Weight-Space Physics: Interpretable Hypernetworks for Lattice Quantum Field Theories

Tobias Göbel [*1 2] **Julian R. Ebelt** [*1 2] **Zier Mensch** [*1 2] **Mathis Gerdes** [3 4] **Miranda C.N. Cheng** [1 5]

Abstract

Lattice field theory is the workhorse of non-perturbative physics, used to simulate phenomena from the strong nuclear force to critical phenomena in materials. Its Boltzmann distributions are parametrized analytically by coupling constants, but these bare parameters are weak predictors of observables—extracting physics typically requires extensive simulation. While normalizing flows have emerged as effective samplers at fixed couplings, it remains difficult to interpret what these networks have learned. This raises a natural question: can the flow *network parameters themselves* be generated for new theories, and can their physics be read off directly from the network weights? We propose lattice field theory as a testbed for neural network interpretability: because the target physics is qualitatively well-understood and smoothly varying, it provides ideal synthetic data with known ground truth. To this end, we introduce **JEPAWG**, a Joint-Embedding Predictive Architecture–based Weight Generator that maps couplings directly to flow weights via a learned latent space. On a scalar theory at lattice sizes 6^2, 8^2 and 10^2, the JEPAWG latent space recovers the correct intrinsic dimension of the underlying manifold, locates the phase transition, and encodes a finite-size shift aligned with the 2D Ising exponent $\nu \approx 1$, allowing us to uncover physical structure by studying the network weights alone. This suggests the fascinating idea of treating the network weights as a new type of physical observable. As a generator, JEPAWG also interpolates and extrapolates to unseen couplings effectively and remains robust to weight-space discontinuities introduced by multi-seed training data, outperforming PCA, AE, and VAE baselines.

[*]Equal contribution [1]Institute of Physics, University of Amsterdam, Amsterdam, Netherlands [2]Department of Physics, National Taiwan University, Taipei, Taiwan [3]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA [4]The NSF Institute for Artificial Intelligence and Fundamental Interactions [5]Institute for Mathematics, Academia Sinica, Taipei, Taiwan. Correspondence to: Miranda C.N. Cheng <miranda.cheng.w@gmail.com>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Lattice field theory (LFT) discretizes a quantum field onto a finite lattice, providing a first-principles approach to non-perturbative phenomena across high-energy and condensed matter physics. The action $S_g(\phi)$ is parametrized by bare coupling constants g, and physics is extracted by sampling the Boltzmann distribution $p_g(\phi) \propto e^{-S_g(\phi)}$ numerically.

Flow-based generative models have emerged as samplers for these distributions (Albergo et al., 2021; Kanwar, 2024; Cheng & Stratikopoulou, 2026), learning an invertible map from a tractable base to $p_g(\phi)$ and producing independent samples that sidestep MCMC autocorrelation. A trained flow at coupling g is parametrized by a vector θ_g, and we expect these parameters to carry information about the physical structure of the theory. For instance, near a phase transition, where p_g varies rapidly with g, θ_g must respond accordingly. This raises two complementary questions. *As a diagnostic*—our primary focus— does the geometry of these parameters in the weight space recover physical structure that the bare couplings themselves obscure? *As a generator*, instrumentally, can θ_g be predicted from g, amortizing training across couplings and extrapolating to unseen ones?

Lattice field theory is a uniquely well-suited testbed for these questions of *weight interpretation* and *weight generation*, because it supplies the kind of ground truth that interpretability research typically lacks. Understanding what neural networks have learned is a central open problem (Räuker et al., 2023; Sharkey et al., 2025), but progress is often bottlenecked by the absence of ground truth: for typical ML tasks, we cannot independently verify what structure a learned representation *should* encode. Physical theories supply this ground truth. The target distributions $p_g(\phi)$ are known analytically, the relevant coupling space is low-dimensional and naturally smooth, and the renormalization

group makes quantitative predictions about its phase structure (Section 3.1). A representation of weight space can therefore be checked against falsifiable physical predictions, not only against qualitative intuitions. Physics, in this sense, acts as a supplier of natural synthetic data: rich enough to be non-trivial, structured enough to be interpretable.

To this end, we introduce **JEPAWG**, a JEPA-based Weight Generator (LeCun, 2022; Assran et al., 2023) that maps couplings to flow parameters through a learned latent space. The same architecture that enables weight generation at unseen couplings also yields a latent representation whose geometry can be probed for physical structure. We apply JEPAWG to two-dimensional ϕ^4 theory and make four contributions.

(i) **Recovery of physical dimensionality:** the JEPAWG latent space matches the 2D coupling manifold (MLE intrinsic dimension 2.01 ± 0.06; 98.5% of variance in two principal axes), whereas AE and VAE baselines inflate this to 3.68 and 4.88.

(ii) **Phase transition from weights alone:** the pullback metric on $\theta(g)$ exhibits a ridge tracking the critical line of ϕ^4 theory.

(iii) **Scale transformation from latent geometry:** comparing weights of flows trained at different lattice sizes through the learned embedding, and requiring physically similar theories to lie nearby in the embedding space, we find that the latent geometry is organized by the finite-size scaling variable, with the optimal exponent broadly consistent with the 2D Ising value $\nu = 1$. A shuffled-label control confirms that this signal vanishes when the coupling–weight correspondence is broken.

(iv) **Generation:** JEPAWG produces functional flows at unseen couplings, reaching an effective sample size (ESS) of 0.88 on multi-seed interpolation, where PCA, AE, and VAE baselines reach only 0.30–0.40, and it remains robust under the weight canonicalization procedure.

2. Related Work

Joint-embedding predictive architectures. Joint-embedding predictive architectures (JEPA) learn representations by predicting in latent space rather than reconstructing the input (LeCun, 2022). I-JEPA (Assran et al., 2023) and V-JEPA (Bardes et al., 2024) demonstrated this for images and video, while VL-JEPA (Chen et al., 2025) extended the framework to vision-language tasks, attaching a lightweight decoder to translate predicted embeddings back into text. To our knowledge, JEPA has not previously been applied to neural network weight spaces.

Hypernetworks and weight generation. A hypernetwork (Schmidhuber, 1992; Ha et al., 2017) is a network that produces the weights of another, conditioned on a task or system descriptor. This paradigm has been applied to continual learning (von Oswald et al., 2020) and approximate Bayesian inference over weights (Krueger et al., 2017); see Chauhan et al. (2024) for a survey. Our work instantiates this paradigm for flow models simulating quantum field theories: a learned encoder–predictor–decoder pipeline outputs per-coupling weights, enabling extrapolation to unseen couplings without retraining.

Flow-based sampling for lattice field theory. Normalizing flows have been widely applied as samplers for lattice field theories; see e.g. Cheng & Stratikopoulou (2026); Kanwar (2024) for reviews. Closest to our setting, Gerdes et al. (2023) train a single coupling-conditional flow that generalizes across couplings by conditioning each layer on g while keeping network weights shared. Instead of conditioning a fixed network, our method treats trained flow parameters as data and learns a latent embedding that decodes to flow parameters at unseen couplings.

3. Background

3.1. Scalar Quantum Field Theory on the Lattice

We work with the two-dimensional real scalar ϕ^4 theory on a periodic square lattice $V_L \cong (\mathbb{Z}/L\mathbb{Z})^2$, with site variables $\phi_x \in \mathbb{R}$ and action

$$S_g(\phi) = \sum_{\langle x,y \rangle} \phi_x \Delta_{xy} \phi_y + \sum_{x \in V_L} m^2 \phi_x^2 + \sum_{x \in V_L} \lambda \phi_x^4, \quad (1)$$

where Δ is the discrete Laplacian and $g = (m^2, \lambda)$ collects the two bare couplings[1]. Physical observables are expectations under the Boltzmann measure $p_g(\phi) \propto e^{-S_g(\phi)}$. Despite its simplicity, this theory exhibits a non-trivial phase structure and therefore forms an excellent testbed for studying what neural samplers learn, since the known phase structure provides a falsifiable ground truth against which learned representations can be checked.

Phase structure. Near a phase transition, the physics becomes acutely sensitive to the couplings: a small change in g produces a qualitative change in the theory, which makes critical regions among the hardest to simulate. At the same time, the long-distance physics there becomes *universal*: it is determined only by symmetry and dimensionality, independent of microscopic details. Together, these properties

[1] These are called *bare* couplings because they appear directly in the action, before quantum effects are taken into account; the physically meaningful indicators are the observables whose mapping to bare couplings is typically very complex and depends on the lattice spacing.

make critical points a natural place to test whether a learned representation has captured the physical structure.

The action (1) is invariant under the global $\mathbb{Z}_2$ symmetry $\phi \mapsto -\phi$. For $\lambda > 0$ and $m^2 < 0$, the theory has two phases, separated by a critical line $m^2 = m_c^2(\lambda)$ in the coupling plane. When $m^2 > m_c^2(\lambda)$, the lattice magnetization $\bar{\phi} = L^{-2}\sum_x \phi_x$ averages to zero (the *paramagnetic* phase). When m^2 is sufficiently negative, the potential acquires two degenerate minima; the system settles into one of them, breaking $\mathbb{Z}_2$ spontaneously, and $\langle|\bar{\phi}|\rangle \neq 0$ (the *ferromagnetic* phase). The coupling space (m^2, λ) is two-dimensional and the two phases are separated by a one-dimensional critical line; any representation that captures the physics should reflect this low-dimensional structure.

Observables. The transition is conventionally diagnosed by moments of the magnetization. The susceptibility $\chi(g) = L^2\big(\langle\bar{\phi}^2\rangle - \langle|\bar{\phi}|\rangle^2\big)$ peaks along the critical line, while the Binder cumulant $U_4(g) = 1 - \langle\bar{\phi}^4\rangle/(3\langle\bar{\phi}^2\rangle^2)$ interpolates between 0 in the paramagnetic phase and $2/3$ in the ferromagnetic one, with curves at different L crossing at the critical point. Two-dimensional ϕ^4 theory belongs to the 2D Ising universality class, with the correlation-length critical exponent $\nu = 1$. A learned representation that has captured the physical content of the theory should therefore organize the coupling plane by RG-invariant combinations rather than by g itself, a property we test in Section 5.3.

Normalizing flows for lattice field theory. Estimating these observables requires drawing samples from the Boltzmann measure $p_g(\phi) \propto e^{-S_g(\phi)}$, a task that becomes increasingly difficult near the critical line, where local Monte Carlo updates suffer from critical slowing down. In recent years, flow-based methods have been introduced as successful samplers of lattice field theories (Albergo et al., 2019), offering the promise of mitigating critical slowing down. A normalizing flow (Tabak & Vanden-Eijnden, 2010; Papamakarios et al., 2021; Kobyzev et al., 2020) learns an invertible map $f_\theta : \mathbb{R}^{L^2} \to \mathbb{R}^{L^2}$ that transforms a tractable base distribution ρ (typically a standard Gaussian) into an approximation of the target Boltzmann measure $p_g(\phi)$. The map is trained by minimizing the reverse KL divergence between the pushforward of ρ under f_θ and p_g, using only the known action S_g and samples from the flow itself. A conditional flow architecture (Gerdes et al., 2023) extends this by making the parameters depend smoothly on the couplings $g = (m^2, \lambda)$, so that a single model produces a sampler for each theory in a continuous region of coupling space.

3.2. Joint-embedding predictive architectures

A joint-embedding predictive architecture (JEPA) (Assran et al., 2023; LeCun, 2022) learns representations by predicting the embedding of one signal from the embedding of a related one, rather than reconstructing the input directly. Given two such signals, x and y, that describe the same underlying object, encoders f_x and f_y produce embeddings $z_x = f_x(x)$, $z_y = f_y(y)$ and a predictor h maps between them. The encoders and the predictor are trained jointly, so that the embeddings and the predictive map between them are shaped together by a single objective enforcing $h(z_y) \approx z_x$.

Unlike generative architectures, JEPA operates entirely in representation space, encouraging the embeddings to capture high-level structure rather than input-level detail. In our setting (Section 4.1), the underlying object is a lattice theory at a fixed coupling, and the two signals are its bare coupling g and a flow's parameter vector θ trained to sample at g.

Without additional regularization, both encoders can collapse to a constant mapping, trivially satisfying the alignment objective. The VICReg objective (Bardes et al., 2022) prevents this with three terms: an *invariance* term $s = \frac{1}{N}\|h(z_y) - z_x\|_2^2$ that enforces alignment, a *variance* term that penalizes small variance and hence representational collapse, and a *covariance* term that penalizes correlations between latent dimensions and thereby encourages the utilization of all available latent dimensions. Full expressions are given in Appendix A.2.

4. Methods

We introduce JEPAWG (Section 4.1), our JEPA-based weight generator that maps couplings to flow parameters through a learned latent space, and compare it against three baselines (PCA, AE, VAE), trained without coupling information (Section A.1).

4.1. JEPAWG

We slightly modify the joint-embedding predictive architecture (JEPA) introduced in Section 3.2, and augment it with a decoder that enables *weight generation* at unseen couplings. We refer to the resulting model as JEPAWG (JEPA-based Weight Generator), combining JEPA-style representation learning with a hypernetwork-style mapping from coupling space to flow parameters. We use a JEPA backbone, rather than a coupling-conditional regressor $g \mapsto \hat{\theta}$ trained with reconstruction loss, because JEPA shapes the latent space from both signals jointly: the predictor is trained to recover the weight embedding $\mathbf{z}_\theta$ from the coupling embedding $\mathbf{z}_g$, while the variance term prevents either embedding from collapsing to a constant. The latent space is therefore incentivized to align with the coupling structure during training rather than fitted to it post-hoc. We will see that this feature is most helpful when the relevant sub-manifold in the weight space is non-smooth (Section 5.2).

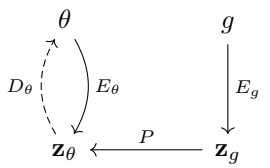

Figure 1. JEPAWG architecture: weight encoder E_θ, coupling encoder E_g, predictor P, and decoder D_θ.

Architecture. The model consists of four components: (i) a weight encoder $E_\theta : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_z}$ (with $d_\theta = 19{,}689$, the dimension of the neural ODE weights; see Appendix B.1) mapping canonicalized [2] parameter vectors θ to weight embeddings $\mathbf{z}_\theta = E_\theta(\theta)$, (ii) a coupling encoder $E_g : \mathbb{R}^2 \to \mathbb{R}^{d_z}$ mapping couplings $g = (m^2, \lambda)$ to coupling embeddings $\mathbf{z}_g = E_g(g)$, (iii) a predictor $P : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z}$ mapping coupling embeddings into the weight embedding space, trained so that $P(\mathbf{z}_g) \approx \mathbf{z}_\theta$, and (iv) a decoder $D_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_\theta}$ mapping predicted weight embeddings back to parameter vectors. At inference, the composition $D_\theta \circ P \circ E_g : g \mapsto \hat{\theta}$ defines a hypernetwork from couplings to flow parameters.

Training. The encoders E_θ, E_g and predictor P are trained using the VICReg objective (7), with the covariance term removed. The reason is that this term deliberately decorrelates latent dimensions and encourages the representation to spread across all d_z available directions; however, we wish to retain and exploit the low-dimensional structure of the problem—our object of study is a two-parameter family of Boltzmann distributions. This property is probed directly in our experiments, as discussed in Section 5.3. For completeness, we also evaluate the variant that retains the covariance term (i.e., the full VICReg objective), which we denote JEPAWG*. After training the encoders and the predictor, we freeze (E_θ, E_g, P) and train the decoder D_θ. This two-stage procedure ensures that the latent space is shaped entirely by the JEPA objective; training the decoder jointly would allow reconstruction pressure to leak into the weight encoder, encouraging it to retain θ-specific detail beyond what is predictable from g. The decoder minimizes

$$\mathcal{L}_{\text{dec}} = \frac{1}{N} \sum_{i=1}^{N} \left\| D_\theta(P(E_g(g_i))) - \theta_i \right\|_2^2 \tag{2}$$

on training pairs $\{(\theta_i, g_i)\}_{i=1}^N$. Throughout, we set $d_z = 16$, deliberately over-sizing the latent space relative to the two-dimensional coupling manifold so that any concentration onto fewer dimensions is a learned property rather than an architectural constraint.

[2] See Section 5.1

Inference. Given a new coupling $g^\star$, we generate flow parameters by applying the map $\hat{\theta}(g^\star) = D_\theta\big(P(E_g(g^\star))\big)$. The generated parameters $\hat{\theta}(g^\star)$ can then be used to instantiate a flow model at the desired coupling.

4.2. Baselines

We compare against three baselines: weight-space PCA, an autoencoder (AE) and a variational autoencoder (VAE); see Section A.1. As an ablation, we also include JEPAWG*, which uses the full VICReg loss with nonzero covariance term. Each method operates on flattened parameter vectors $\theta \in \mathbb{R}^{d_\theta}$, standardized per dimension as $\tilde{\theta}_i = (\theta_i - \bar{\theta}_i)/\sqrt{\mathrm{Var}(\theta_i)}$. Coupling information enters only at inference, through a post-hoc regressor mapping $g = (m^2, \lambda)$ to the learned embedding, which allows the baseline to act as a conditional generator. All baselines are trained on the same parameter vectors as JEPAWG, isolating the architectural comparison from data-pipeline effects. We report results in Table 1 and discuss the asymmetric effect of canonicalization in Section 5.2.

5. Experiments

A central question we address is whether a single framework can achieve both competent flow-model generation for unseen couplings and meaningful representation learning whose geometry reflects the underlying physics. We evaluate JEPAWG on both criteria. We focus our analysis on the multi-seed setting (see Section B) where the training data occupy disjoint regions of weight space, to ensure that the representations JEPAWG learns are not merely a result of the specific way of data generation.

For **weight generation**, we measure the effective sample size (ESS; (19)) of decoded flows on interpolation and extrapolation points (Section 5.2). For **representation geometry**, we estimate the intrinsic dimension of the latent space, visualize its structure, and show that the geometry of the parameter map recovers the location of the phase boundary directly, localizing the criticality signal to non-local components of the network (Section 5.3). Crucially, JEPAWG is trained only on weight reconstruction; this physical structure emerges without supervision from physical observables.

5.1. Data Preparation

We use the conditional flow of (Gerdes et al., 2023) as the first step in our data-generation machinery. On a 19×19 grid in the space of couplings (m^2, λ), we fine-tune the models generated by the conditional flow so that the resulting ESS is at least 0.95. We repeat the process for lattices with sizes 6^2, 8^2, and 10^2. The network weights have various symmetries, including changes of basis, permutations, and sign flips, which change the parameter vector θ while lead-

ing to the same input-output map. We fix this redundancy with a **weight canonicalization** procedure: we map each θ to a unique canonical representative per symmetry orbit (Appendix A.3) to facilitate non-ambiguous feature extraction from weights; the importance of this step is quantified in Section 5.2. More details of the data preparation procedure are described in Appendix B.

5.2. Weight Generation

In the data preparation stage, we first train two conditional flows with independent random initializations, each covering the full λ interval but a disjoint half of the m^2 range: seed 1 covers $m^2 \in [-5.5, -3.5]$ and seed 2 covers $m^2 \in (-3.5, -1.5]$, both over $\lambda \in [3.0, 6.0]$. The training set for all methods is assembled by combining the (g, θ_g) pairs from the two parts. Because the two conditional flows converge to different optima, the weight manifold has a structural discontinuity at the boundary, even though the underlying physics varies continuously. This multi-seed approach thus helps to ensure the robustness of our experiments: JEPAWG does not simply inherit coherent structure from the conditional network. We evaluate on 256 coupling pairs (m^2, λ) uniformly sampled within their respective full training intervals, and 32 coupling pairs in each of eight extrapolation regions, each extending a distance $\delta = 1.5$ beyond the training window along the relevant axis (Figure 10). These eight regions are divided into four edge regions (one of the coupling constants beyond the training window) and four corner regions (both couplings outside the training window); see Figure 10 for an illustration. Table 1 shows that JEPAWG produces the strongest flows in this setting, achieving an interpolation ESS of $\mathbf{0.88}$. JEPAWG* reaches 0.61, while the reconstruction-trained baselines (VAE, AE, PCA) lag substantially at 0.30–0.40. The gap widens on extrapolation: JEPAWG retains 0.56 on edges and 0.39 on corners, where the weakest baselines fall below 0.10. The embeddings (Figure 2) are consistent with this result: the AE and VAE encoders separate the two seeds into disjoint clusters rather than forming a continuous manifold, while JEPAWG maps both seeds onto a single smooth sheet. Table 1 reveals an asymmetry in how the methods respond to canonicalization. For the JEPAWG family, canonicalization is helpful (JEPAWG: $+0.093$, JEPAWG*: $+0.204$), consistent with the VICReg objective benefiting from a coupling-aligned representation rather than absorbing symmetry variation. For the reconstruction-trained baselines, canonicalization is actively harmful: raw-weight VAE and AE reach interpolation ESS of 0.910 and 0.915 respectively, dropping by 0.5–0.6 once canonicalized. These methods exploit the symmetry orbit as feature signal, and removing it strips the decoder of the variance it relies on. Even at this best raw-weight setting, however, the baselines lack the structural properties JEPAWG provides elsewhere (intrinsic dimension matching the coupling manifold, smooth extrapolation, and recovery of the phase boundary; Section 5.3). A single-seed variant of this experiment is reported in Section C.

Table 1. Decoded-flow ESS on the split-seed grid. The coupling plane is split into a training/interpolation window, four edge regions, and four corner regions, as illustrated in Figure 10. Interpolation: ESS estimation on 256 uniformly sampled couplings (m^2, λ) with $m^2 \in [-5.5, -1.5]$ and $\lambda \in [3.0, 6.0]$. Edges: mean±std over the four edge regions at $\delta = 1.5$ (32 points/region). Corners: mean±std over the four corner regions at $\delta = 1.5$ (32 points/region). Middle three columns use canonicalized weights; "Raw Interp." reports interpolation ESS without weight canonicalization. See Table 8 for the per-region breakdown.

Method	Can. Interp.	Can. Edges	Can. Corner	Raw Interp.
JEPAWG	$\mathbf{0.88\pm0.18}$	$\mathbf{0.56\pm0.19}$	$\mathbf{0.39\pm0.19}$	0.78 ± 0.22
JEPAWG*	0.61 ± 0.32	0.35 ± 0.12	0.25 ± 0.09	0.41 ± 0.33
VAE	0.40 ± 0.34	0.23 ± 0.07	0.13 ± 0.14	0.91 ± 0.18
AE	0.33 ± 0.33	0.15 ± 0.05	0.08 ± 0.07	$\mathbf{0.92\pm0.18}$
PCA	0.30 ± 0.31	0.09 ± 0.02	0.10 ± 0.07	0.75 ± 0.27

Table 2. Intrinsic dimension estimates of z_θ at encoder width H=128 and latent dimension d_z=16. Top-5 PCA denoising; mean ± standard deviation across 5 seeds.

	MLE (K=10)	PCA-95%
$\dim((m^2, \lambda))$	2.00	2
JEPAWG	$\mathbf{2.01\pm0.06}$	$\mathbf{2.0\pm0.0}$
JEPAWG*	2.36 ± 0.04	13.2 ± 0.8
VAE	4.88 ± 0.15	14.8 ± 0.4
AE	3.68 ± 0.09	11.6 ± 0.5
PCA	2.34 ± 0.01	15

5.3. Interpretability

Intrinsic dimension. We estimate the intrinsic dimension of each learned representation using the maximum likelihood estimator (MLE) of Levina & Bickel (2004) (details in Appendix B.3). Table 2 shows that JEPAWG is the only method whose MLE estimate matches the ground truth of d=2 (the dimension of the coupling space we used), and the only one whose latent variance is captured almost entirely by two principal axes (PCA-95%=2). The VAE and AE inflate the intrinsic dimension well beyond the physical ground truth, with MLE estimates of 4.88 and 3.68 respectively, indicating that their latent spaces retain structure unrelated to the coupling parameters.

Latent-space embedding. Figure 2 shows the first two principal components of each representation, colored by the Binder cumulant U_4; see Figure 11 for the full comparison across all five methods and additional colorings by m^2 and λ. For JEPAWG, PC_1 and PC_2 capture 98.5% of the latent variance and form a smooth sheet aligned with the coupling

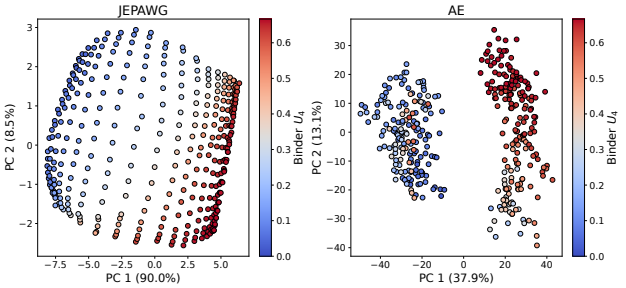

Figure 2. PCA of latent representations (d_z=16) for JEPAWG (left) and autoencoder (right), trained on 361 canonicalized flow checkpoints from two independently trained source flows. Points are colored by the Binder cumulant U_4. JEPAWG organizes the latent space along an axis aligned with the phase structure, whereas the autoencoder embedding exhibits less structure and splits into two disconnected components reflecting the seed boundary inherited from the 2-seed setup with which the training data are generated.

parameters; U_4, which is never seen during training, varies coherently along the same axes. The autoencoder places only 51.0% of variance in its leading two components, and neither coupling nor U_4 traces a coherent direction; instead, the two training seeds separate into distinct clusters, indicating that the encoder retains seed identity rather than learning a continuous representation of the physics. This concentration of the JEPAWG representation onto two physically relevant axes is consistent with the PCA-95%=2 result in Table 2. The smoothness of the JEPAWG embedding also holds in the single-seed setting (Section C).

Phase transition. We study how the parameter map $g \mapsto \theta(g)$ varies across coupling space by computing the pullback metric $G_{ij}(g) = \langle \partial_i \theta(g), \partial_j \theta(g) \rangle$, with $\langle \cdot, \cdot \rangle$ the Euclidean inner product on the weight space and derivatives estimated by finite differences in (m^2, λ). The area element $A(g) = \sqrt{\det G(g)}$ quantifies how fast the learned parameters deform with the bare couplings; a peak is expected near the critical line, where $\theta(g)$ must accommodate rapid changes in the physics.

To localize where this deformation originates within the architecture, we partition the weights θ into six functional groups (see Section B.1): two that act non-locally on configurations (the zero mode scaling $\theta_{\text{zero_mode}}$ and the Fourier scaling θ_{scale}) and four bulk groups (θ_{conv}, $\theta_{\text{feat_super}}$, $\theta_{\text{time_kernel}}$, $\theta_{\text{feat_map}}$) that act mostly locally. Computing the area element per group, $A_c(g) = \sqrt{\det G_c(g)}$, we find that only the non-local groups develop a sharp ridge tracking the critical line, while the bulk groups remain featureless (Figure 3). Table 3 confirms this quantitatively: $\theta_{\text{zero_mode}}$ and θ_{scale} exhibit positive correlations between $A_c(g)$ and all three criticality observables ξ, $|\nabla U_4|$, and χ. The bulk groups, by contrast, show no positive correlation with any

Table 3. Pearson correlation between the per-group area element $A_c(g) = \sqrt{\det G_c(g)}$ and physical observables across the (m^2, λ) grid. Bold entries mark $r \geq 0.5$. θ_{bulk} aggregates the four bulk groups; the column reports the range across these groups. Per-group values are given in Table 7.

Observable	$\theta_{\text{zero_mode}}$	θ_{scale}	θ_{bulk} (range)
ξ	$+\mathbf{0.57}$	$+0.36$	$[-0.26, -0.12]$
$\lvert\nabla U_4\rvert$	$+\mathbf{0.61}$	$+\mathbf{0.56}$	$[-0.11, +0.07]$
χ	$+\mathbf{0.67}$	$+\mathbf{0.51}$	$[-0.24, -0.06]$
$\lvert m\rvert$	-0.04	-0.10	$[-0.05, +0.14]$
λ	-0.11	-0.14	$[-0.76, -0.59]$

criticality observable; their dominant signal is a strong negative correlation with λ ($r \in [-0.76, -0.59]$), indicating that they track the quartic coupling rather than proximity to the transition.

The physics is consistent with this observation: a transition driven by the divergence of the correlation length must be absorbed by the long-range degrees of freedom of the model. Restricted to those non-local groups, the geometry of the parameter map recovers the phase boundary without drawing any samples.

5.4. Cross-lattice RG structure

Beyond organizing weight vectors by phase, we investigate whether the learned latent representation also captures aspects of renormalization-group (RG) structure across lattice sizes. Near a continuous phase transition, finite-size scaling (FSS) predicts that observables depend primarily on a small number of relevant scaling fields. At fixed λ and close to the transition, we define the relevant scaling field

$$t = a(\lambda)\left(m^2 - m_c^2(\lambda)\right), \tag{3}$$

which measures the distance from the critical line $m^2 = m_c^2(\lambda)$, with $a(\lambda)$ a non-universal normalization. Finite-size scaling predicts that the dimensionless combination

$$x = tL^{1/\nu} \tag{4}$$

with ν the correlation-length critical exponent, is the physical coordinate whose meaning remains approximately invariant under changes of lattice size. For the two-dimensional ϕ^4 theory studied here, the transition is known to belong to the 2D Ising universality class, for which $\nu = 1$. Hence we postulate that, if the embedding space truly encodes physical information, the embedding of the weights $\theta_{g_{L=6}}$ associated with the couplings $g_{L=6}$ at $L = 6$ should be close in embedding space to $\theta_{g_{L=8}}$ where $g_{L=8}$ corresponds to the same x as the given $g_{L=6}$.

To test this hypothesis, we define an FSS-predicted partner point between two lattice sizes L_1 and L_2 at fixed quartic coupling λ by requiring the scaling variable of Eq. (4) to be

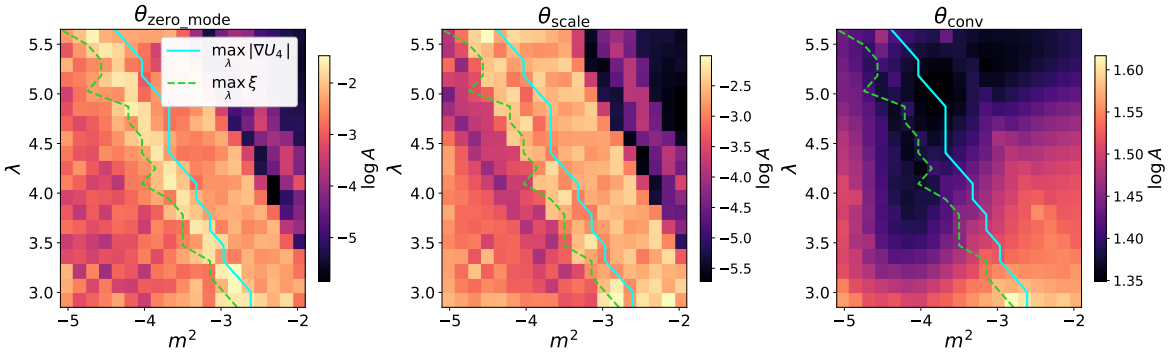

Figure 3. Pullback area $A(g)$ on the (m^2, λ) grid for the 8^2 lattice for three parameter groups; $\theta_{\text{zero_mode}}$, θ_{scale} and θ_{conv}. Overlaid are the per-row maxima of $|\nabla U_4|$ (cyan, solid) and ξ (lime, dashed), serving as transition proxies. The parameter groups ($\theta_{\text{zero_mode}}$, θ_{scale}) exhibit a sharp A-ridge tracking the critical line.

preserved across lattices, $t_1 L_1^{1/\nu} = t_2 L_2^{1/\nu}$. Solving for the source mass yields

$$m^2_{\text{pred}} = m_c^2(\lambda; L_1) + \left(m_2^2 - m_c^2(\lambda; L_2)\right) \left(\frac{L_2}{L_1}\right)^{1/\nu} . \quad (5)$$

We test whether this relation is reflected in the learned latent space by comparing, for a given target point (m_2^2, λ) on lattice L_2, two matching procedures: the naive *same-(bare) coupling* matching ($m_1^2 = m_2^2$, no RG rescaling) and *FSS* matching (m_1^2 determined by Eq. (5)) using different methods for estimating the critical mass, which we outline in Appendix D.2. For each pairing, we compute the mean latent-space distance between encoded weight vectors, $d_{\text{mean}}(\nu) = \frac{1}{N}\sum_{i=1}^{N} \|z_i^{(1)} - z_i^{(2)}\|_2$, where $z_i^{(j)}$ denotes embedded weight vectors for lattice size L_j. We scan $\nu \in [0.2, 3.5]$ and report $\nu^* = \arg\min_\nu d_{\text{mean}}(\nu)$, with uncertainty estimated by a bootstrap over the N coupling points (Appendix D.3).

Figure 4 shows the latent-distance curves obtained by scanning the scaling exponent ν for the single-step coarse-graining mappings $8 \to 6$ and $10 \to 8$. Table 4 summarizes the corresponding numerical results. In both cases, the FSS-based pairing produces lower distances in the embedding space than same-coupling matching. The minimizers ν^* range from 0.70 to 1.56 across the two single-step mappings and the two critical-line estimators, bracketing the analytic Ising value $\nu = 1$; this spread reflects the sensitivity to the critical-line estimate and the finite-size corrections expected at these small lattices. That the finite-size scaling variable organizes the latent geometry at all, and that the signal vanishes when the coupling–weight correspondence is shuffled (Section G), indicates that the learned representation has captured physical information consistent with the renormalization-group scaling law.

Results for the remaining coarse-graining pairing ($10 \to 6$) are shown in Figure 8 and Table 6. The improvement is weaker for this two-step mapping, consistent with larger

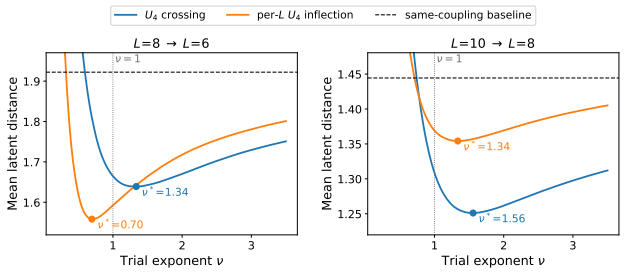

Figure 4. Mean latent distance $d_{\text{mean}}(\nu)$ under FSS-based pairing for the lattice-size mappings $8 \to 6$ and $10 \to 8$. The curves differ in their method for estimating the critical line, further explained in Appendix D.2. The horizontal dashed lines indicate the corresponding same-coupling baseline distances d_{same}. Vertical dotted lines mark the expected Ising value $\nu = 1$. For both lattice-size mappings, FSS-based pairing lowers the latent distance and produces minima near the expected scaling exponent.

finite-size corrections with increasing scale ratio $L_1/L_2 > 1$. Note that this analysis is meaningful only in the coarse-graining direction: both lattice sizes were trained on the same grid of bare couplings, so FSS pairing in the inverse direction maps target points to source couplings that lie outside the training grid, where the encoder has not seen weights.

Additional lattice-pair scans, control directions, and comparisons between critical-line estimators are reported in Appendix D.3.

6. Conclusion

We introduced JEPAWG, a JEPA-based weight generator that maps lattice couplings directly to flow parameters via a learned latent space, and used it to study what neural networks learn about field theories from their weights alone. On 2D ϕ^4 theory, the JEPAWG weight embedding recovers the intrinsic dimensionality of the coupling manifold. Beyond this, the geometry of the trained parameters exposes

Table 4. Latent-space finite-size scaling results for single-step coarse-graining mappings ($L_1 \to L_2 < L_1$). m_c^2 estim.: method used to estimate the critical mass (Appendix D.2). d_{same}: mean latent distance under same-coupling matching. d^*_{FSS}: minimum latent distance under FSS matching. ν^*: minimizing exponent, with 95% CI from bootstrap resampling of the N=361 coupling points (see Appendix D.3).

Mapping	m_c^2 estim.	d_{same}	d^*_{FSS}	ν^* [95% CI]
$8 \to 6$	U_4 crossing	1.92	1.64	1.34 $[1.17, 1.46]$
$8 \to 6$	per-L U_4	1.92	1.56	0.70 $[0.64, 0.80]$
$10 \to 8$	U_4 crossing	1.44	1.25	1.56 $[1.41, 1.75]$
$10 \to 8$	per-L U_4	1.44	1.35	1.34 $[1.16, 1.59]$

physical structure: the pullback metric exhibits a ridge that traces the critical line of the theory, localized to the network parameters that are responsible for non-local transformations on the generated configurations. Across lattice sizes, the encoder organizes network weights according to the finite-size scaling variable predicted by the renormalization group. More quantitatively, the scaling exponent that yields the most coherent latent geometry is broadly consistent with the exact $\nu = 1$ prediction of the Ising model. Notably, none of these properties are supervised by physical observables; they emerge from the geometry of weight space and that of the JEPA embedding space alone. As a generator, JEPAWG also produces functional flows at unseen couplings in a multi-seed setting where PCA, AE, and VAE baselines collapse.

From this work, the picture that emerges is a simple one: the geometry of trained weights reflects the geometry of the physical theory, and it concentrates on the components of the network where the physics actually lives. These results validate our JEPAWG framework and support our broader claim on the effectiveness of physics as natural synthetic data for developing interpretable AI.

Limitations and Future Directions. Our experiments are confined to a single 2D theory at small lattices, and our interpretability claims rest on agreement with one universality class. We see two natural extensions that we are currently pursuing. The first is *scale*: our experiments are confined to small lattices ($L \leq 10$) due to the cost of training the underlying flows, and extending the analysis to larger lattices will further validate our results and in particular sharpen the weight-space RG results. The second is *diversity*: applying JEPAWG to other quantum field theories, or other scientific systems would test whether the framework remains effective in richer physical settings. In favorable cases, the same approach could even uncover qualitative features of systems that are not yet fully understood theoretically, turning JEPAWG into an experimental tool.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

The work of Julian Ebelt, Tobias Göbel, Zier Mensch, Mathis Gerdes, and Miranda Cheng was supported by the AI-application grant AS-IAIA-114-M02 from Academia Sinica, and the Vici grant (number VI.C.232.117) from the Dutch Research Council (NWO). Mathis Gerdes was also partially supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/) Finally, we thank the Wistron Computing Power Donation Program for compute support.

References

Albergo, M. S., Kanwar, G., and Shanahan, P. E. Flow-based generative models for markov chain monte carlo in lattice field theory. *Phys. Rev. D*, 100: 034515, Aug 2019. doi: 10.1103/PhysRevD.100.034515. URL https://link.aps.org/doi/10.1103/PhysRevD.100.034515.

Albergo, M. S., Boyda, D., Hackett, D. C., Kanwar, G., Cranmer, K., Racanière, S., Rezende, D. J., and Shanahan, P. E. Introduction to normalizing flows for lattice field theory. *arXiv preprint arXiv:2101.08176*, 2021.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. URL https://arxiv.org/abs/2105.04906.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Chauhan, V. K., Zhou, J., Lu, P., Molaei, S., and Clifton, D. A. A brief review of hypernetworks in deep learning. *Artificial Intelligence Review*, 57(9), 2024.

Chen, D., Shukor, M., Moutakanni, T., Chung, W., Yu, J., Kasarla, T., Bolourchi, A., LeCun, Y., and Fung, P.

VL-JEPA: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

Cheng, M. C. and Stratikopoulou, N. Lecture notes on normalizing flows for lattice quantum field theories. *SciPost Physics Lecture Notes*, pp. 110, 2026.

Facco, E., d'Errico, M., Rodriguez, A., and Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017.

Gerdes, M. and Cheng, M. C. N. Bijx: Bijections and normalizing flows with jax/nnx. *Journal of Open Source Software*, 10(116):9521, 2025. doi: 10.21105/joss.09521. URL https://doi.org/10.21105/joss.09521.

Gerdes, M. and Cheng, M. C. N. Analytic bijections for smooth and interpretable normalizing flows, 2026. URL https://arxiv.org/abs/2601.10774.

Gerdes, M., de Haan, P., Rainone, C., Bondesan, R., and Cheng, M. C. Learning lattice quantum field theories with equivariant continuous flows. *SciPost Physics*, 15 (6):238, 2023.

Ha, D., Dai, A., and Le, Q. V. HyperNetworks. In *International Conference on Learning Representations*, 2017.

Kanwar, G. Flow-based sampling for lattice field theories. *arXiv preprint arXiv:2401.01297*, 2024.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2014.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

Krueger, D., Huang, C.-W., Islam, R., Turner, R., Lacoste, A., and Courville, A. Bayesian hypernetworks. *arXiv preprint arXiv:1710.04759*, 2017.

LeCun, Y. A path towards autonomous machine intelligence. *Open Review*, 62(1), 2022.

Levina, E. and Bickel, P. J. Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*, 17, 2004.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing Flows for Probabilistic Modeling and Inference, 2021. URL http://arxiv.org/abs/1912.02762.

Räuker, T., Ho, A., Casper, S., and Hadfield-Menell, D. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks, 2023. URL https://arxiv.org/abs/2207.13243.

Schmidhuber, J. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., Casper, S., Tegmark, M., Saunders, W., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., and McGrath, T. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Tabak, E. G. and Vanden-Eijnden, E. Density estimation by dual ascent of the log-likelihood. 8(1): 217–233, 2010. ISSN 1539-6746, 1945-0796. URL https://projecteuclid.org/journals/communications-in-mathematical-sciences/volume-8/issue-1/Density-estimation-by-dual-ascent-of-the-log-like cms/1266935020.full.

von Oswald, J., Henning, C., Grewe, B. F., and Sacramento, J. Continual learning with hypernetworks. In *International Conference on Learning Representations*, 2020.

A. Experimental Details

A.1. Baseline Methods

Weight-space PCA. We perform PCA on the standardized training set $\{\tilde{\theta}_i\}_{i=1}^N$, retaining the top-$k$ principal components $U_k \in \mathbb{R}^{d_\theta \times k}$ and representing each parameter vector by its projection $\mathbf{z}_i = U_k^\top \tilde{\theta}_i$.

Autoencoder and variational autoencoder. We train a symmetric MLP encoder $e_\phi : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_z}$ and decoder $d_\xi : \mathbb{R}^{d_z} \to \mathbb{R}^{d_\theta}$ with a Gaussian reparameterized latent $z = \mu_\phi(\tilde{\theta}) + \sigma_\phi(\tilde{\theta}) \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, using the VAE objective (Kingma & Welling, 2014)

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left\| d_\xi(z_i) - \tilde{\theta}_i \right\|_2^2 + \beta \, D_{\mathrm{KL}}\big(q_\phi(z|\tilde{\theta}_i) \| \mathcal{N}(0, I)\big). \tag{6}$$

Setting $\beta = 0$ gives the deterministic autoencoder (AE); $\beta = 1$ gives the standard VAE. In both cases, the embedding of point i is the encoder mean $\mathbf{z}_i = \mu_\phi(\tilde{\theta}_i)$, and no coupling information is used during training.

Conditional generation. At inference, all three baselines introduce coupling information through a shared post-hoc step: we fit a 2-layer MLP from couplings to the learned weight embeddings. Given a new coupling $g^\star$, we predict $\hat{\mathbf{z}}(g^\star)$ and apply the corresponding inverse to recover parameters.

In contrast, JEPAWG requires no post-hoc fit: the coupling encoder E_g and predictor P are trained jointly with the weight encoder via the VICReg invariance, so the learned composition $P \circ E_g$ plays the role of the post-hoc regressor used by the baselines.

A.2. VICReg objective

For a batch of N paired samples, we denote by $Z_x, Z_y \in \mathbb{R}^{N \times d_z}$ the stacked latent representations produced by the two encoders, where each row corresponds to one sample and each column to one latent dimension. The predictor is applied row-wise per sample, yielding $\hat{Z}_y = h(Z_x) \in \mathbb{R}^{N \times d_z}$.

The VICReg loss (Bardes et al., 2022) is defined as

$$\mathcal{L} = \alpha \, s(Z_x, Z_y) + \beta \big[v(Z_x) + v(Z_y) \big] + \gamma \big[c(Z_x) + c(Z_y) \big], \tag{7}$$

with weights $\alpha = \beta = 25$ and $\gamma = 1$.

Invariance term. The invariance term encourages alignment between predicted and target representations,

$$s(Z_x, Z_y) = \frac{1}{N} \sum_{i=1}^{N} \| \hat{\mathbf{z}}_{y,i} - \mathbf{z}_{y,i} \|_2^2 \,, \tag{8}$$

where $\hat{\mathbf{z}}_{y,i}$ and $\mathbf{z}_{y,i}$ denote the i-th rows of $\hat{Z}_y$ and Z_y, respectively.

Variance term. Let $z^j \in \mathbb{R}^N$ denote the j-th column of Z, i.e. the values of latent dimension j across the batch. The variance term penalizes dimensions whose empirical standard deviation falls below one,

$$v(Z) = \frac{1}{d_z} \sum_{j=1}^{d_z} \max\Big(0, \; 1 - \sqrt{\mathrm{Var}(z^j) + \varepsilon}\Big) \,, \tag{9}$$

where

$$\mathrm{Var}(z^j) = \frac{1}{N} \sum_{i=1}^{N} \big(Z_{ij} - \mu_j\big)^2, \quad \mu_j = \frac{1}{N} \sum_{i=1}^{N} Z_{ij}. \tag{10}$$

This term prevents representational collapse by ensuring that each latent dimension varies across the batch.

Covariance term. Let $\bar{Z} \in \mathbb{R}^{N \times d_z}$ denote the mean-centered representations,

$$\bar{Z}_{ij} = Z_{ij} - \mu_j. \tag{11}$$

The covariance matrix of the latent representation is then

$$C(Z) = \frac{1}{N-1} \bar{Z}^\top \bar{Z} \in \mathbb{R}^{d_z \times d_z}. \tag{12}$$

The covariance term penalizes correlations between distinct latent dimensions,

$$c(Z) = \frac{1}{d_z} \sum_{i \neq j} \big[C(Z)\big]_{ij}^2. \tag{13}$$

This encourages different latent dimensions to capture independent factors of variation.

Covariance ablation. Throughout the main text we set $\gamma = 0$, omitting the covariance term so the latent space can concentrate onto fewer dimensions when the data manifold admits a lower-dimensional structure (Section 4.1). For comparison, we also report results for the standard VICReg setting $\gamma = 1$, denoted JEPAWG* in the tables. The JEPAWG configuration thus refers to the covariance-off variant combined with the decoder D_θ.

A.3. Weight symmetries and canonicalization

The factorization (17) introduces bond indices d' and f' that are contracted out, so the triple $(\widetilde{W}, W^K, W^H)$ is determined only up to invertible changes of basis: for any $M \in \mathrm{GL}(F')$, replacing $W^H \mapsto M\, W^H$ and contracting $\widetilde{W}$ with $(M^{-1})^\top$ on f' leaves the physical weight tensor W_{xydf} invariant, and analogously for $\mathrm{GL}(D')$ on the time bond. Two further symmetries act on the Fourier index $f \geq 2$: permutations $\sigma \in S_{F-1}$ of the pairs $(\omega_f, W^H_{\cdot,f})$, and paired sign flips $\omega_f \mapsto -\omega_f$, $W^H_{\cdot,f} \mapsto -W^H_{\cdot,f}$ (since $\sin$ is odd). Together these span a symmetry group

$$G \;=\; \mathrm{GL}(F') \times \mathrm{GL}(D') \times S_{F-1} \times \{\pm 1\}^{F-1} \tag{14}$$

of continuous dimension $(F')^2 + (D')^2$, under which distinct parameter vectors θ can encode the same input-output map.

We select a unique orbit representative as follows. Compute the QR decomposition $W^H = QR$ with $Q \in \mathbb{R}^{F' \times F'}$ orthogonal and $R \in \mathbb{R}^{F' \times F}$ upper trapezoidal, using the convention $\mathrm{diag}(R) \geq 0$ to fix the sign ambiguity. Since $Q \in \mathrm{O}(F') \subset \mathrm{GL}(F')$, the map $W^H \mapsto Q^\top W^H = R$ is a valid symmetry transformation. The first F' columns of R form an upper-triangular matrix R_1 with positive diagonal, so a second transformation $R \mapsto R_1^{-1} R = [\, I_{F'} \mid R_1^{-1} R_2 \,]$ brings W^H to identity-block form. This selects the $\mathrm{GL}(F')$ orbit representative uniquely: if two matrices on the same orbit are both in this form, $[\, I \mid A \,] = M [\, I \mid A' \,] = [\, M \mid MA' \,]$, hence $M = I$ and hence $A = A'$. The combined transformation composes on the $\widetilde{W}$ side as $\widetilde{W} \mapsto \widetilde{W} \cdot R_1^\top Q^\top$ on f', which preserves (17). The same construction applied to W^K selects the $\mathrm{GL}(D')$ orbit representative. Finally, the discrete factor $S_{F-1} \times \{\pm 1\}^{F-1}$ is canonicalized by sorting $\{|\omega_f|\}_{f \geq 2}$ in ascending order and flipping signs to enforce $\omega_f \geq 0$, with columns of W^H permuted and sign-flipped accordingly.

B. Data Generation

The dataset analyzed in this work is a collection of *trained network parameters* of the coupling-conditional flow of Section 3.1, each near-optimal as a sampler for the two-dimensional ϕ^4 lattice field theory at a different point of the coupling space $g = (m^2, \lambda)$. We generate it in two stages: a coupling-conditional pre-training that produces a single model covering the whole coupling box (Section B.1), followed by per-point fine-tuning on a regular grid. The output is an $N \times N$ array of parameter vectors $\{\theta_{ij}\}_{i,j=1}^N$, all in the same architecture and indexed by the underlying coupling $g_{ij} = (m_i^2, \lambda_j)$.

Throughout the paper, we work on a periodic lattice of sizes $L \in \{6, 8, 10\}$ and restrict the couplings to the box $\Omega = [-5.5, -1.5] \times [3.0, 6.0] \ni (m^2, \lambda)$, which straddles the symmetric and broken phases of the model. The target density is $p_g(\phi) \propto \exp[-S_g(\phi)]$ with the standard nearest-neighbor ϕ^4 action.

B.1. Coupling-conditional pre-training

We build on the continuous normalizing flow architecture for sampling from the ϕ^4 lattice field theory (Gerdes et al., 2023; Gerdes & Cheng, 2025). The flow is preceded by a Fourier-space preconditioner: a learnable per-mode diagonal rescaling in

Fourier space, whose form is motivated by the free-theory propagator, which diagonalizes the quadratic action. Writing $\hat{\phi}(k)$ for the FFT of ϕ, the preconditioner takes the form

$$\hat{\phi}(k) \;\mapsto\; e^{s_k(g)}\,\hat{\phi}(k)\,, \qquad k \neq 0\,, \tag{15}$$

with one learnable real scalar s_k per reduced Fourier slot, while the zero mode $\hat{\phi}(0) \propto \sum_x \phi(x)$ is mapped through a separate stack of 1-D $\sinh$-based bijections (Gerdes & Cheng, 2026) followed by a global scaling. This diagonal rescaling has exactly the form that reproduces the Gaussian per-mode variance $1/(\hat{k}^2 + m^2)$ of the quadratic action; rather than fixing the s_k to those free values, we let them be learned, so the preconditioner can settle at the effective per-mode variances of the full interacting theory. In the conditioned model both s_k and the parameters of the zero-mode stack are stored as stacks of A copies and contracted against the embedding $J(g)$, so the preconditioner becomes a smooth function of the couplings. Note that the number of parameters in this layer depends on the lattice size L. We therefore align networks across $L \in \{6, 8, 10\}$ to a common parameter shape by zero-padding smaller lattices and truncating larger ones in Fourier space.

The consecutive neural ODE then defines an invertible map $f_\theta\colon \mathbb{R}^{L^2} \to \mathbb{R}^{L^2}$ as the solution to a vector field of the form

$$\frac{d\phi(t)_x}{dt} = \sum_{y,d,f} W_{xydf}\, K(t)_d\, H(\phi(t)_y)_f\,, \tag{16}$$

where H is a field basis of dimension F defined by $H(\phi)_1 = \phi$ and $H(\phi)_f = \sin(\omega_f \phi)$ for $f \geq 2$, with learnable frequencies $\omega \in \mathbb{R}^{F-1}$, K a Fourier time kernel of dimension D, and W a learnable weight tensor constrained by the lattice symmetry group $C_L^2 \rtimes D_4$. To reduce the number of free parameters, W is parameterized as

$$W_{xydf} = \sum_{d',f'} \widetilde{W}_{xyd'f'}\, W^K_{d'd}\, W^H_{f'f}\,. \tag{17}$$

Parameter groups. The flow's parameter vector θ decomposes naturally into six groups, referenced by name in Section 5.3 and Table 3:

- $\theta_{\text{zero_mode}}$: parameters of the dedicated zero-mode flow acting on $\hat{\phi}(0)$, governing the global magnetization mode and its bimodal distribution in the broken phase.
- θ_{scale}: the per-mode preconditioner scalars s_k for $k \neq 0$, which set the overall variance of each Fourier mode.
- θ_{conv}: the spatial weight tensor $\widetilde{W}_{xyd'f'}$ in (17), which acts as a learnable convolution on the lattice.
- $\theta_{\text{time_kernel}}$: the time-kernel factor $W^K_{d'd}$ in (17).
- $\theta_{\text{feat_super}}$: the feature-superposition factor $W^H_{f'f}$ in (17).
- $\theta_{\text{feat_map}}$: the learned frequencies $\omega \in \mathbb{R}^{F-1}$ that define the field basis H.

The first two groups parametrize global, non-local degrees of freedom (zero mode and per-mode scale); the remaining four parametrize the local, mode-mixing operator inside the neural ODE.

To learn a continuous family of theories parametrized by couplings $g = (m^2, \lambda)$, each factor in (17) is extended with an additional index contracted against an embedding $J(g) \in \mathbb{R}^A$, making the weight tensor $W_{xydf}(g)$ a smooth function of the theory parameters. At fixed couplings, the conditioned flow is fully determined by a parameter vector $\theta(g)$ obtained by flattening the factors $(\widetilde{W}, W^K, W^H)$ and the learned frequencies ω, together with the preconditioner parameters (θ_{scale}, $\theta_{\text{zero_mode}}$). The JEPAWG weight encoder E_θ is applied to the neural ODE factors $(\widetilde{W}, W^K, W^H, \omega)$ only ($d_\theta = 19{,}689$); the preconditioner is excluded from the embedding because it is near-invariant across fine-tuned couplings and is reattached at generation from a reference checkpoint.

Embedding. The embedding $J(g) \in \mathbb{R}^A$ that conditions the factors of (17) is implemented as sinusoidal positional features applied independently to each coupling, concatenated, and projected by a single (orthogonally initialized) linear layer. Concretely, for $g_a \in \{m^2, \lambda\}$ we form $\mathrm{PE}_a(g_a) = [\sin(\boldsymbol{\omega}_a g_a), \cos(\boldsymbol{\omega}_a g_a), g_a]$, with frequencies $\boldsymbol{\omega}_a$ on a logarithmic ladder up to a maximum position $g_a^{\max}$, then set $J(g) = W[\mathrm{PE}_1(m^2) \,\|\, \mathrm{PE}_2(\lambda)] + \boldsymbol{b}$. We use $A = 16$, 40 Fourier features per coupling, and $g^{\max} = (2.1,\ 4.0)$.

Objective. At each gradient step, we draw M couplings $g^{(1)}, \ldots, g^{(M)}$ uniformly in Ω, and for each, B samples $\phi^{(b)} \sim q_{\theta(g)}$ from the conditional flow. The loss is the sample-mean reverse Kullback–Leibler divergence between flow and target,

$$\mathcal{L} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{B}\sum_{b=1}^{B}\Big[\log q_{\theta(g^{(m)})}\big(\phi^{(m,b)}\big) + S_{g^{(m)}}\big(\phi^{(m,b)}\big)\Big]. \tag{18}$$

Gradients flow through both the per-factor base tensors and the embedding parameters $W, \boldsymbol{b}$ via the conditioning of (17).

Optimization. We optimize with Adam (β_1=0.85, β_2=0.9), initial learning rate 3×10^{-4}, decayed exponentially by $\frac{1}{2}$ every 8000 steps. We train for 30,000 steps with $M = 8$ couplings per step and $B = 32$ samples per coupling (total batch size 256). The neural ODE is integrated with `diffrax` using the `Tsit5` adaptive stepper (atol = rtol = 10^{-5}). Convergence is monitored on a fixed set of reference couplings via the per-coupling effective sample size,

$$\mathrm{ESS}(g) = (\sum_b w^{(b)})^2/(N\sum_b (w^{(b)})^2) \tag{19}$$

with $w^{(b)} = p_g(\phi^{(b)})/q_{\theta(g)}(\phi^{(b)})$.

Fine-tuning loop. The unconditional flow is fine-tuned at fixed g_{ij} by minimizing the single-coupling reverse KL,

$$\mathcal{L}_{ij}(\theta) = \frac{1}{B}\sum_{b=1}^{B}\Big[\log q_\theta(\phi^{(b)}) + S_{g_{ij}}(\phi^{(b)})\Big], \quad \phi^{(b)} \sim q_\theta. \tag{20}$$

We use Adam with initial learning rate 5×10^{-4} (otherwise identical betas), exponential decay with half-life 2000 steps, batch size $B = 256$, and a step budget of $T_{\max} = 1000$. Every 25 steps we estimate the ESS on 2048 samples; training stops as soon as

$$\mathrm{ESS}(g_{ij}) \;\geq\; \tau, \qquad \tau = 0.95, \tag{21}$$

or after $T_{\max}$ steps, whichever comes first. The early-stopping rule makes the per-point training cost itself an observable: points where the conditional initialization is already good require few or no gradient steps, while points near difficult regions of Ω exhaust the budget.

Compute. All experiments ran on a single NVIDIA A100 or H100 GPU. Conditional flow pre-training takes a few hours; per-coupling fine-tuning, JEPAWG training, and the baselines each take on the order of minutes per run.

B.2. Resulting dataset

The procedure above yields, for the box $\Omega = [-5.5, -1.5] \times [3.0, 6.0]$ on a 6×6 lattice, a 19×19 grid of fine-tuned flow parameter vectors, all in a common architecture and all initialized from the same conditional model. By construction the parameters vary smoothly across Ω, a property inherited from the conditional embedding and only mildly perturbed by the per-point fine-tuning, which is the structural prerequisite for the analyses presented in the main text.

Figure 5 summarizes the physical content of the dataset: the Binder cumulant U_4 tracks the symmetric-to-broken transition, and the susceptibility χ peaks along the same crossover ridge.

B.3. Data Analysis Tools

Intrinsic dimension estimation. We estimate the intrinsic dimension of each learned representation using the *Maximum Likelihood Estimator* (MLE) (Levina & Bickel, 2004), which assumes the data lies near a d-dimensional manifold and estimates d from the rate at which the number of neighbors grows with radius. For a point x_i with K-th nearest neighbor distance $R_K(x_i)$ and j-th nearest neighbor distance $R_j(x_i)$, the local dimension estimate is

$$\hat{d}_K(x_i) = \left(\frac{1}{K-1}\sum_{j=1}^{K-1}\log\frac{R_K(x_i)}{R_j(x_i)}\right)^{-1}, \tag{22}$$

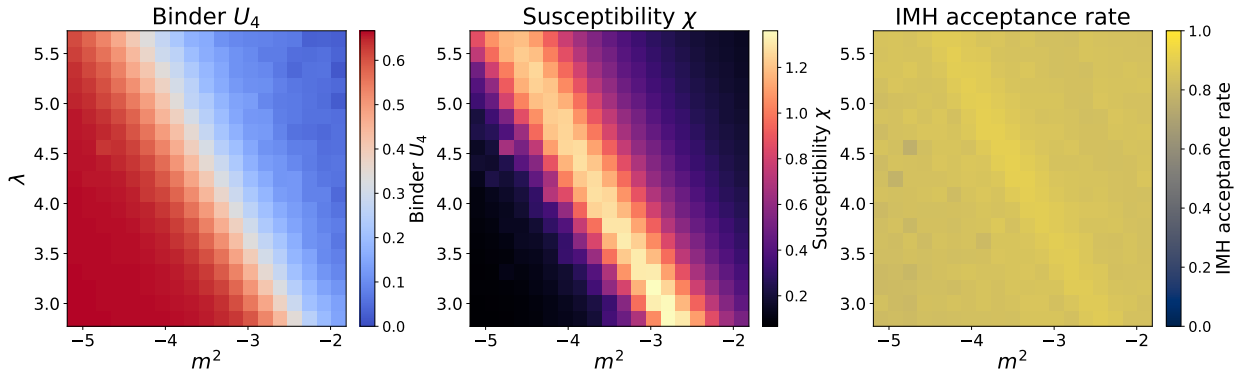

Figure 5. Phase-structure observables on the 19×19 fine-tuned grid for the 8×8 lattice ($m^2 \in [-5.5, -1.5]$, $\lambda \in [3.0, 6.0]$). **Left:** Binder cumulant $U_4 = 1 - \langle\bar{\phi}^4\rangle/(3\langle\bar{\phi}^2\rangle^2)$. The high-$U_4$ ridge marks the broken phase, whereas the low-U_4 basin ($U_4 \to 0$) corresponds to the symmetric phase. The crossover between them traces out the critical line. **Center:** Susceptibility $\chi = L^2\big(\langle\bar{\phi}^2\rangle - \langle|\bar{\phi}|\rangle^2\big)$ peaks along the same crossover, as expected at a continuous phase transition. **Right:** Independent Metropolis Hastings acceptance rate at each grid point, computed from 4096 importance proposals from the fine-tuned conditional flow with a 256-step burn-in. Acceptance stays in the 0.76–0.91 range across the grid, confirming that the fine-tuned flow proposal remains a good match to the target distribution everywhere on the grid (no localized quality drops along the critical line). All quantities are estimated from samples drawn from the fine-tuned single-coupling flows, with further unbiasing using Independent Metropolis Hastings.

and the global estimate is the average over all points. We use $K = 10$ throughout. We also considered the TwoNN estimator (Facco et al., 2017) but found it unreliable on our data, producing estimates that did not stabilize with subsample size; we therefore report only MLE results. The estimator is applied after projecting each representation onto its top-5 PCA components, a standard denoising step for noisy high-dimensional manifolds.

C. Single Seed Experiment

We repeat the weight generation experiment using flow parameters from a single conditional flow, removing the seed-boundary discontinuity present in the multi-seed setting of Section 5.2. The training data consists of a 19×19 grid of fine-tuned flows over the coupling box Ω, all initialized from the same conditional model. Table 5 reports the decoded-flow ESS across 9 regions of the coupling plane, averaged over 5 random seeds.

In the single-seed setting, all methods benefit from the smooth weight manifold. On the interior, JEPAWG and JEPAWG* both achieve ESS above 0.97, and even PCA with an MLP head reaches 0.95. The baselines that failed in the multi-seed setting (VAE: 0.74, AE: 0.78, PCA: 0.79) now produce functional flows, confirming that their earlier failure was due to the seed discontinuity rather than insufficient model capacity.

On extrapolation, JEPAWG remains the strongest method overall (Figure 7). Along the λ_{hi} edge it reaches 0.92, and at the $(m^2_{\text{hi}}, \lambda_{\text{hi}})$ corner 0.73, both substantially above the baselines. The $(m^2_{\text{lo}}, \lambda_{\text{lo}})$ corner remains difficult for all methods (ESS ≤ 0.07). JEPAWG* performs comparably to JEPAWG on the interior but falls behind on several extrapolation regions, consistent with the observation that the covariance-off variant concentrates its representation onto fewer latent dimensions and extrapolates more smoothly. Embeddings for all methods are shown in Figure 6.

D. Finite-size scaling details

This appendix provides additional details for the finite-size scaling (FSS) analysis discussed in Section 5.4, including the derivation of the lattice-size matching rule, critical-line estimation procedures, and additional latent-distance scans for additional coarse-graining mappings.

D.1. Derivation of the FSS partner map

Near a continuous phase transition, finite-size scaling predicts that observables depend primarily on the scaling combination

$$x = tL^{1/\nu}, \tag{23}$$

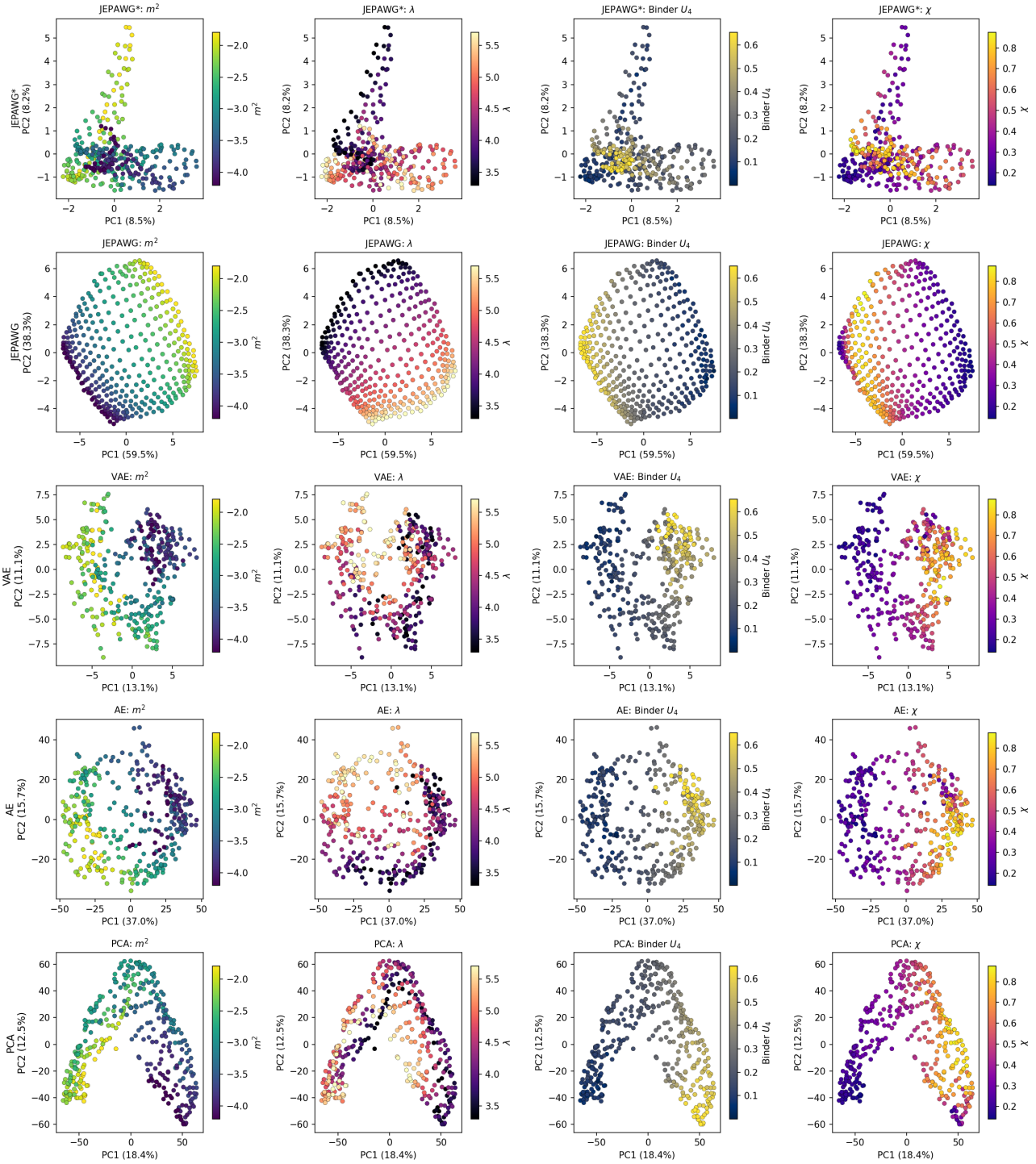

Figure 6. Latent-space embeddings (first two PCs) for JEPAWG*, JEPAWG, VAE, AE, and PCA (top to bottom) in the single-seed setting, colored by m^2, λ, and the Binder cumulant U_4 (left to right).

Table 5. Single-seed decoded-flow ESS across 9 regions of the coupling plane. Interior: training window; edges and corners: δ=1.5 extrapolation. Each cell is the mean over 60 uniformly sampled evaluation couplings, $\pm$ standard deviation across 5 random seeds. JEPAWG* refers to the covariance-on ablation ($\gamma = 1$); JEPAWG is the main configuration ($\gamma = 0$). JEPAWG* and JEPAWG use the trained coupling predictor $P \circ E_g$; PCA+MLP uses a 2-layer MLP fitted post-hoc on frozen PCA coordinates; PCA, VAE, and AE use a linear least-squares head $g \to \mathbf{z}$. VAE uses $\beta = 1$, AE uses $\beta = 0$.

Region	PCA	PCA+MLP	JEPAWG*	JEPAWG	VAE	AE
interior	0.793±0.033	0.947±0.007	0.973±0.002	**0.976±0.001**	0.737±0.030	0.775±0.065
m^2_{lo}	0.205±0.148	0.287±0.056	0.430±0.105	**0.440±0.046**	0.148±0.070	0.344±0.165
m^2_{hi}	0.393±0.157	0.400±0.112	0.667±0.134	**0.820±0.038**	0.441±0.047	0.395±0.201
λ_{lo}	0.230±0.049	0.397±0.040	0.386±0.056	**0.414±0.045**	0.167±0.043	0.174±0.021
λ_{hi}	0.511±0.107	0.687±0.074	0.825±0.076	**0.920±0.012**	0.763±0.066	0.781±0.041
$(m^2_{\mathrm{lo}}, \lambda_{\mathrm{lo}})$	0.029±0.044	0.042±0.031	**0.065±0.045**	0.059±0.016	0.001±0.001	0.029±0.028
$(m^2_{\mathrm{lo}}, \lambda_{\mathrm{hi}})$	0.297±0.294	0.311±0.154	0.325±0.114	0.599±0.079	**0.652±0.119**	0.517±0.111
$(m^2_{\mathrm{hi}}, \lambda_{\mathrm{lo}})$	0.330±0.248	0.437±0.155	0.412±0.188	0.650±0.117	**0.738±0.078**	0.498±0.190
$(m^2_{\mathrm{hi}}, \lambda_{\mathrm{hi}})$	0.186±0.130	0.239±0.150	0.458±0.164	**0.732±0.036**	0.329±0.066	0.280±0.124

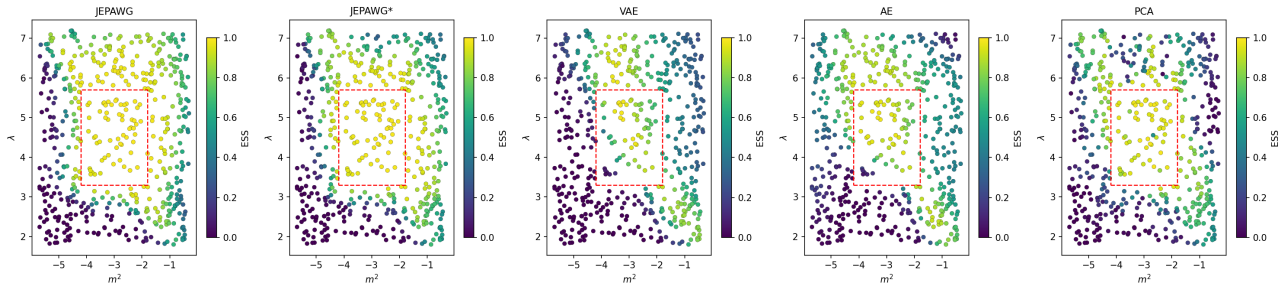

Figure 7. Per-coupling ESS scatter heatmaps for all methods in the single-seed setting, covering interpolation in the training region and extrapolation.

where t is the thermal scaling field and ν is the correlation-length critical exponent. Close to the critical line, the thermal direction is approximately aligned with the bare mass scaling field,

$$t \approx a(\lambda)\left(m^2 - m_c^2(\lambda)\right), \tag{24}$$

up to higher-order corrections and nonlinear mixing between scaling fields.

Configurations with approximately equal values of the scaling variable are expected to exhibit similar long-distance physics. Therefore, to construct FSS-related partner points between two lattice sizes L_1 and L_2, we require the scaling variable to be preserved,

$$t_1 L_1^{1/\nu} = t_2 L_2^{1/\nu}. \tag{25}$$

Substituting the leading-order relation for the thermal scaling field and solving for the source-lattice mass yields

$$m_{\text{pred}}^2 = m_c^2(\lambda; L_1) + \left(m_2^2 - m_c^2(\lambda; L_2)\right)\left(\frac{L_2}{L_1}\right)^{1/\nu}. \tag{26}$$

This construction defines the FSS-predicted source point used to test whether latent-space proximity aligns with finite-size scaling trajectories.

D.2. Critical-line estimation

The FSS matching procedure requires an estimate of the critical line $m_c^2(\lambda)$. We consider two estimators:

1. U_4 **crossing estimator:** the critical mass is obtained from Binder-cumulant crossings between neighboring lattice sizes.
2. **Per-L Binder estimator:** the critical mass is estimated independently for each lattice size from the location of the Binder-cumulant inflection point.

These two procedures differ by finite-size effects and therefore provide a simple robustness check for the latent-space FSS analysis.

D.3. Complete coarse-graining latent-distance scans

For each coarse-graining pairing, we scan the scaling exponent $\nu \in [0.2, 3.5]$ and compute the mean latent-space distance

$$d_{\text{mean}}(\nu) = \frac{1}{N}\sum_{i=1}^{N}\left\|z_i^{(1)} - z_i^{(2)}\right\|_2. \tag{27}$$

The optimal scaling exponent is defined by

$$\nu^* = \arg\min_{\nu} d_{\text{mean}}(\nu), \tag{28}$$

while

$$d_{\text{FSS}}^* = \min_{\nu} d_{\text{mean}}(\nu). \tag{29}$$

Figure 8 shows the latent-distance curves for all three coarse-graining pairings. Table 6 summarizes the corresponding numerical results. The single-step mappings ($8 \to 6$ and $10 \to 8$) show the clearest minima near $\nu = 1$; the two-step mapping ($10 \to 6$) shows a smaller improvement, consistent with larger finite-size corrections at the greater scale ratio.

Uncertainty on ν^*. To assess the stability of the minimizer ν^* with respect to the choice of coupling points, we use a nonparametric bootstrap. We draw B=500 resamples of size N with replacement from the N coupling points, recompute $d_{\text{mean}}(\nu)$ for each resample, and record its minimizer ν_b^*. The 95% confidence interval on ν^* is the $[2.5, 97.5]$ percentile interval of $\{\nu_b^*\}_{b=1}^{B}$. An interval that stays well away from the scan boundaries $\nu \in [0.2, 3.5]$ indicates that the minimum is a genuine interior feature of the curve and is not an artifact of a particular subset of coupling points. Note that the bootstrap quantifies sensitivity to the set of coupling points used; it does not capture the systematic difference between critical-line estimators, which is a separate source of uncertainty already visible from the two curves shown in each panel.

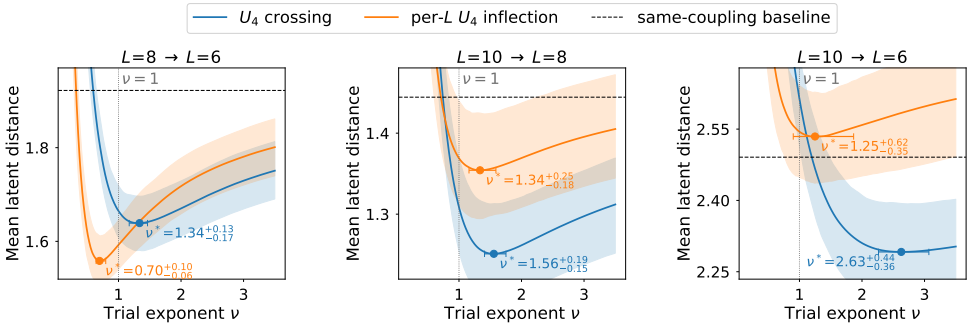

Figure 8. Latent-distance scans for all three coarse-graining mappings. Each panel shows the mean latent distance $d_{\text{mean}}(\nu)$ under FSS-based matching using two methods for estimating the critical mass (Appendix D.2), and under bare-coupling matching, as a function of the trial scaling exponent ν. Shaded bands are pointwise 95% bootstrap intervals on $d_{\text{mean}}(\nu)$. Filled circles mark ν^*; horizontal error bars show the 95% bootstrap confidence interval on ν^* (B=500 resamples with replacement over the N=361 coupling points; see Appendix D.3). Error bars that lie away from the scan boundaries confirm that the minimum is a robust feature of the curve, not driven by individual coupling points. Horizontal dashed lines indicate the same-coupling baseline d_{same}.

Table 6. Finite-size scaling pairing results for all coarse-graining lattice-size mappings ($L_{\text{src}} > L_{\text{tgt}}$). m_c^2 estimation refers to the two methods for estimating the critical line explained in Appendix D.2. d_{same} denotes the mean latent distance from same-coupling matching; d^*_{FSS} is the minimum latent distance under FSS matching at $\nu^* = \arg\min_{\nu\in[0.2,3.5]}(d_{\text{FSS}}(\nu))$. Δ denotes the relative improvement of FSS matching over same-coupling matching. 95% CI: bootstrap interval on ν^* (B=500 resamples).

$L_{\text{src}} \to L_{\text{tgt}}$	m_c^2 estimation	d_{same}	d^*_{FSS}	ν^*	Δ	95% CI
$8 \to 6$	U_4 crossing	1.922	1.639	1.34	+14.7%	[1.17, 1.46]
$8 \to 6$	per-L Binder	1.922	1.558	0.70	+18.9%	[0.64, 0.80]
$10 \to 8$	U_4 crossing	1.444	1.251	1.56	+13.4%	[1.41, 1.75]
$10 \to 8$	per-L Binder	1.444	1.354	1.34	+6.2%	[1.16, 1.59]
$10 \to 6$	U_4 crossing	2.491	2.292	2.63	+8.0%	[2.26, 3.07]
$10 \to 6$	per-L Binder	2.491	2.534	1.25	−1.7%	[0.90, 1.87]

Table 7. Per-group Pearson correlations between the per-component area element $A_c(g) = \sqrt{\det G_c(g)}$ and physical observables across the (m^2, λ) grid. Full breakdown of the four bulk groups aggregated as θ_{bulk} in Table 3. Bold entries mark $r \geq 0.5$.

	ξ	$\lvert\nabla U_4\rvert$	χ	$\lvert m\rvert$	λ
θ_{conv}	-0.12	$+0.07$	-0.06	$+0.11$	-0.75
$\theta_{\text{feat_super}}$	-0.26	-0.11	-0.24	$+0.14$	-0.72
$\theta_{\text{time_kernel}}$	-0.21	-0.05	-0.19	$+0.14$	-0.76
$\theta_{\text{feat_map}}$	-0.20	-0.03	-0.17	-0.05	-0.59
$\theta_{\text{zero_mode}}$	$\mathbf{+0.57}$	$\mathbf{+0.61}$	$\mathbf{+0.67}$	-0.04	-0.11
θ_{scale}	$+0.36$	$\mathbf{+0.56}$	$\mathbf{+0.51}$	-0.10	-0.14

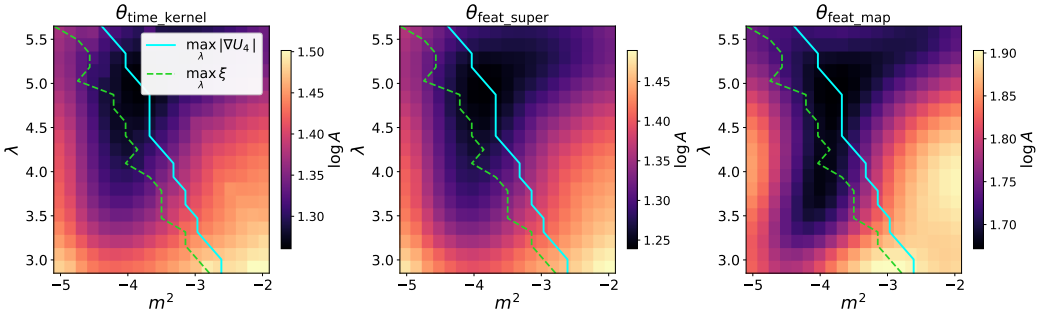

Figure 9. Per-parameter-group pullback area $A(g)$ on the (m^2, λ) grid for the 8^2 lattice. Here, the remaining parameter-groups, complementary to the groups mentioned in Figure 3. Overlaid are the per-row maxima of $|\nabla U_4|$ (cyan, solid) and ξ (lime, dashed), serving as transition proxies.

E. Phase Transition Detection

Table 7 expands the aggregated bulk row of Table 3 into per-group correlations between the area element $A_c(g)$ and the physical observables. The four bulk groups behave consistently: each shows a strong negative correlation with λ ($r \in [-0.76, -0.59]$) and no positive correlation with any criticality observable. The non-local groups $\theta_{\text{zero_mode}}$ and θ_{scale} instead correlate positively with ξ, $|\nabla U_4|$, and χ, and only weakly with λ. The complementary spatial maps for the bulk groups are shown in Figure 9.

F. Additional Data

Figure 10 visualizes the split of the coupling space (m^2, λ) into training, interpolation, and extrapolation regions. Table 8 reports decoded-flow ESS broken down across all 9 regions of the coupling plane, comparing canonicalized and raw weights for all five methods. Figure 11 extends the latent-space comparison of the main text to all five methods.

G. Shuffled Data Control Experiment

To verify that JEPAWG recovers coupling structure from the flow weights rather than exploiting architectural inductive biases, we train three shuffled-label controls. In each, the 361 fine-tuned checkpoints are randomly permuted before training so that the encoder sees a broken correspondence between couplings and weights. *Global shuffle*: a uniformly random permutation of all 361 points. *Local shuffle* k=5 / k=25: a greedy random matching in which each coupling is swapped only with one of its k nearest neighbors in normalized (m^2, λ) space, limiting how far the pairing deviates from the true layout. The permutation is fixed at the start of training and not re-drawn during training. Three independent shuffle seeds are run per condition; the model initialization seed is held fixed across them.

All shuffled models reach the same training-time $R^2 \approx 0.97$ as the clean model, confirming sufficient capacity. To evaluate whether the learned latent space encodes the true coupling, we apply each trained encoder to the *original*, unshuffled checkpoints and measure R^2 of a linear coupling predictor (Table 9). The global control collapses to near-chance ($R^2 \approx 0.03$–0.06). The local controls retain substantial signal that decreases monotonically with neighborhood radius: $R^2 \approx 0.96$ for k=5 and $R^2 \approx 0.91$ for k=25, reflecting that very local swaps preserve much of the coupling–weight structure.

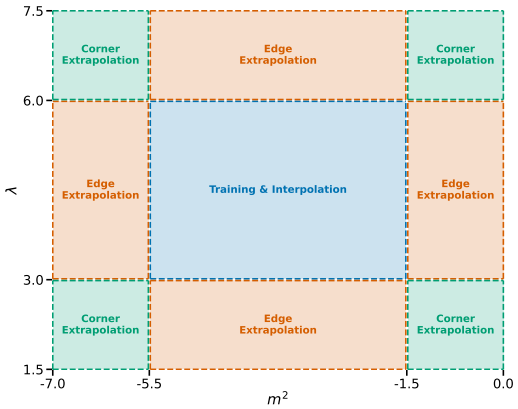

Figure 10. Split of the coupling space (m^2, λ) into training and interpolation (blue), and extrapolation regions. The extrapolation regions are further divided into edges (orange) and corners (green).

Table 8. Decoded-flow ESS on the split-seed grid across 9 regions of the coupling plane (6×6, $m^2 \in [-5.5, -1.5]$, $\lambda \in [3.0, 6.0]$). Interpolation: 256 random (m^2, λ) over the joint window; edges and corners: δ=1.5 extrapolation, 32 random points per region. Each cell is mean±std over the points in that region (single seed). JEPAWG* refers to the covariance-on ablation (γ=1); JEPAWG is the main configuration (γ=0). JEPAWG* and JEPAWG use the trained coupling predictor $P \circ E_g$; VAE, AE, and PCA use a MLP head $g \to \mathbf{z}$. VAE uses β=1, AE uses β=0. All evaluations use 4096 Metropolis Hastings samples per point.

Region	PCA	JEPAWG*	JEPAWG	VAE	AE
Can. Interp.	0.298±0.31	0.613±0.32	**0.876±0.18**	0.395±0.34	0.327±0.33
$m^2_{\rm lo}$	0.127±0.22	0.518±0.30	**0.881±0.06**	0.325±0.29	0.132±0.19
$m^2_{\rm hi}$	0.091±0.19	0.205±0.25	**0.439±0.37**	0.133±0.27	0.070±0.18
$\lambda_{\rm lo}$	0.070±0.18	0.303±0.37	**0.380±0.40**	0.248±0.35	0.188±0.29
$\lambda_{\rm hi}$	0.071±0.16	0.379±0.28	**0.532±0.34**	0.208±0.29	0.198±0.27
$(m^2_{\rm lo}, \lambda_{\rm lo})$	0.048±0.10	0.206±0.26	**0.216±0.32**	0.072±0.21	0.002±0.00
$(m^2_{\rm lo}, \lambda_{\rm hi})$	0.217±0.24	0.278±0.32	**0.696±0.15**	0.373±0.26	0.120±0.24
$(m^2_{\rm hi}, \lambda_{\rm lo})$	0.082±0.17	**0.368±0.26**	0.254±0.32	0.035±0.12	0.183±0.32
$(m^2_{\rm hi}, \lambda_{\rm hi})$	0.048±0.16	0.138±0.27	**0.405±0.34**	0.021±0.10	0.019±0.09
Raw Interp.	0.749±0.27	0.409±0.32	0.783±0.22	0.910±0.18	**0.915±0.18**
$m^2_{\rm lo}$	0.398±0.28	0.286±0.33	0.828±0.12	0.892±0.11	**0.930±0.06**
$m^2_{\rm hi}$	0.126±0.28	0.101±0.20	0.318±0.35	0.237±0.34	**0.384±0.32**
$\lambda_{\rm lo}$	0.305±0.41	0.152±0.19	0.445±0.38	**0.480±0.42**	0.477±0.41
$\lambda_{\rm hi}$	0.155±0.30	0.179±0.26	0.536±0.22	0.475±0.34	**0.668±0.29**
$(m^2_{\rm lo}, \lambda_{\rm lo})$	0.041±0.10	0.205±0.26	0.208±0.31	**0.536±0.36**	0.512±0.42
$(m^2_{\rm lo}, \lambda_{\rm hi})$	0.184±0.27	0.119±0.16	0.552±0.32	**0.691±0.26**	0.663±0.26
$(m^2_{\rm hi}, \lambda_{\rm lo})$	0.120±0.28	0.035±0.08	**0.231±0.33**	0.221±0.32	0.198±0.30
$(m^2_{\rm hi}, \lambda_{\rm hi})$	0.049±0.16	0.105±0.21	**0.380±0.36**	0.249±0.29	0.323±0.28

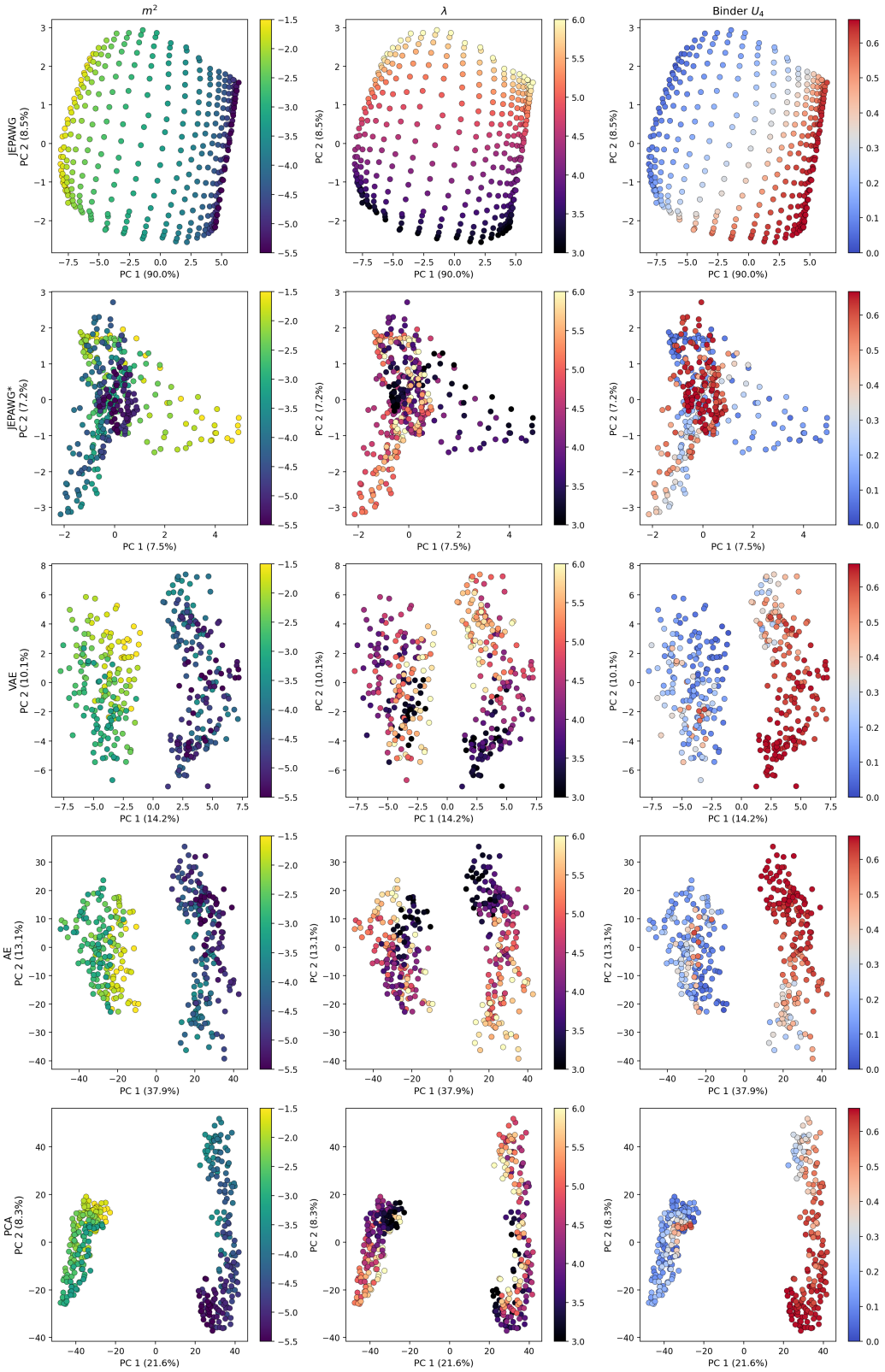

Figure 11. PCA of latent representations (d_z=16) for JEPAWG, JEPAWG*, VAE, AE, and PCA (top to bottom), trained on 361 canonicalized flow checkpoints. Points are colored by m^2, λ, and the Binder cumulant U_4. JEPAWG organizes the latent space along axes aligned with the coupling parameters and phase structure; the other methods do not.

Table 9. Shuffled-label control: R^2 of a linear coupling predictor applied to embeddings trained on shuffled checkpoints, then evaluated on the original (unshuffled) assignments. Mean±std over 3 shuffle seeds; training-time R^2 is identical (≈0.97) for all conditions.

Condition	$R^2(m^2)$	$R^2(\lambda)$
Clean (no shuffle)	0.992	0.991
Local shuffle k=5	0.960 ± 0.004	0.963 ± 0.001
Local shuffle k=25	0.909 ± 0.008	0.902 ± 0.006
Global shuffle	0.035 ± 0.009	0.055 ± 0.012

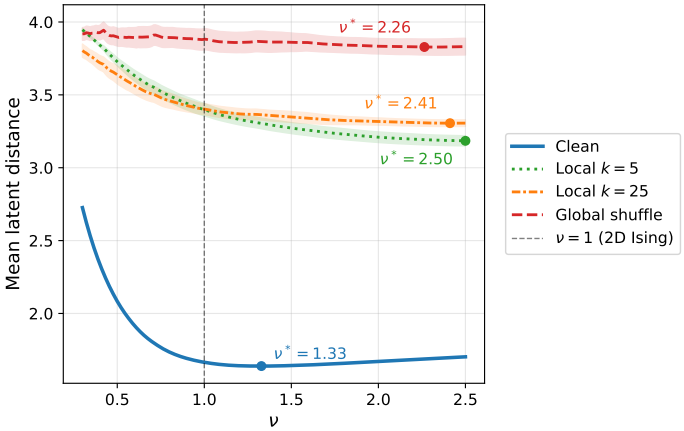

Figure 12. FSS ν scan: mean latent distance between 8×8 embeddings and their FSS-predicted 6×6 partners, as a function of the assumed scaling exponent ν. The minimum locates ν^* for each condition; shaded bands show ±std over 3 shuffle seeds. The clean JEPAWG encoder recovers $\nu^* = 1.33$, broadly in line with the analytic value of $\nu = 1$ of the 2D Ising model. All shuffled-label controls produce ν^* at or near the scan ceiling, indicating that the FSS signal requires a faithful coupling–weight correspondence.

As a second check we run the FSS ν scan (Section D; Figure 12) on each shuffled embedding. The clean JEPAWG recovers $\nu^* = 1.33$, consistent with the 2D Ising universality class. All three shuffled conditions produce ν^* at or near the scan ceiling of 2.5, indicating that the FSS signal is destroyed when the coupling–weight correspondence is broken.