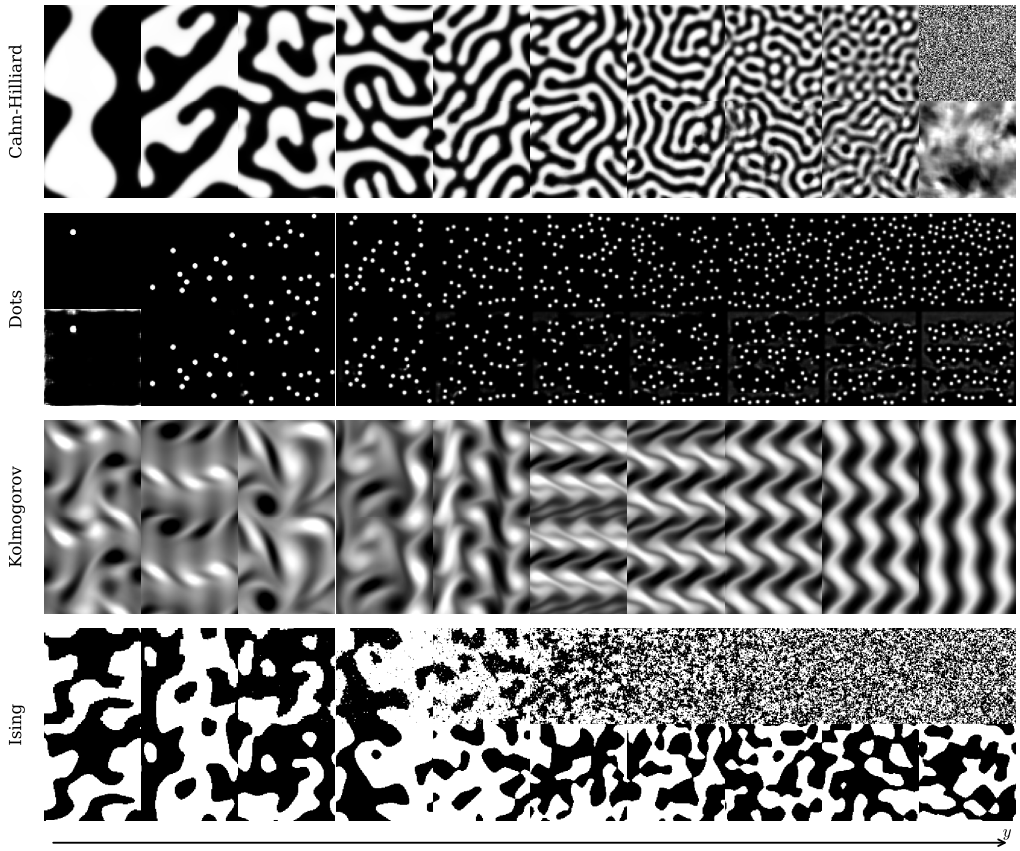

Figure 7: WAE reconstruction comparison on scientific datasets. For each dataset, the upper image is the input, while the lower is the reconstruction. In addition, the examples are organised in the order of label y increasing from left to right.

with $d_{G_k}(i,j) = +\infty$ if j is unreachable from i. This is a standard graph approximation of manifold geodesic distances (Tenenbaum et al., 2000). Let $A_{\min} \subseteq V$ be an anchor set near one end of the label range (e.g., the indices of the n_a smallest labels). We define the single-anchor geodesic coordinate

$$g_i := \min_{a \in A_{\min}} d_{G_k}(a, i). \tag{13}$$

We report $\rho(g, y)$ and the reachable fraction

$$\text{Reach.} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[g_i < \infty]. \tag{14}$$

Two-anchor coordinate s. Single-anchor distances can be sensitive to anchor placement and sampling density. We therefore also use anchors at both ends of the label range. Let $A_{\max}$ be an anchor set near the largest labels and define

$$g_i^{\min} := \min_{a \in A_{\min}} d_{G_k}(a, i), \qquad g_i^{\max} := \min_{a \in A_{\max}} d_{G_k}(a, i). \tag{15}$$

For nodes reachable from both ends, we define the normalized two-anchor coordinate

$$s_i := \frac{g_i^{\min}}{g_i^{\min} + g_i^{\max}} \in [0, 1]. \tag{16}$$

We report $\rho(s, y)$ computed over indices with finite $g_i^{\min}$ and $g_i^{\max}$.

Figure 8: WAE reconstruction comparison on natural images datasets. For each dataset, the upper image is the input, while the lower is the reconstruction.

Oriented local violation rate. Global rank correlations can be high even if local neighborhoods violate label order. We therefore measure local consistency on a symmetrized edge set E_k^{sym} derived from the kNN graph. We orient the geodesic coordinate using the sign of its global association with the label,

$$\tilde{g}_i := \operatorname{sign}\big(\rho(g, y)\big)\, g_i, \tag{17}$$

and count an undirected edge $\{i, j\} \in E_k^{\mathrm{sym}}$ as a violation when the label difference disagrees with the oriented coordinate difference:

$$\text{Viol.} := \frac{1}{|E_k^{\mathrm{sym}}|} \sum_{\{i,j\} \in E_k^{\mathrm{sym}}} \mathbb{I}\big[(y_j - y_i)(\tilde{g}_j - \tilde{g}_i) < 0\big]. \tag{18}$$

This is a local analogue of discordant-pair counting underlying Kendall-type rank statistics (Kendall, 1938).

Averaged encodings. For all datasets, we additionally compute "avg" statistics by applying 1024 label-preserving transforms to each sample, encoding each transform, and averaging the latent vectors before computing diagnostics. This reduces nuisance variation and can reveal an underlying ordered structure.

E.3 Full latent-space monotonicity diagnostics

Table 7 shows that several scientific datasets exhibit a measurable monotone organization of unsupervised WAE latents with respect to a continuous label. The purpose of this analysis is to connect the observed structure to established conditioning mechanisms in deep generative models and to leverage this connection to guide architectural design. For visualisation, we also provide PCA projections in Figures 9–10.

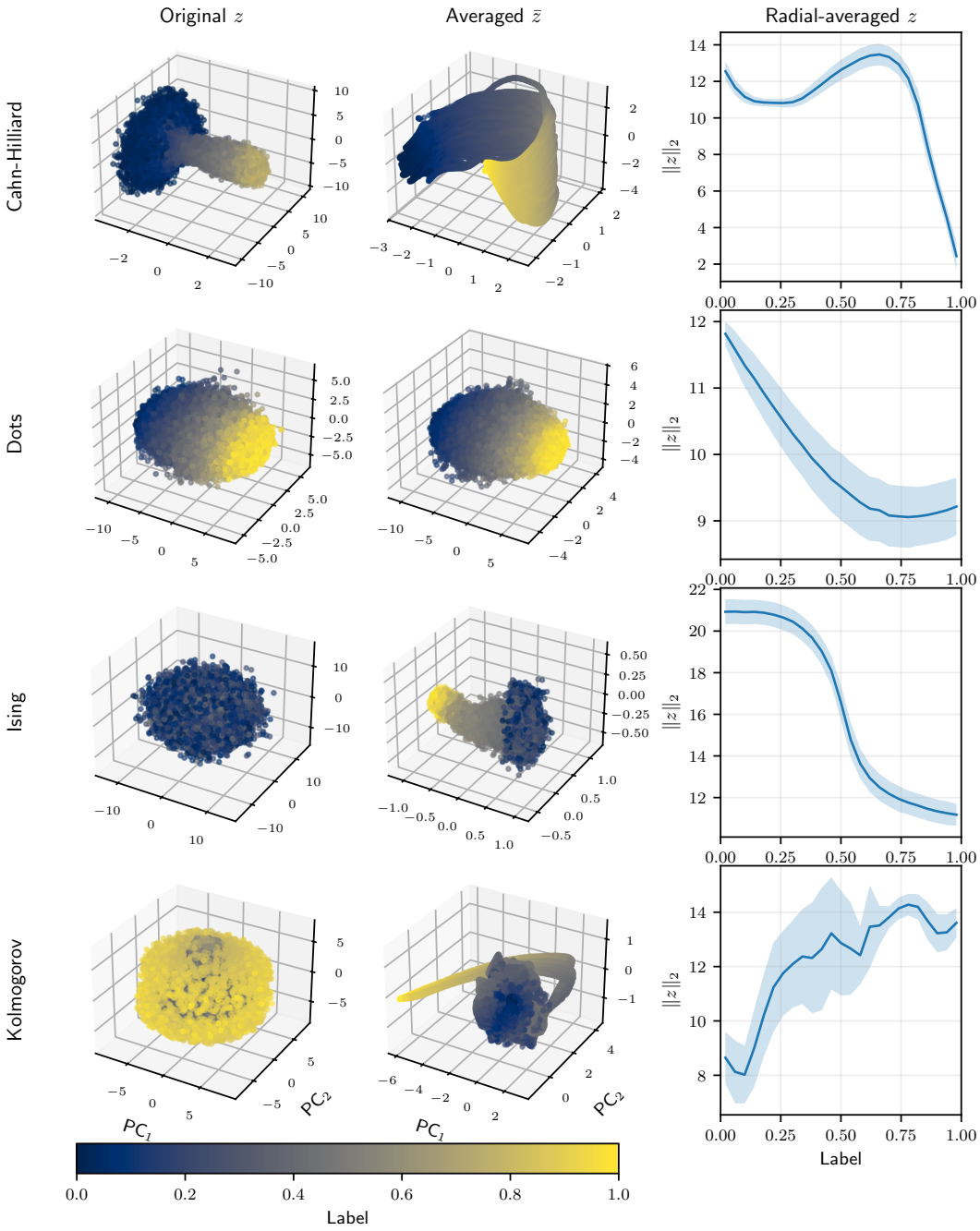

Figure 9: Latent representations for the scientific datasets (rows). **Left:** PCA of the original latent codes z, colored by the normalized label. **Middle:** PCA of label-consistent averaged codes $\bar{z}$ as Eq. 25 (same color scale). **Right:** radial statistics computed in the original latent space, showing the mean $\pm$ standard deviation of $\|z\|_2$ as a function of the label (binned).

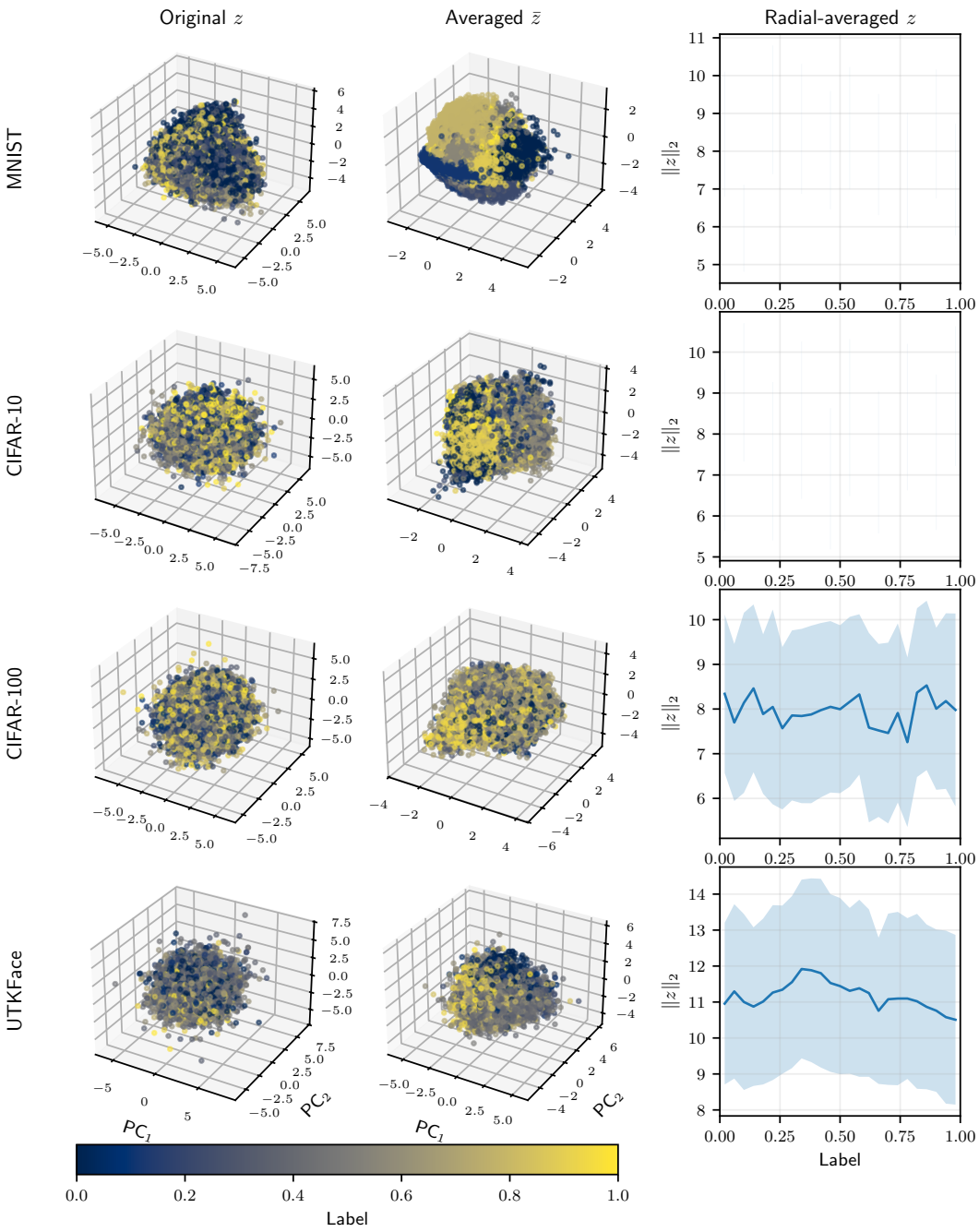

Figure 10: Latent representations for the natural images datasets (rows). **Left:** PCA of the original latent codes z, colored by the normalized label. **Middle:** PCA of label-consistent averaged codes $\bar{z}$ as Eq. 25 (same color scale). **Right:** radial statistics computed in the original latent space, showing the mean $\pm$ standard deviation of $\|z\|_2$ as a function of the label (binned).

Table 7: Latent-space monotonicity diagnostics for WAE computed on a kNN graph with $k = 20$. We report rank correlations between the label y and the single-anchor geodesic distance g, the two-anchor normalized coordinate s, and the latent radius $r = \|z\|_2$, together with an oriented local violation rate on kNN edges and the reachable fraction from the anchor set. Columns with subscript "avg" are computed after averaging per-sample latents across label-preserving transforms.

Dataset	$\rho(g, y)$	$\rho(g, y)_{\text{avg}}$	$\rho(s, y)$	$\rho(s, y)_{\text{avg}}$	$\rho(r, y)$	$\rho(r, y)_{\text{avg}}$	Viol.	Viol.$_{\text{avg}}$	Reach.
CH	0.5002	**0.9982**	0.3070	0.9963	0.3953	0.0535	0.8226	0.3276	1.0000
Dots	0.9414	0.9877	0.9586	**0.9879**	-0.8124	-0.3815	0.0120	0.0031	1.0000
Ising	-0.7800	-0.8630	0.3742	0.8982	**-0.9419**	-0.9209	0.1838	0.4610	1.0000
Kolmogorov	0.5159	**0.8930**	0.4414	**0.8927**	0.7360	0.3410	0.3574	0.4067	1.0000
MNIST	0.2185	0.2492	0.2364	**0.5722**	-0.0105	-0.2120	0.1090	0.040	1.0000
CIFAR-10	0.0647	**0.0746**	0.0623	0.0674	0.0509	0.0124	0.3378	0.3067	1.0000
CIFAR-100	**-0.0136**	0.0110	-0.0067	0.0112	-0.0103	-0.0061	0.4291	0.4543	1.0000
UTKFace	0.0804	0.2802	0.2506	**0.3981**	-0.0822	-0.0675	0.4852	0.4308	1.0000
Mirrored Ising	-0.0052	-0.0005	0.0061	0.0027	-0.0020	-0.0010	0.500	0.4997	1.0000

F ESTIMATING THE SHIFT $\beta(y)$ FROM WAE LATENTS VIA AUGMENTATION AVERAGING

This appendix describes a simple procedure for extracting the label-dependent shift $\beta(y)$ from a trained WAE and formalizes why averaging multiple label-preserving transforms of the same input improves the accuracy of the recovered $\beta(y)$.

F.1 SETUP AND NOTATION

Let $P_{X,Y}$ be a data distribution on $\mathcal{X} \times \mathcal{Y}$, where $y \in \mathcal{Y}$ is a continuous label. Let the trained WAE encoder be a Markov kernel $Q_\phi(z \mid x)$ on $\mathcal{Z}$, and let $z \in \mathbb{R}^{d_Z}$ denote the latent code.

We study the label-conditional aggregated posterior

$$Q_{Z|Y=y}(A) := \int_{\mathcal{X}} Q_\phi(A \mid x)\, dP_{X|Y=y}(x), \qquad A \subseteq \mathcal{Z}. \tag{22}$$

We focus on the label-dependent mean of $Q_{Z|Y=y}$, which we denote by

$$\beta^\star(y) := \mathbb{E}[Z \mid Y = y], \qquad Z \sim Q_\phi(\cdot \mid X),\ X \sim P_{X|Y=y}. \tag{23}$$

In the main text, $\beta^\star(y)$ corresponds to the FiLM shift term suggested by the observed latent structure.

F.2 LABEL-PRESERVING TRANSFORMS AND PER-IMAGE AVERAGING

Let $\mathcal{T}$ be a distribution over measurable transforms $T : \mathcal{X} \to \mathcal{X}$ such that T preserves the label. Formally, for $(X, Y) \sim P_{X,Y}$, assume

$$(X, Y) \sim P_{X,Y},\ T \sim \mathcal{T} \quad \Rightarrow \quad (T(X), Y) \sim P_{X,Y}. \tag{24}$$

In practice, T can represent standard data augmentations that do not change the label y (e.g. translations, flips, mild photometric changes, or simulation symmetries).

For a fixed labeled sample (x_i, y_i), draw K i.i.d. transforms $T_{i,1}, \ldots, T_{i,K} \sim \mathcal{T}$ and define

$$x_{i,k} := T_{i,k}(x_i), \qquad z_{i,k} \sim Q_\phi(\cdot \mid x_{i,k}), \qquad \bar{z}_i := \frac{1}{K} \sum_{k=1}^{K} z_{i,k}. \tag{25}$$

The averaged code $\bar{z}_i$ is the quantity used to fit $\beta(y)$ in the procedure below.

F.3 A SIMPLE NOISE MODEL FOR AUGMENTED LATENTS

Assume that, conditional on y, the encoder outputs for label-preserving transforms can be modeled as

$$z_{i,k} = \beta^\star(y_i) + \varepsilon_{i,k}, \tag{26}$$

where the residual satisfies

$$\mathbb{E}[\varepsilon_{i,k} \mid y_i] = 0, \qquad \mathrm{Cov}(\varepsilon_{i,k} \mid y_i) = \Sigma(y_i), \qquad \varepsilon_{i,1}, \ldots, \varepsilon_{i,K} \text{ i.i.d. given } y_i. \tag{27}$$

This model captures variability caused by applying different transforms, as well as possible stochasticity in Q_ϕ.

Lemma F.1 (Averaging preserves the mean and reduces variance)**.** *Under 26–27,*

$$\mathbb{E}[\bar{z}_i \mid y_i] = \beta^\star(y_i), \qquad \mathrm{Cov}(\bar{z}_i \mid y_i) = \frac{1}{K}\Sigma(y_i). \tag{28}$$

Consequently,

$$\mathbb{E}\big[\|\bar{z}_i - \beta^\star(y_i)\|_2^2 \,\big|\, y_i\big] = \frac{1}{K}\,\mathrm{tr}\big(\Sigma(y_i)\big), \tag{29}$$

which is a factor $1/K$ smaller than the corresponding quantity for a single latent code $z_{i,1}$.

Proof. The mean identity follows from linearity of expectation and $\mathbb{E}[\varepsilon_{i,k} \mid y_i] = 0$. The covariance identity follows from independence and the scaling of covariance under averaging: $\mathrm{Cov}(\frac{1}{K}\sum_k \varepsilon_{i,k} \mid y_i) = \frac{1}{K^2}\sum_k \Sigma(y_i) = \frac{1}{K}\Sigma(y_i)$. The MSE identity follows from $\mathbb{E}\|U\|_2^2 = \mathrm{tr}(\mathrm{Cov}(U))$ for zero-mean U. □

Lemma F.1 formalizes the basic reason why averaging helps with accuracy. The averaged code $\bar{z}_i$ is an unbiased estimate of the label-dependent mean $\beta^\star(y_i)$ with reduced noise.

F.4 Fitting $\beta(y)$ by Gaussian negative log-likelihood

We fit a parametric model $\beta_\eta : \mathcal{Y} \to \mathbb{R}^{d_Z}$ to approximate $\beta^\star(y)$. Optionally, we also fit a diagonal covariance model $\sigma_\eta^2 : \mathcal{Y} \to \mathbb{R}^{d_Z}_{>0}$, which captures label-dependent dispersion.

For a training pair $(y_i, \bar{z}_i)$, define the Gaussian negative log-likelihood (GNLL)

$$\ell(\eta; y_i, \bar{z}_i) := \frac{1}{2} \sum_{j=1}^{d_Z} \left[\log \sigma_{\eta,j}^2(y_i) + \frac{\big(\bar{z}_{i,j} - \beta_{\eta,j}(y_i)\big)^2}{\sigma_{\eta,j}^2(y_i)} \right]. \tag{30}$$

We train by empirical risk minimization:

$$\hat{\eta} \in \arg\min_\eta \ \frac{1}{n} \sum_{i=1}^{n} \ell(\eta; y_i, \bar{z}_i). \tag{31}$$

If $\sigma_\eta^2(y)$ is held fixed, 31 reduces to a weighted least-squares regression of $\bar{z}$ onto y.

F.5 Why averaging improves the accuracy of β estimated by GNLL

The benefit of averaging can be stated in two complementary ways.

1) Reduced observation noise implies improved target estimation. Lemma F.1 already shows that $\bar{z}_i$ concentrates around $\beta^\star(y_i)$ at rate $1/K$. Any reasonable regression procedure that estimates $\beta(y)$ from noisy observations inherits this improvement.

2) GNLL is maximum likelihood under a Gaussian model, and averaging increases Fisher information. Assume in addition that the conditional distribution of $z_{i,k}$ given y_i is Gaussian:

$$z_{i,k} \mid y_i \sim \mathcal{N}\big(\beta^\star(y_i), \Sigma(y_i)\big). \tag{32}$$

Then $\bar{z}_i \mid y_i \sim \mathcal{N}(\beta^\star(y_i), \Sigma(y_i)/K)$, and GNLL is the negative log-likelihood for that model. For clarity, consider any differentiable parameterization $\beta_\eta(y)$ and suppose $\Sigma(y)$ is known. For a fixed y, the Fisher information for η from one observation $\bar{z}$ is

$$\mathcal{I}_K(\eta \mid y) = \mathbb{E}\Big[\big(\nabla_\eta \log p(\bar{z} \mid y; \eta)\big) \big(\nabla_\eta \log p(\bar{z} \mid y; \eta)\big)^\top \Big] = K\, \mathcal{I}_1(\eta \mid y), \tag{33}$$

because the log-likelihood curvature scales with the inverse covariance, which scales by K. Therefore, by the Cramér-Rao lower bound, the covariance of any unbiased estimator $\tilde{\eta}$ satisfies

$$\mathrm{Cov}(\tilde{\eta} \mid y) \ \succeq \ \mathcal{I}_K(\eta \mid y)^{-1} \ = \ \frac{1}{K}\, \mathcal{I}_1(\eta \mid y)^{-1}. \tag{34}$$

Equation 34 shows that using K transforms per image improves the best achievable parameter accuracy by a factor $1/K$, under the stated assumptions. In particular, the estimation error of $\beta_\eta(y)$ decreases because the uncertainty in the fitted parameters decreases, leading to a closer match to the underlying data distribution.

G $\beta(y)$ estimation for Cahn-Hilliard and Dots datasets

The following section provides additional details regarding the linearity of $\beta(y)$ estimated from WAE encoder. The experiments are conducted according to the practical recipe listed below, while the theoretical reasoning for averaging is explained in Appendix F.

1. Train a WAE encoder $Q_\phi(z \mid x)$ on the dataset.
2. For each labeled sample (x_i, y_i), draw K label-preserving transforms $T_{i,k}$ and compute $\bar{z}_i$ using Eq. 25.
3. Fit $\beta_\eta(y)$ and $\sigma_\eta^2(y)$ by minimizing the GNLL in Eq. 31.
4. Compute $R^2(\beta_\eta(y), ay + b)$, where $a, b \in \mathbb{R}^{d_Z}$.

Table 8: GNLL and R^2 values for scientific datasets

Dataset	**GNLL** $\downarrow$	$R^2(\beta_\eta(y), ay + b) \uparrow$
Cahn-Hilliard	-2.3686	0.493
Dots	-1.8280	0.961
Ising	-2.8274	0.885
Kolmogorov	-2.8595	0.703

For this task, we train a 3-layer MLP with SiLU activation function between them for 5000 epochs. The network receives y as input and predicts $\beta_\eta(y)$ and $\sigma^2_\eta(y)$. We use the Adam optimizer with $lr = 1e-4$ with ReduceLROnPlateau scheduler with min_lr $= 1e-6$, factor$= 0.5$ and patience$= 5$.

The experimental values are available on Table 8, the low GNLL value indicates a good fit. Dots shows near-linearity ($R^2 \approx 0.96$), while Ising is moderately close to linear ($R^2 \approx 0.89$).

H GAN DETAILS

This appendix summarizes the GAN training protocols. We use the original code implementation from Huang et al. (2024). Unless otherwise specified, we used the same procedure as in the original paper. Both the generator and the discriminator only have one inverted bottleneck ResNet block as we found it to be sufficient for our experiments. Models are trained using AdamW optimization with $\beta_1 = 0$ and $\beta_2 = 0.99$. The initial learning rate is set to 2×10^{-4} and utilizes a cosine annealing scheduler for the first 50 epochs and held constant at 5×10^{-5}. We train for 500 epochs using a batch size of 1536 with the relativistic pairing loss with R_1 and R_2 regularization, and the corresponding discriminator regularization strength γ for each dataset can be found in Table 9.

	Ising	Dots	Cahn-Hilliard	Kolmogorov
γ for R_1 and R_2	10	200	20	20

Table 9: Table for the different γ strength used in the experiments. For the Ising holdout and mirrored version, we use the same γ as for the Ising dataset.

The ILI embeddings are trained following the procedure from Ding et al. (2020) for each dataset and are frozen during GANs training.

I SIMILARITY PROCRUSTES ANALYSIS FOR COMPARING LABEL-EMBEDDING SHAPES

This appendix introduces similarity Procrustes analysis and explains how we use it to compare the *shape* of learned label embeddings $c(y)$ across datasets. We apply the analysis to embeddings produced by ILI and by our lightweight MLP encoder (a linear layer with ReLU), together with a linear reference curve $Ay + b$. Similarity Procrustes alignment removes translation, isotropic scaling, and rotation, which are exactly the transformations that a subsequent linear layer can absorb. The remaining discrepancy, therefore, measures differences in embedding *shape* that cannot be eliminated by a linear readout and can influence conditional generation. The orthogonal Procrustes solution via SVD is due to Schönemann (1966), and standard treatments of similarity Procrustes and statistical shape analysis can be found in Dryden & Mardia (2016).

Embedding trajectories as shapes. Fix a dataset and a label encoder $c : \mathcal{Y} \to \mathbb{R}^{d_c}$. We sample a uniform grid of $n = 1000$ labels $\{y_i\}_{i=1}^n$ spanning $[0, 1]$ and form the matrix $C \in \mathbb{R}^{n \times d_c}$ whose ith row is $c(y_i)^\top$. We interpret the rows of C as a discrete trajectory in embedding space and compare trajectories obtained from ILI, MLP, and the linear reference $c_{\text{lin}}(y) = Ay + b$.

Similarity Procrustes alignment. Given two trajectories $A, B \in \mathbb{R}^{n \times d_c}$ defined on the same label grid, similarity Procrustes finds a translation vector $t \in \mathbb{R}^{d_c}$, a scale $s > 0$, and an orthogonal matrix

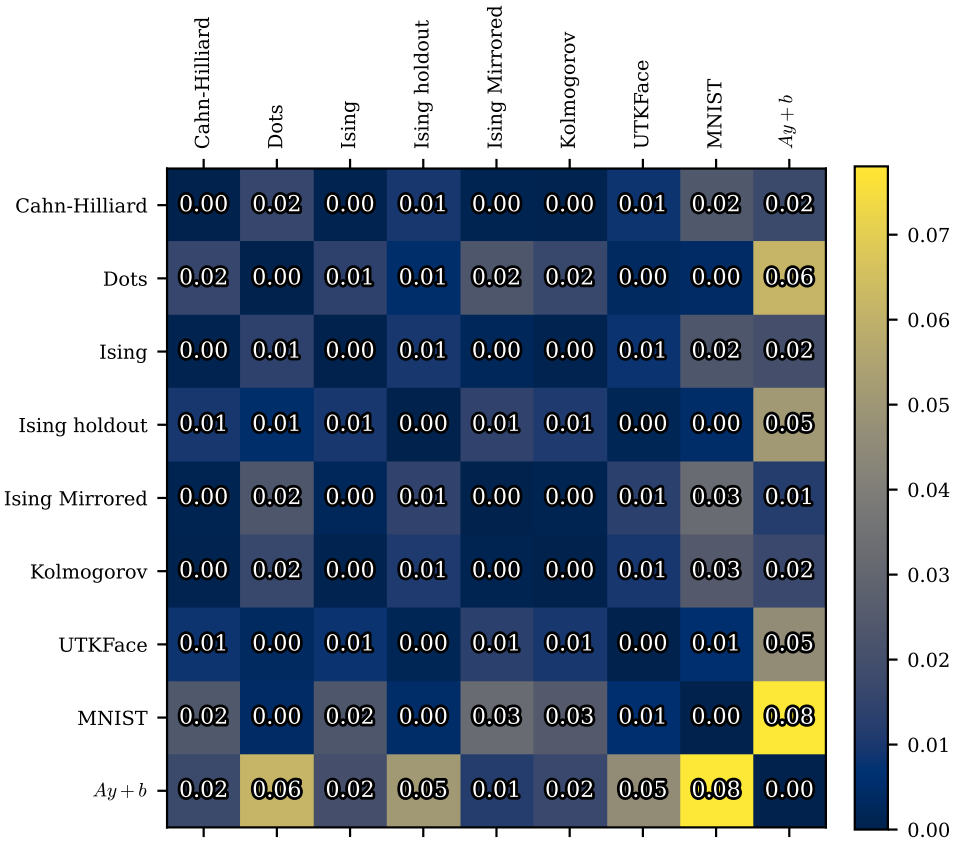

Figure 11: Similarity Procrustes distance matrix between ILI embedding trajectories learned on different datasets. We sample $n = 1000$ points uniformly in $y \in [0, 1]$ and compute distances after optimal translation, rotation, and isotropic scaling alignment.

$R \in \mathbb{R}^{d_c \times d_c}$ with $R^\top R = I$ that minimize

$$\min_{s>0,\ R^\top R=I,\ t\in\mathbb{R}^{d_c}} \left\| s\, AR + \mathbf{1}t^\top - B \right\|_F^2, \tag{35}$$

where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector. The standard solution centers both shapes, solves an orthogonal Procrustes problem by SVD to obtain R Schönemann (1966), and then chooses s to minimize the residual Dryden & Mardia (2016). We report the minimized value of Eq. 35 normalized by $\|B\|_F^2$ as a scale-invariant Procrustes discrepancy.

Why similarity Procrustes matches our conditioning setup. In our conditioning blocks, the label embedding is followed by a linear map that produces modulation parameters, for example $(\gamma(y), \beta(y)) = Wc(y) + b$ in a FiLM block. If we replace the embedding by a similarity-transformed version $c'(y) = s\, c(y)R + t$, then there exist parameters (W', b') such that $W'c'(y) + b' = Wc(y) + b$ for all y. The translation term is absorbed by the bias, and the scaling and rotation are absorbed by the weight matrix. Similarity Procrustes therefore removes precisely the degrees of freedom that are not identifiable when the embedding is followed by a linear readout. After alignment, remaining differences reflect shape properties such as curvature, folding, or changes in parameterization that cannot be removed by a single linear layer.

Conclusions from the measured distances. Figures 11 and 12 reveal two distinct behaviors. First, ILI embeddings exhibit very small pairwise Procrustes distances across datasets, including between Ising and Mirrored Ising, which indicates that the learned trajectory shape is nearly dataset-invariant once similarity transforms are removed. This supports the interpretation that ILI behaves like a fixed template whose differences across datasets are dominated by translation, rotation, or scale that can be absorbed by the subsequent linear readout. Second, the MLP embeddings remain close to a

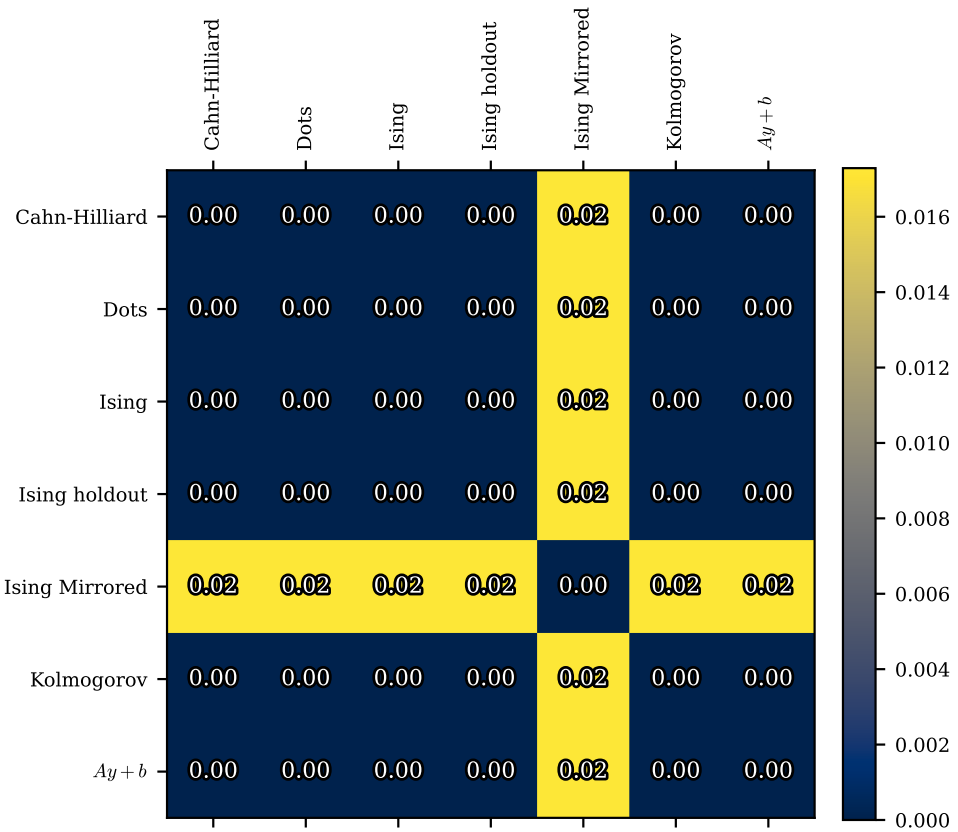

Figure 12: Similarity Procrustes distance matrix between MLP embedding trajectories learned on different datasets. The same $n = 1000$ label grid and similarity alignment protocol as in Fig. 11 is used.

linear reference $Ay + b$ on most datasets, but Mirrored Ising shows a consistently larger Procrustes discrepancy to the other datasets and to the linear baseline. This indicates that MLP adapts its embedding *shape* when the label–attribute relation requires it, which is consistent with the GAN results where MLP+FiLM remains accurate on Mirrored Ising while ID+FiLM fails and ILI+FiLM degrades. Together, the Procrustes analysis provides a geometric explanation for why ILI can underperform in non-monotonic settings: its embedding shape does not change enough to represent folding or other non-injective label structure, whereas the lightweight MLP can deviate from an affine trajectory when required.

J Evaluation Metrics for Physics-Based Image Distributions

Limitations of FID. The FID Heusel et al. (2017) is a commonly used metric for evaluating generative image models. It measures the Fréchet distance between two multivariate Gaussian distributions fitted to features extracted by an ImageNet-pretrained Inception-V3 network. While effective for natural images, FID has several limitations when applied to physics-based image distributions. First, FID relies on the assumption that feature embeddings are well approximated by Gaussian distributions. Recent studies have shown that deep feature representations, including those produced by Inception and CLIP models, often exhibit significant non-Gaussian structure, leading to biased or unstable distance estimates Chong & Forsyth (2020); Jayasumana et al. (2024). This issue is particularly pronounced for physical systems, where data typically lie on continuous low-dimensional manifolds parameterized by quantities such as temperature or correlation length. Second, the Inception-V3 encoder introduces semantic biases specific to natural images, which are largely misaligned with the morphological and statistical structures present in lattice-based or PDE-generated fields. As a result, FID may fail to capture physically meaningful discrepancies between generated and reference samples.

From FID to MMD-based metrics. Recent work has questioned the suitability of Fréchet-based metrics for evaluating generative models in deep feature spaces. In particular, CLIP-based Maximum Mean Discrepancy (CMMD) was introduced to address limitations of FID arising from the mismatch between Fréchet distance assumptions and the empirical structure of deep embeddings Jayasumana et al. (2024). By replacing the Fréchet distance with MMD, CMMD avoids parametric assumptions on feature distributions and has been shown to provide improved robustness and sensitivity when comparing complex image distributions. These results suggest that MMD-based metrics offer a more appropriate framework than FID in settings where Gaussianity assumptions are violated.

Power spectral density as a physics-informed feature representation. Motivated by these observations regarding Fréchet-based metrics, we adopt the power spectral density (PSD) as the primary feature representation for evaluating generative models of physical systems. The PSD is computed as the squared magnitude for the Fourier transform of the field, and then averaged over the angular direction in the Fourier space. For spatially extended fields with periodic boundary conditions, the PSD provides a natural and physically meaningful characterization of spatial correlations. By construction, the PSD is invariant under spatial translations, and its radially averaged form is also invariant under rotations. Furthermore, the PSD is directly related to two-point correlation functions via the Wiener–Khinchin theorem, making it sensitive to characteristic length scales, correlation lengths, and domain morphology.

We compute MMD in PSD space (PSD-MMD) to compare generated and reference distributions. This approach enables a direct comparison of spatial statistics without reliance on semantic features or parametric assumptions, and aligns naturally with the symmetries and statistical structure of the underlying physical systems.

K VISUALS FROM THE GENERATORS

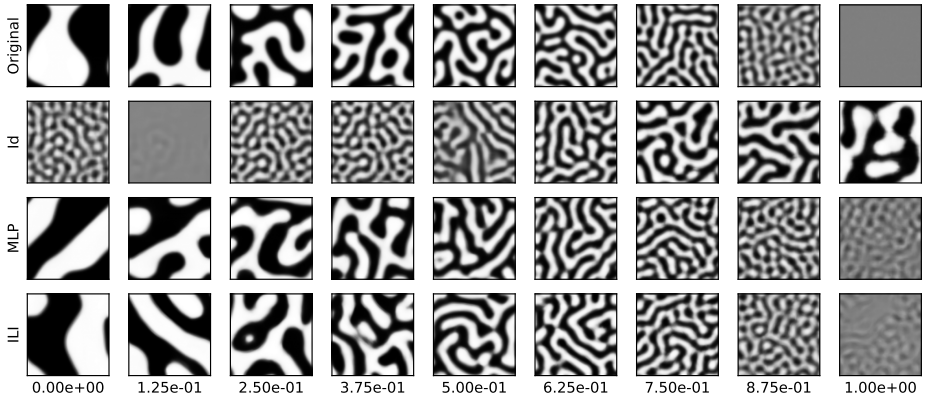

Figure 13: Qualitative comparison of phase-field evolution for the Cahn–Hilliard obtained by the different generators.

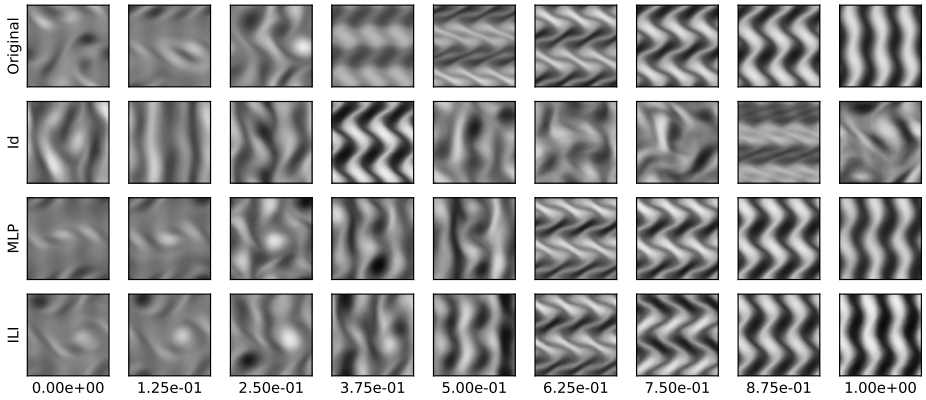

Figure 14: Qualitative comparison of the evolution of the Kolmogorov obtained by the different generators.

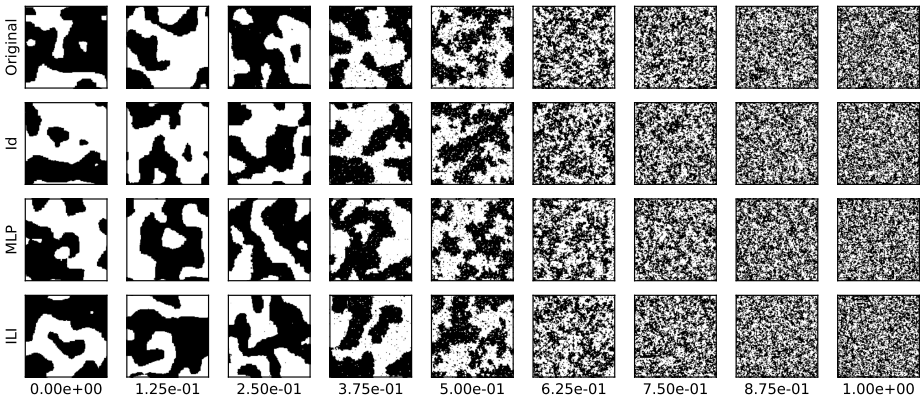

Figure 15: Qualitative comparison of lattice evolution for the Ising obtained by the different generators.

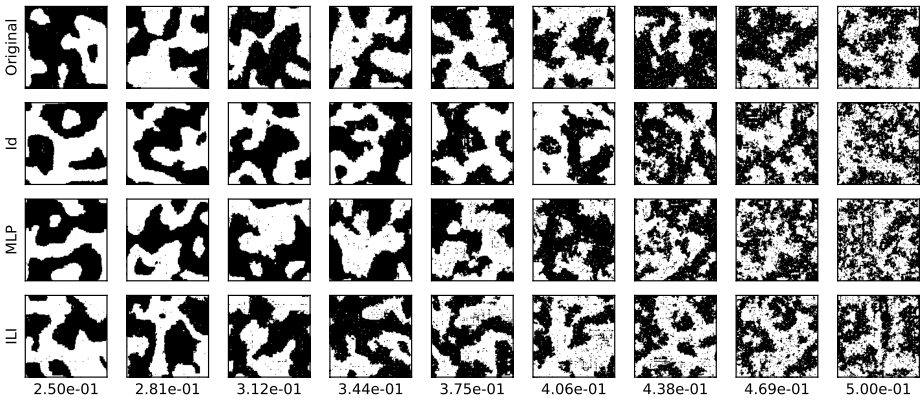

Figure 16: Qualitative comparison of lattice evolution for the Ising for the cropped section obtained by the different generators.

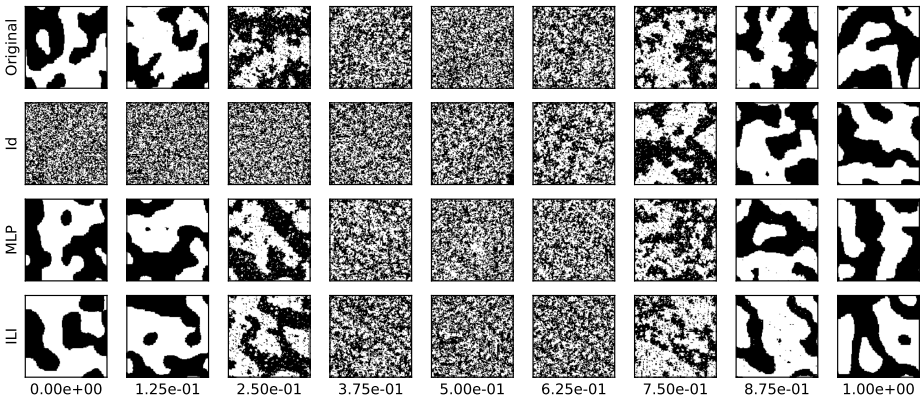

Figure 17: Qualitative comparison of lattice evolution for the mirrored Ising obtained by the different generators.

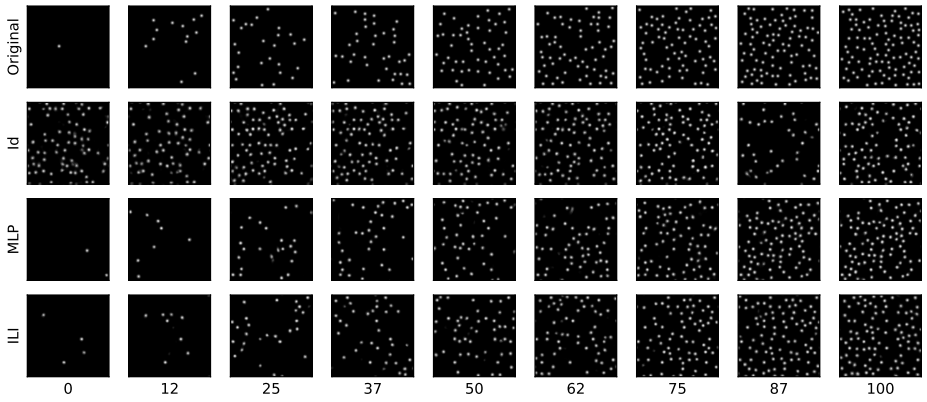

Figure 18: Qualitative comparison of lattice evolution for the Dots obtained by the different generators.