Latent Energy-Score Training with Geometry-Preserving Decoders for Epistemic UQ in PDE Surrogates

Ramon Daniel Regueiro-Espino[1], Nicolas Thome[1,2], Patrick Gallinari[1,3]
[1] Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
[2] Institut universitaire de France (IUF) [3] Criteo AI Lab, Paris, France

Abstract

Modeling uncertainty in neural surrogates for partial differential equations (PDEs) is crucial for reliable scientific prediction, yet current approaches remain computationally expensive and sensitive to decoder-induced distortions. We show that probabilistic PDE surrogates can be trained entirely in latent space by minimizing the Energy Score (ES)—a strict proper scoring rule—without decoding Monte Carlo samples. Our key insight is that for approximately geometry-preserving (bi-Lipschitz) encoder-decoder pairs with small reconstruction error, improvements in latent-space ES transfer to physical space with explicit constants. We propose a two-stage framework: (i) learn a geometry-preserving latent representation, and (ii) train the latent processor using ES directly in latent space. This eliminates per-sample decoding and auxiliary uncertainty heads, reducing training cost while improving physical-space ES. We derive an explicit ES transfer bound for geometry-preserving decoders and demonstrate empirical improvements under both in-domain (ID) and out-of-domain (OOD) shifts across PDE benchmarks, with reduced wall-clock time.

1 Introduction

Recent progress in computational science has driven the rapid development of deep learning (DL) surrogate models for systems governed by partial differential equations (PDEs) Li et al. [2020], Lu et al. [2019], Brandstetter et al. [2022]. Among these, the *encode-process-decode* framework has become particularly influential, demonstrating the potential of evolving complex physical dynamics in compact latent spaces Serrano et al. [2023, 2024b], Wang and Wang [2024], Zhou et al. [2024].

Most existing frameworks, however, are designed for deterministic forecasting, which limits their ability to quantify predictive reliability. Recent developments have begun to address this gap through *latent generative processors* Serrano et al. [2024a], Koupaï et al. [2025] and auxiliary *uncertainty decoders* Wu et al. [2024]. While these approaches have improved the robustness and safety of DL-based dynamical surrogates, two important challenges remain. First, latent generative models often overlook the geometry of the decoder—the mapping from latent variables to physical space. In practice, this mapping can be *anisotropic*: it stretches some directions of the latent space while compressing others. Such distortions mean that small, isotropic uncertainty in the latent space can translate into highly skewed or miscalibrated uncertainty in the physical space, even when the latent posterior itself is well modeled. More critically for training, optimizing a latent-space ES ES_Z is *not* guaranteed to reduce the physical-space score ES_X unless the decoder is near-isometric; we empirically observe a “crossing” regime in which tightening the decoder's geometry increases ES_Z yet decreases ES_X (Table 9). This misalignment risk motivates explicit geometry control. Second, many existing probabilistic frameworks rely on additional decoding modules or multi-

1st Workshop on Epistemic Intelligence in Machine Learning (EIML 2025) .

branch architectures to estimate uncertainty, which substantially increases computational cost and complicates training.

Motivated by this latent-physical ES misalignment, our central hypothesis is that preserving the geometric structure between the latent and physical spaces is key to capturing meaningful and reliable uncertainties—without resorting to additional decoding modules. To this end, we introduce a framework that learns a *geometry-preserving latent space*, establishing a near-isometric/approximate correspondence between latent and physical distributions. This formulation simultaneously mitigates the computational overhead of multi-branch uncertainty decoders and reduces distortions induced by anisotropic mappings.

Concretely, we introduce a *geometric regularization term* into the encoder-decoder objective, encouraging distance preservation up to a global scale factor. This constraint ensures consistent uncertainty propagation across spaces. Uncertainty is then quantified using a *strict proper scoring rule (SPSR)*, which evaluates the fit of the learned probabilistic predictions to observations. We integrate geometric regularization with SPSR-based uncertainty learning in latent spaces, coupling near-isometric decoders with ES minimization.

Unlike prior approaches that estimate uncertainties directly in the high-dimensional physical space Wu et al. [2024], Bülte et al. [2025], our framework learns to forecast and assess uncertainty directly within a low-dimensional, geometry-preserving latent manifold. This results in improved scalability, reduced computational cost, and enhanced interpretability of uncertainty estimates.

Contributions.

- We establish a formal connection between geometric deep learning and latent predictive uncertainty, highlighting how geometric consistency improves uncertainty reliability.
- We introduce a latent-space training scheme for spatiotemporal forecasting that optimizes ES without decoding Monte Carlo (MC) samples during training, together with a geometry regularizer to promote near-isometric decoders.
- We demonstrate on representative spatiotemporal forecasting tasks under ID and OOD parameter shifts, the method attains competitive point accuracy and improved proper scores at lower training cost.

2 Background

Problem We study time–dependent parametric PDEs on a spatial domain Ω and time interval $[0, T]$, with solution field $u^t(x) := u(t, x)$ for $t \in [0, T]$, $x \in \Omega$. We assume for simplification that only the PDE coefficient parameters change in this parametric setting, all other aspects of the dynamics being fixed. Let $\mathcal{S}_C^{\Delta t}$ denote the one–step solution operator (flow map) that advances the state by Δt under a context C (e.g., past states $u^{0:t-1}$), i.e., $u^{t+\Delta t} = \mathcal{S}_C^{\Delta t}(u^t)$ in the deterministic case.

Let $u \sim P_U(\,\cdot\,|C)$, denote a probability distribution on the space state domain of the PDE. We seek models $\widehat{\mathcal{G}}$ that produce *probabilistic* next–step predictions.

$$\widehat{\mathcal{G}} :\ (u^t, C) \longmapsto P_U(u^{t+\Delta t}|C)$$

In the deterministic setting the ground–truth conditional law degenerates to a point mass, $\mathsf{Law}\big(u^{t+\Delta t} \mid C\big) = \delta_{\mathcal{S}_C^{\Delta t}(u^t)}$, where δ_x is the Dirac measure at x. Any stochasticity in P_U then reflects model/epistemic uncertainty (e.g., finite data, model misspecification, numerical error) rather than inherent process noise.

Encode–process–decode framework We adopt an encode–process–decode paradigm with two training phases. First, an encoder e and decoder d are trained to reconstruct states in a low-dimensional latent space. Second, a processor $\mathcal{F}$ is trained to model the *latent* time evolution, producing a conditional latent distribution that we can later decode when needed. This separation targets sample efficiency and allows us to score models directly in latent space.

Scoring-rule minimization A (proper) scoring rule S evaluates a distribution P against an outcome x drawn from a distribution Q by returning a real score $S(P, x)$; its *expected* value is minimized in truth at $Q = P$ Gneiting and Raftery [2007]. If this minimum is unique, S is strictly proper. We focus on the ES,

$$ES^{\beta}(P, x) \;=\; \mathbb{E}_{X\sim P}\big[\,\|X - x\|^{\beta}\big] \;-\; \tfrac{1}{2}\,\mathbb{E}_{X,X'\sim P}\big[\,\|X - X'\|^{\beta}\big], \qquad \beta \in (0, 2), \tag{1}$$

which is strictly proper over broad classes of distributions. Prior work Bülte et al. [2025] applies SPSR minimization with Neural Operators (NOs), directly in the physical space through:

$$\arg\min_{\theta}\; \mathbb{E}_{C\sim P_{\mathcal{C}}}\, \mathbb{E}_{u\sim P_U(\cdot|C)} S\big(P_\theta(\cdot|C),\, u\big), \tag{2}$$

where $P_{\mathcal{C}}$ is the prior distribution of the context. Equation (2) attains $P_\theta(\cdot|C) = P_U(\cdot|C)$ almost everywhere when S is strictly proper. In practice we use the empirical ES estimator with M samples $\{Z^{(m)}\}_{m=1}^{M} \sim P(\cdot)$:

$$\widehat{ES}^{\beta}_{Z} \;=\; \frac{1}{M}\sum_{m=1}^{M}\|Z^{(m)} - z\|^{\beta} - \frac{1}{2M(M-1)}\sum_{m\neq n}\|Z^{(m)} - Z^{(n)}\|^{\beta}. \tag{3}$$

3 Methodology

We show that scoring stochastic predictors with ES *in latent space* is both theoretically justified and computationally cheaper than in physical space. Our key result is a transfer bound: under a near–scaled-isometry decoder and small reconstructions, minimizing latent ES reduces physical ES up to explicit constants (Equation (4)). This yields a simple two-step training recipe.

Latent scoring We evaluate SPSR objectives *directly in latent space*. Let $e : X \to Z$ be an encoder, $g : Z \to X$ a decoder, and let the processor output a predictive distribution $P(\cdot \mid e(x))$ on Z. For each input x, we draw M latent samples $Z^{(m)} \sim P(\cdot \mid e(x))$ and train by minimizing the SPSR between these samples and the target $e(x)$ *in* Z—i.e., we optimize $ES^{\beta}_{Z}(P, e(x))$ rather than decoding with g to score against x in data space (optimize $ES^{\beta}_{X}(g(P), x)$). This keeps optimization in one geometry, avoids decoding M samples per step, and reduces variance from decoder nonlinearities. In Appendix C.2, we illustrate a “crossing” example where latent ES worsens yet physical ES improves as geometry tightens.

Transfer bounds on latent Energy Score Let g be $(\ell_{\mathrm{Lip}}, L_{\mathrm{Lip}})$-bi-Lipschitz and $r(x) = \|g(e(x)) - x\|$. Assume $\mathbb{E}\,\|Z - Z'\|^{\beta} < \infty$. The following result shows that the physical ES is bounded, upper and lower, by the latent ES with constants α_β and τ_β as in Equation (4). This will allow us, by ensuring that the bounds are tight to control ES in the physical space through ES in the latent space.

Theorem 1 (Transfer bounds on latent Energy Score)**.** *Let $e : X \to Z$ be measurable, $g : Z \to X$ be $(\ell_{\mathrm{Lip}}, L_{Lip})$-bi-Lipschitz, $\beta \in (0, 2)$, $Y = g(Z)$, and $r(x) = ||g(e(x)) - x||$. Assume $\mathbb{E}_{Z,Z'\sim P}||Z - Z'||^{\beta} < \infty$. Define*

$$\alpha_\beta = \begin{cases} 1, & \textit{if } \beta \in (0, 1], \\ 2^{\beta-1}, & \textit{if } \beta \in (1, 2), \end{cases} \qquad \tau_\beta = \begin{cases} 1, & \textit{if } \beta \in (0, 1], \\ 2^{1-\beta}, & \textit{if } \beta \in (1, 2). \end{cases}$$

Then, we have that ES in the space X, ES^{β}_{X}, can be bounded in relation with ES in the space Z, ES^{β}_{Z}, by

$$\begin{aligned} ES^{\beta}_{X}(g(P), x) &\leq \alpha_\beta L^{\beta}_{Lip} ES^{\beta}_{Z}(P, e(x)) + \tfrac{\alpha_\beta L^{\beta}_{Lip} - \ell^{\beta}_{\mathrm{Lip}}}{2}\mathbb{E}_{Z,Z'\sim P}||Z - Z'||^{\beta} + \alpha_\beta r(x)^{\beta}, \\ ES^{\beta}_{X}(g(P), x) &\geq \tau_\beta \ell^{\beta}_{\mathrm{Lip}} ES^{\beta}_{Z}(P, e(x)) - \tfrac{L^{\beta}_{Lip} - \tau_\beta \ell^{\beta}_{\mathrm{Lip}}}{2}\mathbb{E}_{Z,Z'\sim P}||Z - Z'||^{\beta} - r(x)^{\beta}. \end{aligned} \tag{4}$$

The constant associated with $\mathbb{E}_{Z,Z'\sim P}||Z - Z'||^{\beta}$ is unimprovable for $\beta \in (0, 1]$.

The proof is deferred to Appendix A. Theorem 1 also covers the Continuous Ranked Probability Score (CRPS),

$$\mathrm{CRPS}(P, x) = \mathbb{E}_{X\sim P}[\,|X - x|\,] - \tfrac{1}{2}\,\mathbb{E}_{X,X'\overset{\text{i.i.d.}}{\sim} P}[\,|X - X'|\,]\,,$$

as the special case of ES with $d = 1$ and $\beta = 1$ (Euclidean norm in one dimension).

Idealized scaled isometry If $\ell_{\text{Lip}} = L_{\text{Lip}}$ (relative distances preserved) and $r(x) = 0$ (perfect reconstruction), the bounds in Equation (4) tighten markedly. For $\beta \in (0, 1]$ they collapse to $ES_X^\beta(g(P), x) = L_{\text{Lip}}^\beta \, ES_Z^\beta(P, e(x))$; for $\beta \in (1, 2)$ an additive constant proportional to $\mathbb{E}\|Z - Z'\|^\beta$ remains because $\alpha_\beta > 1$. In other words, exact equality generally requires both near-scaled isometry and small pairwise dispersion in Z. In practice, perfect distance preservation is not reachable, and one will seek approximate preservation.

The bounds in Theorem 1 say that, up to bi-Lipschitz constants of the decoder g and the reconstruction error $r(x)$, the ES computed in physical space X is controlled by the same score computed in latent space Z. If decoder g does not distort the relative distances too much (approximate isometry preservation) and x is well reconstructed then reducing $ES_Z^\beta(P, e(x))$ necessarily reduces $ES_X^\beta(g(P), x)$ by a comparable amount. This result allows us to leverage latent ES as a *surrogate loss* for minimizing the physical ES. It can then be used as a minimization objective for directly training a latent processor. This is the basis for the algorithm proposed below. In a first step, one trains a geometry preserving auto-encoder, ensuring that the decoder is approximately a scaled isometry. In a second step, latent ES is used as a training objective for the processor.

Algorithmic framework: two steps *Step 1: Geometry-preserving autoencoding.* We fit the parametric encoder $e_\theta : \mathbb{R}^{d_{\text{phys}}} \to \mathbb{R}^{d_{\text{lat}}}$ and the parametric decoder $g_\theta : \mathbb{R}^{d_{\text{lat}}} \to \mathbb{R}^{d_{\text{phys}}}$, where d_{lat} is the dimension of the latent space and d_{phys} is the dimension of the physical space. The goal is to minimize

$$\mathcal{L}_{\text{geo-VAE}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}} \, \mathcal{L}_{\text{KL}} + \lambda_{\text{iso}} \, \mathcal{L}_{\text{iso}},$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss, $\mathcal{L}_{\text{KL}}$ the Kullback–Leibler term defined for two continuous random variables P and Q with densities on the same domain p(x) and q(x) respectively:

$$\text{D}_{\text{KL}}(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} dx,$$

and $\mathcal{L}_{\text{iso}}$ is the regularization term introduced in Lee et al. [2022], described for a probability measure P_θ as:

$$\mathcal{T}(g_\theta; P_\theta) := d_{\text{lat}}^2 \frac{\mathbb{E}_{z \sim P_\theta}[\text{Tr}[H_\theta^2(z)]}{\mathbb{E}_{z \sim P_\theta}[\text{Tr}[H_\theta(z)]^2},$$

where $H_\theta(z) := J_{g_\theta}^T(z) J_{g_\theta}(z)$, with J_{g_θ} the Jacobian of g_θ. The objective of term $\mathcal{T}(g_\theta; P_\theta)$ is to make g_θ approximately a scaled isometry so that latent distances (and hence ES) control those in physical space.

Step 2: Latent ES training. With (e_θ, g_θ) fixed, we train the processor $\mathcal{F}$ by minimizing the latent ES

$$\min_\theta \; \widehat{ES}_Z^\beta\big(P_\theta(\cdot \mid e_\theta(x)),\, e_\theta(x)\big),$$

using M samples $Z^{(m)} \sim P_\theta(\cdot \mid e_\theta(x))$ and *no decoding at train time*. By Theorem 1, improvements in latent ES transfer to physical ES. Crucially, we *do not decode* these M samples during training scaling compute to $t_{\text{lat}} \approx T_{\text{proc}}$ (latent ES) versus the original $t_{\text{phys}} \approx M T_{\text{dec}} + T_{\text{proc}}$ (physical ES).

Geometry diagnostics. We report robust surrogates of decoder geometry—lower/upper Lipschitz bands $\hat{\ell}_{\text{Lip}}, \hat{L}_{\text{Lip}}$, scale alignment α^*, STRESS, and reconstruction error $r(x)$—to indicate when latent transfers to physical should hold; full definitions and estimation details are in Appendix C.1. In practice, low anisotropy ($\hat{L}_{\text{Lip}} / \hat{\ell}_{\text{Lip}} \approx 1$), and small $r(x)$ predict strong transfer.

4 Experiments

We analyze our methods on two parametric dynamical systems. First, the 1D time-dependent Combined Equation, which is a one-dimensional PDE Brandstetter et al. [2022], without forcing terms. The Combined Equation is thus described by the following PDE:

$$[\partial_t u + \partial_x(\alpha u^2 - \beta \partial_x u + \gamma \partial_{xx} u)](t, x) = f(t, x),$$

$$f(t, x) = 0, \quad u_0(x) = \sum_{j=1}^{J} A_j \sin(2\pi \ell_j x / L + \phi_j).$$

We generate 4000/2000/2000 ID samples for training/validation/test with parameter range $\alpha \in [0, 1],\ \beta \in [0.01, 0.3],\ \gamma \in [0.0, 0.7]$, and 2000 OOD samples ($\alpha$ shifted to $[1.5, 5.0]$).

Our second system is the 2D time-dependent Burgers Equation from Zhou and Farimani [2024]:

$$[\partial_t u + u(c \cdot \nabla u) - \nu \nabla^2 u](t, x, y) = 0,$$
$$u(0, x, y) = \sum_{j=1}^{J} A_j \sin(2\pi \ell_{x_j} x/L + 2\pi \ell_{x_y} y/L + \phi_j).$$

We generate 1200/400/400 ID samples for training/validation/test with parameter range $\nu \in [0.0075, 0.015],\ c \in [0.5, 1]$, and 400 OOD samples ($\nu$ shifted to $[0.020, 0.050]$). The shift in ν produces smoother, more diffusive dynamics with fewer, shallower shocks. This lets us test whether models adapt their predictive distributions rather than merely benefiting from an easier target.

Baselines We compare against (i) a deterministic backbone trained with MSE, (ii) the Probabilistic Neural Operator (PNO) Bülte et al. [2025] trained by minimizing the ES in physical space on the same encode–process–decode backbone and (iii) Latent Evolution of PDEs with Uncertainty Quantification (LE-PDE-UQ) Wu et al. [2024] trained by minimizing an uncertainty-aware error in the physical space and a prediction error in the latent space on the same backbone but with an additional decoder for the uncertainty.

Training We use identical backbones and hyperparameters for encoder–decoder and processor across methods, and report metrics with $M = 16$ Monte-Carlo samples for ES estimation across PNO and ours. We employ a teacher-forcing strategy, utilizing linearly scheduled sampling that decays the teacher-forcing probability from 1 to 0.15 over the first 70% of epochs, then holds it constant. For both PDEs, we follow the PNO reparameterization approach Bülte et al. [2025]: two projection layers corresponding to the mean and standard deviation of a normal distribution are included to learn a pointwise normal distribution in the last layer of the network by the reparameterization trick Kingma et al. [2015].

Evaluation We use ES ($\beta = 1$) for theoretical consistency (the idealized scaled isometry collapse $ES^{\beta}_{X}(g(P), x) = L^{\beta}_{\text{Lip}}\, ES^{\beta}_{Z}(P, e(x))$ is only valid if $\beta \leq 1$). An ablation varying β shows small gains at $\beta = 1.5$ while correlation between latent and physical ES remains almost one (Table 10). Denote by $\mathcal{D}$ the spatio-temporal domain of the governing equation, by u the corresponding true distribution and u^i each of the predictions sampled from the predictive distribution $P_{\theta}^{M_{\text{meth}}}$, where M_{meth} is the number of samples for the considered method ($M_{\text{meth}} = 1$ for the deterministic case). Let $\bar{u}_{M_{\text{meth}}} := \frac{1}{M_{\text{meth}}} \sum_{m=1}^{M_{\text{meth}}} u^m$ denote the mean prediction and $\hat{q}^{\alpha}_{\theta}$ denote the empirical (pointwise) quantiles of $P_{\theta}^{M_{\text{meth}}}$ at the level α. We consider the additional following evaluation metrics to assess different aspects of the probabilistic prediction:

$$\begin{aligned}
\text{MSE}(P_{\theta}^{M}, u) :=& \frac{1}{|\mathcal{D}|} \sum_{(x,t)\in\mathcal{D}} \left(\bar{u}_M(x,t) - u(x,t)\right)^2, \\
\text{IS}_{\alpha}(P_{\theta}^{M}, u) :=& \frac{1}{|\mathcal{D}|} \sum_{(x,t)\in\mathcal{D}} \frac{1}{|\Omega|} \sum_{p\in\Omega} \Big[\big(\hat{q}_{\theta}^{1-\alpha/2}(x,t) - \hat{q}_{\theta}^{\alpha/2}(x,t)\big) \\
&+ \frac{2}{\alpha}\left(\hat{q}_{\theta}^{\alpha/2}(x,t) - u(x,t)\right)_{+} + \frac{2}{\alpha}\left(u(x,t) - \hat{q}_{\theta}^{1-\alpha/2}(x,t)\right)_{+}\Big],
\end{aligned}$$

$$\text{CRPS}\big(P_{\theta}^{M}, u\big) := \frac{1}{|\mathcal{D}|} \sum_{(x,t)\in\mathcal{D}} \left(\int_0^1 \text{QS}_{\alpha}\big(\hat{q}_{\theta}^{\alpha}(x,t),\, y\big)\, d\alpha\right),$$

$$\text{Corr}\big(ES\big) := \frac{\sum_{i=1}^{L}(ES_{Z}^{\beta,(i)} - \bar{ES}_{Z}^{\beta})(ES_{X}^{\beta,(i)} - \bar{ES}_{X}^{\beta})}{\sqrt{\sum_{i=1}^{L}(ES_{Z}^{\beta,(i)} - \bar{ES}_{Z}^{\beta})^2}\,\sqrt{\sum_{i=1}^{L}(ES_{X}^{\beta,(i)} - \bar{ES}_{X}^{\beta})^2}},$$

where $QS_{\alpha}(q_{\alpha}, y) := 2(\alpha - \mathbb{1}\{y < q_{\alpha}\})(y - q_{\alpha})$ is the quantile score. Each of the above mentioned evaluation metrics is focused on different aspects of the probabilistic prediction.

	Setup	MSE (10^{-4}) $\downarrow$	ES (10^{-2}) $\downarrow$	IS$_{0.05}$ $\downarrow$	CRPS $\downarrow$	Corr (ES) $\uparrow$
Ours	Test	34.62	**3.58**	**0.447**	**0.029**	**0.9995**
	OOD	**153.67**	**7.46**	1.325	**0.053**	**0.9941**
VAE + Backbone	Test	**32.28**	4.69	1.542	0.039	0.9842
	OOD	159.54	9.32	2.640	0.066	0.9777
PNO	Test	93.08	5.90	0.544	0.048	0.6481
	OOD	230.26	8.59	**1.027**	0.066	0.5776
LE-PDE-UQ	Test	44.36	5.49	1.819	0.045	0.4069
	OOD	169.32	9.80	2.818	0.070	0.5632

Table 1: Combined Equation results (ten rollout steps, five-step context). Metrics are reported in physical space. MSE and ES are scaled as indicated; IS$_{0.05}$ and CRPS are in natural units (lower is better). Best results are shown in **bold**.

	Setup	MSE (10^{-4}) $\downarrow$	ES (10^{-2}) $\downarrow$	IS$_{0.05}$ $\downarrow$	CRPS $\downarrow$	Corr (ES) $\uparrow$
Ours	Test	167.34	**6.92**	**1.125**	**0.069**	**0.9885**
	OOD	85.36	**6.23**	**0.640**	**0.049**	**0.9796**
VAE + Backbone	Test	**130.24**	8.16	3.265	0.082	0.9616
	OOD	87.71	6.91	2.763	0.069	0.9747
PNO	Test	1028.37	21.04	2.215	0.172	0.1848
	OOD	543.57	11.86	1.312	0.119	0.1378
LE-PDE-UQ	Test	142.51	11.16	3.439	0.086	0.4702
	OOD	**82.48**	8.75	2.689	0.067	0.5598

Table 2: Burgers Equation results (ten rollout steps, five-step context). Metrics are reported in physical space. OOD increases viscosity, yielding smoother dynamics. MSE and ES are scaled as indicated; IS$_{0.05}$ and CRPS are in natural units (lower is better). Best results are shown in **bold**.

Across both PDEs we obtain the lowest ES (ID and OOD) and lower CRPS than baselines (Tables 1, 2); IS$_{0.05}$ scores are competitive with baselines and better for Burgers. Considering OOD extrapolation, our ES/CRPS gains hold both in situations where the OOD problem is simpler such as for Burgers with high viscosity) or situations where the OOD dynamics is more complex as for our experiments with the Combined Equation. The proposed method then provides us with uncertainty quantifiers on par with or better than baselines. The high correlation between latent and physical space ES demonstrate that operating on the latent space allow us to control ES in the physical space. Additionally, as shown in Appendix C.5, randomized Probability Integral Transform (PIT) histograms on Combined Equation one-step prediction show that our model is close to a uniform law, indicating that the probabilistic forecast matches well the distribution of observations. In contrast, PNO is over-dispersed indicating a calibration of lower quality.

Compute & scalability. We summarize time-per-step versus sample count M in Appendix C.4. The observed trend matches the simple cost model above: latent-space scoring scales independently of M during training, while physical-space scoring grows roughly linearly with M.

5 Conclusion

We introduce a new theoretical link between predictive uncertainty, latent space, and geometrical deep learning. This link is validated both theoretically and practically, providing both a proof and experimental results. Based on this theoretical contribution, the presented framework offers a novel way to characterize latent uncertainty, enhancing its interpretability based on a geometry-preserving decoder. It is able to produce accurate uncertainty predictions from different, easier, and more difficult covariate shifts. Through a relative distance decoder, the UQ-aware processor training can be done in the latent space without the additional computational cost of decoding it. This framework shows potential for data-driven deep learning surrogates with better uncertainty estimation.

Acknowledgments and Disclosure of Funding

This work has benefited from state aid managed by the Agence Nationale de la Recherche (ANR) under the France 2030 programme, reference 22-PETA-0002. This work was granted access to the HPC resources of IDRIS under the allocation AD011015480 made by GENCI.

References

J. Brandstetter, D. E. Worrall, and M. Welling. Message passing neural pde solvers. *International Conference on Learning Representations*, 2022.

C. Bülte, P. Scholl, and G. Kutyniok. Probabilistic neural operators for functional uncertainty quantification. *arXiv preprint arXiv:2502.12902*, 2025.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

A. K. Koupaï, L. L. Boudec, L. Serrano, and P. Gallinari. Enma: Tokenwise autoregression for generative neural pde operators. *arXiv preprint arXiv:2506.06158*, 2025.

Y. Lee, S. Yoon, M. Son, and F. C. Park. Regularized autoencoders for isometric representation learning. In *International Conference on Learning Representations*, 2022.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

L. Lu, P. Jin, and G. E. Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

M. McCabe, B. R.-S. Blancard, L. H. Parker, R. Ohana, M. Cranmer, A. Bietti, M. Eickenberg, S. Golkar, G. Krawezik, F. Lanusse, et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

L. Serrano, L. Le Boudec, A. Kassaï Koupaï, T. X. Wang, Y. Yin, J.-N. Vittaut, and P. Gallinari. Operator learning with neural fields: Tackling pdes on general geometries. *Advances in Neural Information Processing Systems*, 36:70581–70611, 2023.

L. Serrano, A. K. Koupaï, T. X. Wang, P. Erbacher, and P. Gallinari. Zebra: In-context and generative pretraining for solving parametric pdes. *arXiv preprint arXiv:2410.03437*, 2024a.

L. Serrano, T. X. Wang, E. Le Naour, J.-N. Vittaut, and P. Gallinari. Aroma: Preserving spatial structure for latent pde modeling with local neural fields. *Advances in Neural Information Processing Systems*, 37:13489–13521, 2024b.

T. Wang and C. Wang. Latent neural operator for solving forward and inverse pde problems. *Advances in Neural Information Processing Systems*, 37:33085–33107, 2024.

T. Wu, W. Neiswanger, H. Zheng, S. Ermon, and J. Leskovec. Uncertainty quantification for forward and inverse problems of pdes via latent global evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 320–328, 2024.

A. Zhou and A. B. Farimani. Masked autoencoders are pde learners. *arXiv preprint arXiv:2403.17728*, 2024.

A. Zhou, Z. Li, M. Schneier, J. R. Buchanan Jr, and A. B. Farimani. Text2pde: Latent diffusion models for accessible physics simulation. *arXiv preprint arXiv:2410.01153*, 2024.

A Transferability of strict proper scoring rules

Definition A.1. For normed spaces $(X, ||\cdot||_X)$, and $(Y, ||\cdot||_Y)$, and constants $0 < \ell_{\text{Lip}} \leq L_{\text{Lip}} < \infty$, a mapping $T : \mathcal{M} \subseteq X \to Y$ is $(\ell_{\text{Lip}}, L_{\text{Lip}})$-bi-Lipschitz if

$$\ell_{\text{Lip}}||x - x'||_X \leq ||T(x) - T(x')||_Y \leq L_{\text{Lip}}||x - x'||_X, \quad \forall x, x' \in \mathcal{M}. \tag{5}$$

Lemma 2. *Let $A > 0$, $u, v \geq 0$, and $0 < \beta < 2$. Then*

$$\lim_{a \to 0^+} \frac{(A^2 + ua^2)^{\beta/2} - (A^2 + va^2)^{\beta/2}}{a^\beta} = 0.$$

Proof. Fix $\beta \in (0, 2)$. First, we show that $\lim_{a\to 0^+} \frac{(A^2+ua^2)^{\beta/2}-A^\beta}{a^\beta}$ and $\lim_{a\to 0^+} \frac{A^\beta-(A^2+va^2)^{\beta/2}}{a^\beta}$ exist and are equal to zero. Then, we can rewrite the lemma limit as $\lim_{a\to 0^+} \frac{(A^2+ua^2)^{\beta/2}-(A^2+va^2)^{\beta/2}}{a^\beta} = \lim_{a\to 0^+} \frac{(A^2+ua^2)^{\beta/2}-A^\beta}{a^\beta} + \frac{A^\beta-(A^2+va^2)^{\beta/2}}{a^\beta}$ and apply both limits to show the result.

We have that, by L'Hopital's Rule:

$$\lim_{a\to 0^+} \frac{(A^2 + ua^2)^{\beta/2} - A^\beta}{a^\beta} = \lim_{a\to 0^+} \frac{\frac{\beta}{2}(A^2 + ua^2)^{\beta/2-1}2ua}{\beta a^{\beta-1}} = \lim_{a\to 0^+} (A^2+ua^2)^{\beta/2-1}ua^{2-\beta} = 0,$$

$$\lim_{a\to 0^+} \frac{A^\beta - (A^2 + va^2)^{\beta/2}}{a^\beta} = \lim_{a\to 0^+} -\frac{\frac{\beta}{2}(A^2 + va^2)^{\beta/2-1}2va}{\beta a^{\beta-1}} = \lim_{a\to 0^+} -(A^2+va^2)^{\beta/2-1}va^{2-\beta} = 0.$$

Hence, both limits exist and are equal to zero. So, we have that

$$\begin{aligned}
\lim_{a\to 0^+} \frac{(A^2 + ua^2)^{\beta/2} - (A^2 + va^2)^{\beta/2}}{a^\beta} &= \lim_{a\to 0^+} \frac{(A^2 + ua^2)^{\beta/2} - A^\beta}{a^\beta} + \frac{A^\beta - (A^2 + va^2)^{\beta/2}}{a^\beta} \\
&= \lim_{a\to 0^+} \frac{(A^2 + ua^2)^{\beta/2} - A^\beta}{a^\beta} + \lim_{a\to 0^+} \frac{A^\beta - (A^2 + va^2)^{\beta/2}}{a^\beta} = 0 + 0 = 0.
\end{aligned}$$

□

Proof (Transfer bounds on latent Energy Score (ES)). Let the decoder be $(\ell_{\text{Lip}}, L_{\text{Lip}})$-bi-Lipschitz, i.e. $\ell_{\text{Lip}}||z - z'|| \leq ||g(z) - g(z')|| \leq L_{\text{Lip}}||z - z'||$ with $0 < \ell_{\text{Lip}} \leq L_{\text{Lip}} < \infty$, let $r(x) = ||g(e(x)) - x||$ be the reconstruction error, and define for a considered $\beta \in (0, 2)$:

$$\alpha_\beta = \begin{cases} 1, & \text{if } \beta \in (0, 1], \\ 2^{\beta-1}, & \text{if } \beta \in (1, 2), \end{cases} \qquad \tau_\beta = \begin{cases} 1, & \text{if } \beta \in (0, 1], \\ 2^{1-\beta}, & \text{if } \beta \in (1, 2). \end{cases}$$

We have that for $\beta \in (0, 1]$, $||g(Z) - x||^\beta \leq ||g(Z) - g(e(x))||^\beta + ||g(e(x)) - x||^\beta$, and for $\beta \in (1, 2)$, $||g(Z) - x||^\beta \leq 2^{\beta-1}||g(Z) - g(e(x))||^\beta + 2^{\beta-1}||g(e(x)) - x||^\beta$. Hence,

$$\begin{aligned}
ES_X^\beta(g(P), x) &= \mathbb{E}_{Z\sim P}||g(Z) - x||^\beta - \frac{1}{2}\mathbb{E}_{Z,Z'\sim P}||g(Z) - g(Z')||^\beta \\
&\leq \alpha_\beta \mathbb{E}_{Z\sim P}||g(Z) - g(e(x))||^\beta + \alpha_\beta||g(e(x)) - x||^\beta - \frac{1}{2}\mathbb{E}_{Z,Z'\sim P}||g(Z) - g(Z')||^\beta \\
&\leq \alpha_\beta L_{\text{Lip}}^\beta \mathbb{E}_{Z\sim P}||Z - e(x)||^\beta - \frac{\ell_{\text{Lip}}^\beta}{2}\mathbb{E}_{Z,Z'\sim P}||Z - Z'||^\beta + \alpha_\beta r(x)^\beta \\
&= \alpha_\beta L_{\text{Lip}}^\beta\, ES_Z^\beta(P, e(x)) + \frac{\alpha_\beta L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2}\mathbb{E}_{Z,Z'\sim P}||Z - Z'||^\beta + \alpha_\beta r(x)^\beta.
\end{aligned} \tag{6}$$

In the same way, we have that for $\beta \in (0,1]$, $||g(Z)-g(e(x))||^\beta \leq ||g(Z)-x||^\beta + ||g(e(x))-x||^\beta$, and for $\beta \in (1,2)$, $2^{1-\beta}||g(Z)-g(e(x))||^\beta \leq ||g(Z)-x||^\beta + ||g(e(x))-x||^\beta$. Hence,

$$\begin{aligned}
ES_X^\beta(g(P),x) &= \mathbb{E}_{Z\sim P}||g(Z)-x||^\beta - \frac{1}{2}\mathbb{E}_{Z,Z'\sim P}||g(Z)-g(Z')||^\beta \\
&\geq \tau_\beta \mathbb{E}_{Z\sim P}||g(Z)-g(e(x))||^\beta - ||g(e(x))-x||^\beta - \frac{1}{2}\mathbb{E}_{Z,Z'\sim P}||g(Z)-g(Z')||^\beta \\
&\geq \tau_\beta \ell_{\text{Lip}}^\beta \mathbb{E}_{Z\sim P}||Z-e(x)||^\beta - \frac{L_{\text{Lip}}^\beta}{2}\mathbb{E}_{Z,Z'\sim P}||Z-Z'||^\beta - r(x)^\beta \\
&= \tau_\beta \ell_{\text{Lip}}^\beta\, ES_Z^\beta(P,e(x)) - \frac{L_{\text{Lip}}^\beta - \tau_\beta \ell_{\text{Lip}}^\beta}{2}\mathbb{E}_{Z,Z'\sim P}||Z-Z'||^\beta - r(x)^\beta.
\end{aligned} \tag{7}$$

To show that the constant $\frac{\alpha_\beta L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2}$ for the upper bound and $\frac{L_{\text{Lip}}^\beta - \tau_\beta \ell_{\text{Lip}}^\beta}{2}$ for the lower bound is unimprovable in general if $\beta \in (0,1]$ we will consider a specific $(\ell_{\text{Lip}}, L_{\text{Lip}})$-bi-Lipschitz decoder and show that the bound is not true if we suppose a lower constant value.

We consider the encoder $e(x_1,x_2) = (\frac{x_1}{\ell_{\text{Lip}}}, \frac{x_2}{L_{\text{Lip}}})$ and the decoder $g(z_1,z_2) = (\ell_{\text{Lip}} z_1, L_{\text{Lip}} z_2)$. Hence, we have that g is $(\ell_{\text{Lip}}, L_{\text{Lip}})$-bi-Lipschitz. For both, upper and lower bound, we are going to show that a family of distributions $\{P_a; a > 0\}$ satisfy $\lim_{a\to 0^+} \frac{ES_X^\beta(g(P_a),g(e(x)) - \alpha_\beta L_{\text{Lip}}^\beta ES_Z^\beta(P_a,e(x))}{E_{Z,Z'\sim P}||Z-Z'||^\beta} = \frac{\alpha_\beta L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2}$ and $\lim_{a\to 0^+} \frac{\tau_\beta \ell_{\text{Lip}}^\beta ES_Z^\beta(P_a,e(x) - ES_X^\beta(g(P_a),g(e(x))}{E_{Z,Z'\sim P}||Z-Z'||^\beta} = \frac{L_{\text{Lip}}^\beta - \tau_\beta \ell_{\text{Lip}}^\beta}{2}$ respectively if $\beta \in (0,1]$.

For the upper bound, let $z = (0,3)$ and $P_a = 0.5\delta_{(-a,-3)} + 0.5\delta_{(a,-3)}$. Then, we have that for $x = (0, 3L_{\text{Lip}})$, $e(x) = (0,3)$, $g(e(x)) = x$ and $r(0,3L_{\text{Lip}}) = 0$. Moreover,

$$\begin{aligned}
\mathbb{E}_{Z\sim P}||Z-e(x)||^\beta &= \sqrt{a^2+36}^\beta \\
\mathbb{E}_{Z,Z'\sim P}||Z-Z'||^\beta &= 2^{\beta-1}a^\beta \\
\mathbb{E}_{Z\sim P}||g(Z)-g(e(x))||^\beta &= \sqrt{a^2\ell_{\text{Lip}}^2 + 36L_{\text{Lip}}^2}^\beta \\
\mathbb{E}_{Z,Z'\sim P}||g(Z)-g(Z')||^\beta &= 2^{\beta-1}a^\beta \ell_{\text{Lip}}^\beta \\
ES_Z^\beta(P_a,(0,3)) &= \sqrt{36+a^2}^\beta - 2^{\beta-2}a^\beta \\
ES_X^\beta(P_a,(0,3L_{\text{Lip}})) &= \sqrt{36L_{\text{Lip}}^2 + a^2\ell_{\text{Lip}}^2}^\beta - 2^{\beta-2}a^\beta \ell_{\text{Lip}}^\beta
\end{aligned}$$

We define $c^* = \frac{ES_X^\beta(g(P_a),(0,3L_{\text{Lip}})) - \alpha_\beta L_{\text{Lip}}^\beta ES_Z^\beta(P_a,(0,3))}{E_{Z,Z'\sim P}||Z-Z'||^\beta}$. For $\beta \in (0,1]$, and hence $\alpha_\beta = 1$, we are going to show that $\lim_{a\to 0^+} c^* = \frac{\alpha_\beta L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2}$. Below the second term vanishes by Lemma 2:

$$\begin{aligned}
\lim_{a\to 0^+} c^* &= \lim_{a\to 0^+} \frac{\sqrt{36L_{\text{Lip}}^2 + a^2\ell_{\text{Lip}}^2}^\beta - \ell_{\text{Lip}}^\beta 2^{\beta-2}a^\beta - L_{\text{Lip}}^\beta \sqrt{36+a^2}^\beta + L_{\text{Lip}}^\beta 2^{\beta-2}a^\beta}{2^{\beta-1}a^\beta} \\
&= \lim_{a\to 0^+} \frac{L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2} + \frac{\sqrt{36L_{\text{Lip}}^2 + a^2\ell_{\text{Lip}}^2}^\beta - \sqrt{36L_{\text{Lip}}^2 + a^2L_{\text{Lip}}^2}^\beta}{2^{\beta-1}a^\beta} \\
&= \lim_{a\to 0^+} \frac{L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2} + 0 = \frac{L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2}.
\end{aligned}$$

For the lower bound, let $z = (1, 0)$ and $P_a = 0.5\delta_{(0,-a)} + 0.5\delta_{(0,a)}$. Then, we have that for $x = (\ell_{\text{Lip}}, 0)$, $e(x) = (1, 0)$, $g(e(x)) = x$ and $r(\ell_{\text{Lip}}, 0) = 0$. Then,

$$\begin{aligned}
\mathbb{E}_{Z\sim P}||Z - e(x)||^\beta &= \sqrt{a^2 + 1}^\beta \\
\mathbb{E}_{Z,Z'\sim P}||Z - Z'||^\beta &= 2^{\beta-1}a^\beta \\
\mathbb{E}_{Z\sim P}||g(Z) - g(e(x))||^\beta &= \sqrt{L_{\text{Lip}}^2 a^2 + \ell_{\text{Lip}}^2}^\beta \\
\mathbb{E}_{Z,Z'\sim P}||g(Z) - g(Z')||^\beta &= 2^{\beta-1}L_{\text{Lip}}^\beta a^\beta \\
ES_Z^\beta(P_a, (1, 0)) &= \sqrt{1 + a^2}^\beta - 2^{\beta-2}a^\beta \\
ES_X^\beta(P_a, (\ell_{\text{Lip}}, 0)) &= \sqrt{\ell_{\text{Lip}}^2 + L_{\text{Lip}}^2 a^2}^\beta - 2^{\beta-2}L_{\text{Lip}}^\beta a^\beta
\end{aligned}$$

We define $\hat{c} = \frac{\tau_\beta \ell_{\text{Lip}}^\beta ES_Z^\beta(P_a,(1,0) - ES_X^\beta(g(P_a),(1,0)))}{E_{Z,Z'\sim P}||Z-Z'||^\beta}$ and we are going to show that $\lim_{a\to 0^+} \hat{c} = \frac{L_{\text{Lip}}^\beta - \tau_\beta l_{\text{Lip}}^\beta}{2}$.

Since $\beta \in (0, 1]$, we have that $\tau_\beta = 1$. Following the same idea of above, below the second term vanishes by Lemma 2:

$$\begin{aligned}
\lim_{a\to 0^+} \hat{c} &= \lim_{a\to 0^+} \frac{\ell_{\text{Lip}}^\beta \sqrt{1 + a^2}^\beta - \ell_{\text{Lip}}^\beta 2^{\beta-2}a^\beta - \sqrt{\ell_{\text{Lip}}^2 + L_{\text{Lip}}^2 a^2}^\beta + 2^{\beta-2}L_{\text{Lip}}^\beta a^\beta}{2^{\beta-1}a^\beta} \\
&= \lim_{a\to 0^+} \frac{L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2} + \frac{\sqrt{\ell_{\text{Lip}}^2 + a^2\ell_{\text{Lip}}^2}^\beta - \sqrt{\ell_{\text{Lip}}^2 + L_{\text{Lip}}^2 a^2}^\beta}{2^{\beta-1}a^\beta} \\
&= \lim_{a\to 0^+} \frac{L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2} + 0 = \frac{L_{\text{Lip}}^\beta - \ell_{\text{Lip}}^\beta}{2}.
\end{aligned}$$

□

B Implementation details

Encoder-Decoder $e - g$. based on the Encoder-Decoder from Attentive Reduced Order Model with Attention (AROMA) introduced by Serrano et al. [2024b]. This model encodes variable-size inputs onto a fixed-size compact latent token space aware of local spatial information and, in the decoding, exploits a conditional neural field. This design facilitates the model's ability to query forecast values at any point in the spatial domain of the equation. Table 3 presents the hyperparameters of the Encoder-Decoder.

Latent Processor $\mathcal{F}$. The latent processor adopts an Axial Attention Vision Transformer (AViT) architecture. It features disentangled spatial and temporal attention within its spatiotemporal blocks—specifically, axial attention for capturing spatial dependencies, drawing inspiration from the Multiple Physics Pretraining (MPP) model McCabe et al. [2023], and causal attention for modeling temporal sequences. This separation enables the processor to effectively learn complex dynamics while preserving the directionality of time. The architecture is inherently adaptable to varying input context sizes, allowing it to generalize across diverse spatiotemporal scenarios. We set the same hyperparameters of the latent processor for both equations, which are described in Table 4.

Optimization We detail the optimization hyperparameters for the encoder-decoder training in Table 5 and the optimization hyperparameters for the latent processor training in Table 6. All experiments were performed with a Tesla V100-SXM2-32GB.

Hyperparameter	Combined	Burgers
num_latents	8	16
latent_dim	4	8
normalizer	normal	normal
hidden_dim	128	128
dim	128	128
depth_inr	3	3
num_self_attentions	2	2
latent_heads	4	4
latent_dim_head	32	32
cross_heads	4	4
cross_dim_head	32	32
frequencies	[3, 4, 5]	[3, 4, 5]
num_freq	12	12
dropout_sequence	0.1	0.1
mlp_feature_dim	16	16
bottleneck_index	0	0
encode_geo	False	False
max_pos_encoding_freq	4	4
sample_posterior	True	True
include_pos_in_value	False	False
in_channels	1	1
input_dim	1	2

Table 3: Encoder–decoder hyperparameters for Combined Equation and Burgers Equation setups.

Hyperparameter	Value (shared)
hidden_dim	128
blocks	4
causal_time	True
num_heads	4

Table 4: Processor hyperparameters (identical for both equations).

C Extended results

C.1 Geometry diagnostics: definitions and estimation

To assess how well the learned latent space preserves the geometry of the data, we compute several diagnostic quantities.

Let $z_i = e_\theta(x_i)$ and $\tilde{x}_i = g_\theta(z_i)$. Define pairwise distances $D^Z_{ij} = \|z_i - z_j\|$ and $D^X_{ij} = \|\tilde{x}_i - \tilde{x}_j\|$.

First, we estimate *robust Lipschitz surrogates* $\hat{l}_{\text{Lip}}$ and $\hat{L}_{\text{Lip}}$, which capture local contraction and expansion, respectively. These are obtained as the 5th and 95th percentiles of the pairwise distance ratios

$$\frac{D^X_{ij}}{D^Z_{ij}}, \qquad i \neq j,$$

where D^X_{ij} and D^Z_{ij} denote Euclidean distances between samples i and j in the data and latent spaces, respectively.

Second, we estimate a *global scale factor* α^* that best aligns the pairwise distances between both spaces:

$$\alpha^* := \frac{\sum_{i \neq j} D^X_{ij} D^Z_{ij}}{\sum_{i \neq j} (D^Z_{ij})^2}.$$

Hyperparameter	Combined	Burgers
batch_size	32	32
epochs	3000	4000
scheduler	cosine	cosine
lr	1×10^{-3}	1×10^{-3}
$\eta_{\min}$	5×10^{-7}	5×10^{-7}
λ_{KL}	1×10^{-4}	1×10^{-4}
kl_warmup_prop	0.3	0.3
iso_warmup_epochs	300	200
grad_clip	2.0	2.0

Table 5: Encoder–decoder training hyperparameters.

Equation	batch_size	epochs	scheduler	lr	$\eta_{\min}$	grad_clip
Combined	32	200	cosine	1×10^{-4}	1×10^{-6}	0.0
Burgers	32	300	cosine	1×10^{-4}	1×10^{-6}	0.0

Table 6: Processor training hyperparameters.

Using this optimal scaling, we quantify how well the latent geometry matches the data-space geometry through the *stress* measure,

$$\text{STRESS} := \frac{\|d_X - \alpha^* d_Z\|_2}{\|d_X\|_2},$$

where d_X and d_Z are the vectors of all off-diagonal pairwise distances in data and latent spaces.

Finally, we evaluate the *reconstruction error*

$$r(x) := \|x - g(e(x))\|,$$

which measures the fidelity of the autoencoder mapping.

Together, these quantities provide complementary views of the latent geometry. A low anisotropy ratio ($\hat{L}_{\text{Lip}}/\hat{l}_{\text{Lip}} \approx 1$), a scale factor close to one ($\alpha^* \approx 1$), and a small reconstruction error $r(x)$ indicate that the decoder is approximately scaled–isometric. In this regime, improvements observed in the latent embedding space (ES) are predictive of improvements in the physical space. Tables 7 and 8 report these diagnostics, illustrating the conditions under which latent training correlates with enhanced physical-space performance.

C.2 Geometry regularization tightens latent→physical transfer

Setup. We vary the decoder's geometry weight λ_{iso} and report (i) reconstruction quality, (ii) pullback-geometry diagnostics, and (iii) distributional scores in data space. Lower values of the spread-after-scaling $\frac{\hat{L}_{\text{Lip}}-\hat{l}_{\text{Lip}}}{\alpha^*}$ and STRESS indicate a decoder closer to (single-scale) isometry.

Weight	MSE$_{\text{rec}}$ $(10^{-5}) \downarrow$	$\frac{\hat{L}_{\text{Lip}}-\hat{l}_{\text{Lip}}}{\alpha^*} \downarrow$	STRESS $(10^{-2}) \downarrow$	MSE$_{\text{pred}}$ $(10^{-4}) \downarrow$	ES$_X$ $(10^{-2}) \downarrow$	Corr(ES) $\uparrow$
0.0	1.835	0.120	3.7394	60.82	4.96	0.9867
0.01	3.069	0.094	2.8950	60.21	4.96	0.9924
0.05	2.646	0.0291	0.8964	43.66	4.15	0.9987
0.5	1.637	0.0143	0.4327	34.62	3.58	0.9995

Table 7: **Geometry vs. accuracy across geometry weights** on the Combined Equation *test* set. As λ_{iso} increases, geometry improves (smaller spread-after-scaling and STRESS) and physical-space predictive metrics (MSE$_{\text{pred}}$, ES$_X$) also improve.

In-distribution vs. OOD. From Table 8, we can conclude that, in our experiments, geometry preservation also holds under a parameter shift: diagnostics remain tight out of distribution, even as reconstruction error increases.

Setup	MSE$_{\text{rec}}$ $(10^{-5})\downarrow$	$\hat{l}_{\text{Lip}}$	$\hat{L}_{\text{Lip}}$	α^*	$\frac{\hat{L}_{\text{Lip}}-\hat{l}_{\text{Lip}}}{\alpha^*}\downarrow$	STRESS $(10^{-2})\downarrow$
Standard VAE training						
ID	1.835	3.485	3.930	3.714	0.120	3.70
OOD	76.800	3.375	3.909	3.674	0.145	4.40
Geo-preserving VAE training						
ID	1.637	2.766	2.806	2.789	0.014	0.433
OOD	60.032	2.762	2.806	2.786	0.0159	0.529

Table 8: **Geometry under distribution shift** (Combined Equation). The geometry-preserving decoder maintains a tight Lipschitz band and low STRESS both ID and OOD, despite larger reconstruction error OOD.

Crossing case (latent vs. physical ES). Table 9 highlights that *latent* ES alone can be misleading when geometry is weak. With a small increase in λ_{iso}, latent ES becomes worse but physical ES improves, consistent with tighter latent→physical transfer when the decoder is nearer isometry.

Weight	ES$_Z$ $(10^{-2})\downarrow$	ES$_X$ $(10^{-2})\downarrow$	$r = \frac{\text{ES}_X}{\text{ES}_Z}$	ΔES_Z	ΔES_X
0.0	2.3768	4.9586	2.09	–	–
0.05	2.6344	4.1599	1.58	+10.9%	-16.1%

Table 9: **Crossing example.** Increasing λ_{iso} from 0.0 to 0.05 worsens ES$_Z$ yet *improves* ES$_X$ and tightens the transfer ratio r (from 2.09 to 1.58), in line with Table 7.

Takeaways. (i) Increasing λ_{iso} consistently narrows the Lipschitz band and reduces STRESS; (ii) improvements in ES$_X$ track these geometry diagnostics; (iii) latent-only scores can *invert* unless geometry is controlled (Table 9); and (iv) geometry preservation transfers to OOD even when reconstruction quality degrades.

C.3 Impact of Energy Score β exponent choice

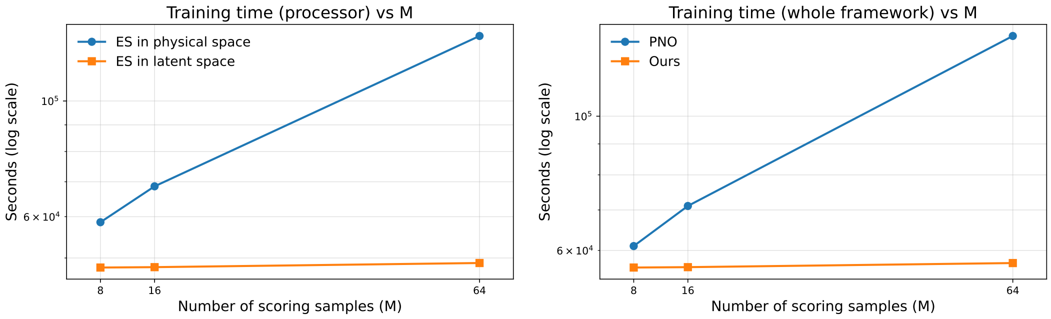
Table 10 varies the ES exponent β in the Combined Equation setting. Despite *milder* theoretical transfer guarantees for $\beta > 1$, the empirical latent–physical alignment *improves*: $\text{Corr}(\text{ES})$ increases to 0.9996 at $\beta \in \{1.5, 1.75\}$. Moreover, $\beta = 1.5$ yields the best overall physical-space performance (lowest MSE $= 34.04 \times 10^{-4}$, lowest IS$_{0.05}$ $= 0.422$, and tied-best CRPS $= 0.029$), whereas $\beta = 1.75$ preserves the correlation gain but degrades MSE/CRPS, and $\beta = 0.75$ worsens all metrics.

β	MSE $(10^{-4})\downarrow$	IS$_{0.05}\downarrow$	CRPS $\downarrow$	Corr (ES) $\uparrow$
0.5	34.92	0.454	0.030	0.9994
0.75	37.46	0.495	0.031	0.9995
1.0	34.62	0.447	**0.029**	0.9995
1.25	35.18	0.461	0.030	0.9995
1.5	**34.04**	**0.422**	**0.029**	**0.9996**
1.75	38.94	0.489	0.032	**0.9996**

Table 10: Effect of ES exponent β on Combined Equation results (ten rollout steps, five-step context). Metrics are reported in physical space. MSE is scaled as indicated; lower is better except Corr(ES). Best results are shown in **bold**.

C.4 Computational scalability for scoring samples

In Figure 1, we can see that the training time scales well with the number of samples M by computing it on the latent space. This suggests that the main computational bottleneck while training the physical ES is decoding the M samples.



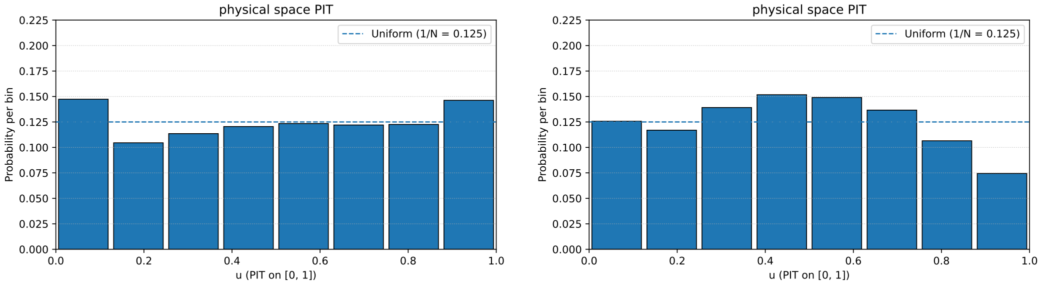
(a)

(b)

Figure 1: **Training time during scoring vs. number of scoring samples** M for (a) processor only and (b) whole framework on the Combined Equation (200 epochs).

C.5 Calibration diagnostics: PIT

We assess calibration in *physical space* using randomized PIT histograms for the Combined Equation next-step prediction. Figure 2 shows that our method's PIT is closer to uniform, whereas PNO exhibits a mild hump pattern consistent with over-dispersion. These diagnostics show that our method is better calibrated, complementing the ES/CRPS results and support that improvements obtained in latent space transfer to the data space when the decoder approximately preserves geometry.



(a) Ours

(b) PNO

Figure 2: **Physical space calibration via PIT** on the Combined Equation, next-step prediction (five-step context). Both plots use randomized PIT; the dashed line is $1/B$. Higher middle bars indicate over-dispersion; edge spikes indicate bias.