EXTRAPOLATION-AWARE PREDICTIONS WITH FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow Matching (FM) is a leading approach for probabilistic prediction in scientific and control applications. When conditioned on inputs outside the training support, however, FM models extrapolate, yielding untrustworthy predictions that cannot be distinguished from valid ones. We trace this to the Lipschitz regularity of neural vector fields, which keeps transport energy nearly constant around the data manifold, so off-manifold inputs follow regular transport paths that produce plausible predictions. We introduce *Diverging Flows*, a training objective that preserves near-optimal transport for on-manifold conditions while making transport deliberately inefficient for off-manifold ones. We show that the resulting energy margin lower-bounds trajectory curvature, so extrapolations can be detected from the sampling trajectory itself, without auxiliary models, with a threshold calibrated by conformal prediction. Across low-dimensional synthetic manifolds, scientific regression on 64×64 and 128×128 fields, and disjoint-domain generation on 32×32 images, a single model reliably detects extrapolations while retaining FM's predictive fidelity, whereas on the same data FM likelihood is near chance.

1 INTRODUCTION

Deep neural networks are increasingly outperforming traditional predictors in safety-sensitive domains, for instance replacing Model Predictive Control (MPC) in robotics (Levine et al., 2016; Miki et al., 2022; Kaufmann et al., 2023; Hwangbo et al., 2019; Akkaya et al., 2019) or physics-based modeling in weather forecasting (Lam et al., 2023; Bi et al., 2023; Kochkov et al., 2024). Their deployment in production is nonetheless hindered by *silent extrapolation*. Under anomalous or off-manifold conditions, deep models do not diverge or abstain but produce predictions that remain structured and downstream-valid, masking failure as nominal behavior. The consequences can be catastrophic. In the fatal 2018 Tempe accident, the perception stack kept relabeling a pedestrian walking a bike, and the vehicle never triggered an emergency stop (National Transportation Safety Board, 2019). In climate science, FourCastNet trained without Category 3–5 tropical cyclones cannot forecast unseen Category 5 events. It reverts to its learned distribution and predicts mild conditions as a catastrophe approaches (Sun et al., 2025; Pasche et al., 2025), rather than flagging the input as outside its competence and deferring to a physics-based model.

These failures directly follow from how these models are trained. Standard supervised and generative models are designed to interpolate within the training distribution and offer no guarantees beyond it; a model must therefore not be trusted on inputs for which it was not trained (Cao & Yousefzadeh, 2023), making extrapolation detection a prerequisite for accountable use (Lahoti et al., 2023). This requirement is increasingly codified in regulation, as the EU AI Act mandates uncertainty estimation for compliance and transparency in high-risk systems (Valdenegro-Toro & Stoykova, 2024; Herrera-Poyatos et al., 2025).

One might hope that the recent shift toward probabilistic regression would mitigate this issue. Because complex physical dynamics are inherently stochastic, deterministic predictors have been largely superseded by conditional generative models, specifically Diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020) and Flow Matching (FM) (Lipman et al., 2022; Tong et al., 2023), which can model complex, multimodal conditional distributions and now set the state of the art for these tasks. FM is particularly favored because regressing the model onto straight conditional paths between source and target distributions yields efficient, numerically stable training targets. FM and its diffusion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

counterparts are consequently displacing prior methods across robotics, weather forecasting, and structure-based drug design, among others, as discussed in Section 2.

Modeling the conditional distribution, however, does not address whether the conditioning input itself is one the model should be predicting on. The structural properties that make FM efficient also make it prone to silent extrapolation. Driven by inductive biases toward spectral smoothness (Rahaman et al., 2019), the learned neural vector fields construct smooth paths between almost any input and output. Queried with physically impossible or semantically invalid off-manifold conditions, the model neither degrades nor diverges; it transports the source distribution to a plausible target, by design. While useful in creative media synthesis (Mariani et al., 2024; Guerreiro et al., 2024; Kim et al., 2025; Ki et al., 2025), this behavior is not acceptable in safety-sensitive regression.

Standard post-hoc safety checks such as likelihood (Nalisnick et al., 2019) also fail to resolve this silent failure (Section 2). Training a separate classifier or one-class detector (Cui et al., 2023; Chalapathy & Chawla, 2019) doubles inference cost and decouples detection from the primary dynamics, while outlier exposure (Hendrycks et al., 2019) relies on a representative set of "unknowns" that is rarely available.

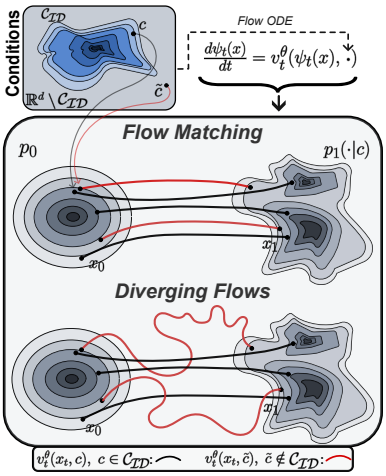


Figure 1: **Conceptual Overview.** A condition c on the valid manifold (left) and an off-manifold query $\tilde{c}$ are both passed through the flow ODE. (Top) Standard FM produces an equally smooth flow for both, silently converging to a plausible but unwarranted output. (Bottom) Diverging Flows is trained so that valid conditions still induce near-straight paths, while off-manifold queries trigger detectable, turbulent divergence from the straight reference path, which the DOT score flags at inference time.

Our starting point is a property of continuous-time conditional modeling, namely that the target probability path $p_t(x|c)$ only needs to accurately capture transport dynamics for valid conditions. Because the vector field is largely unconstrained outside the training support, it presents a unique opportunity to intrinsically encode information about the condition distribution. The same property that makes FM efficient, namely straight conditional paths, can be repurposed as a natural signal of validity. We show that an FM model can be trained to exhibit a *sharp geometric transition*, where valid conditions yield near-optimal paths while off-manifold queries yield strongly curved ones. This demonstrates that detection can emerge directly from the geometry of the generative process, letting the model signal its limits through path turbulence rather than relying on an external statistical score, closely aligning with principles of conservative distribution learning (Ma et al., 2021; Shao et al., 2024; Lubold & Taylor, 2022).

To practically realize this transition, our contributions are as follows: **(i)** We provide a theoretical characterization of silent extrapolation, showing why smooth vector fields keep transport near-optimal for conditions close to the data manifold; **(ii)** We introduce Diverging Flows, a contrastive training objective that makes transport inefficient for off-manifold inputs; **(iii)** We derive a geometric detection criterion that extracts an extrapolation score directly from trajectory turbulence during inference; and **(iv)** We extensively evaluate our framework across synthetic manifolds, physical weather forecasting, and semantic style transfer, demonstrating reliable extrapolation detection with negligible inference overhead.

2 Related Work

Generative Models for Predictive Tasks. In robotics, FM and Diffusion models have become a successful alternative for visuomotor control policies, learning to map observations to precise action trajectories (Chi et al., 2023; Black et al., 2024; Rouxel et al., 2024). In computational biology, these models have been successfully applied to protein backbone design, regressing 3D molecular structures from sequence conditioning (Jing et al., 2023; Bose et al., 2024). Similarly, in earth sciences and healthcare, FM and diffusion models are deployed for global weather forecasting (Price et al., 2025; Andrae et al., 2025; Chan et al., 2024), clinical time-series modeling (Zhang et al., 2024), as well as general time-series forecasting (Kollovieh et al., 2024; 2023). However, while these works

demonstrate the efficacy of generative models for high-fidelity regression, they typically prioritize in-distribution performance. Implicitly, they assume that the training set covers the inference distribution. This assumption often fails in applications deployed at large scale, where rare but unforeseen events are bound to occur. Yet these models lack intrinsic mechanisms to flag off-manifold queries, relying entirely on the model's ability to extrapolate, which poses severe risks in safety-critical deployments.

Off-Manifold Detection with Generative Models. While extensive literature exists for out-of-distribution (OOD) detection in discriminative models (Liang et al., 2018; Sun et al., 2022; Li et al., 2023), including methods that shape a classifier with auxiliary or synthesized boundary outliers (Lee et al., 2017; Hendrycks et al., 2019; Chen et al., 2021; Du et al., 2022) or energy scores (Liu et al., 2020) (Appendix C), addressing this challenge in conditional generative models remains underexplored. Prior unsupervised efforts primarily rely on likelihood criteria or reconstruction scores. However, likelihood-based approaches are not sufficient (Nalisnick et al., 2019; Hendrycks et al., 2019). Attempts to mitigate this via input complexity corrections (Serrà et al., 2020) or generative ensembles (Choi et al., 2018) have met with limited success, often functioning as computationally expensive post-hoc solutions rather than tackling the root cause.

Similarly, reconstruction-based methods using autoencoders or diffusion models (Liu et al., 2023b; Graham et al., 2023) flag a sample as anomalous when the model fails to reconstruct it. Applied to a conditional predictor, they score the generated *output* rather than the conditioning input itself, and typically serve as expensive external wrappers rather than integrated safety mechanisms. Lastly, DiffPath (Heng et al., 2024) detects anomalous inputs from the dynamics of the reverse generation process, but requires a second model for the predictive task. It is also designed for unconditional models, so it cannot assess the conditioning input, and relies on the curvature of diffusion SDE trajectories, which is absent in the smooth paths of FM. Very recent concurrent work (Rao et al., 2026) reads the acceleration of the sampling trajectory of a standard FM model as a training-free uncertainty proxy for failure detection in robot policies. Our findings show that it is complementary to Diverging Flows. For off-manifold conditions standard FM trajectories barely bend, so such passive proxies remain weak, whereas on Diverging Flows they detect extrapolations reliably (Appendix L.2).

3 Problem Formulation and Motivation

Let $\mathcal{C}$ denote the conditioning space and $\mathcal{M} \subset \mathcal{C}$ the manifold of valid conditions (formalized in Section 4.1).

Definition 3.1 (Extrapolation-Aware Probabilistic Regression)**.** Learn a mapping $f_\theta : \mathcal{C} \to \mathcal{P}(\mathcal{X}) \times \mathbb{R}$, $f_\theta(c) = (p_\theta(x|c),\ S_\theta(c))$, that returns a predictive distribution and an extrapolation score, and satisfies (i) *Fidelity*: for $c \in \mathcal{M}$, $p_\theta(\cdot|c)$ approximates $p_{data}(\cdot|c)$ under a chosen statistical distance; and (ii) *Trust*: $S_\theta(\tilde{c}) \gg S_\theta(c)$ for any $\tilde{c} \notin \mathcal{M}$ and $c \in \mathcal{M}$.

Validity is thus defined by manifold membership and not by density. The score should separate conditions outside $\mathcal{M}$, while rare conditions that lie on $\mathcal{M}$ should not be penalized. This mirrors the distinction between covariate and semantic shift in the distribution-shift literature (Yang et al., 2024).

Flow Matching. We realize the predictive component of our formulation using the FM framework (Lipman et al., 2022). While we frame our analysis in terms of probabilistic regression, this formulation is mathematically general and naturally extends to **conditional generation tasks**. The generative process in conditional generation is governed by the flow ODE:

$$\frac{d\psi_t(x)}{dt} = v_t^\theta(\psi_t(x), c), \tag{1}$$

where $t \in [0, 1]$, and v_t^θ is a neural vector field that pushes a noise prior p_0 to the target density $p_1(\cdot|c)$. This probabilistic formulation naturally handles aleatoric uncertainty arising from inherent data noise. By sampling N trajectories $\{\psi_1^{(i)}(c)\}_{i=1}^N \sim p_\theta(\cdot|c)$ generated by solving Eq. 1, we approximate the conditional expectation to obtain a robust prediction $\hat{x}_1 = \mathbb{E}_{p_\theta}[x_1|c]$. While this Monte Carlo integration filters stochastic noise, it fails to ensure model trustworthiness. For an off-manifold condition $\tilde{c}$ disjoint from the training support, the FM model continues to generate a coherent, low-variance predictive distribution, yielding a plausible but extrapolated prediction $\hat{x}_1$, rendering it insufficient for safety-sensitive applications.

4 DIVERGING FLOWS

4.1 THE EXTRAPOLATION PROBLEM

Let $\mathcal{X} = \mathbb{R}^d$ denote the data space and $\mathcal{C} = \mathbb{R}^k$ the conditioning space. We assume valid conditions lie on a lower-dimensional manifold $\mathcal{M} \subset \mathcal{C}$, while its complement $\mathcal{M}^c = \mathbb{R}^k \setminus \mathcal{M}$ represents the off-manifold regime. The conditional OT path of FM (Lipman et al., 2022; Liu et al., 2023a) moves each sample along the straight line from x_0 to x_1 at the constant velocity $u_t^{OT}(x_t|x_1) = x_1 - x_0$. We define the *Transport Energy Excess*, $\mathcal{E}(v^\theta; c)$, as the expected deviation of the learned vector field from this reference velocity:

$$\mathcal{E}(v^\theta; c) = \mathbb{E}_{t,x_t}\left[\|v_t^\theta(x_t, c) - u_t^{OT}(x_t|x_1)\|_2^2\right]. \tag{2}$$

Minimizing $\mathcal{E}(v^\theta; c)$ is mathematically equivalent to the standard FM objective, $\mathcal{L}_{\texttt{FM}}$ (Lipman et al., 2022). Ideally, a safe conditional predictor would keep $\mathcal{E}(v^\theta; c)$ minimal for $c \in \mathcal{M}$ (yielding high fidelity) while making $\mathcal{E}(v^\theta; \tilde{c}) \gg 0$ for $\tilde{c} \in \mathcal{M}^c$ (detectable divergence).

However, achieving this separation is hindered by the structural regularity required of neural ODEs. For the solution ψ_t to be unique, the learned vector field must be locally Lipschitz continuous (Coddington & Levinson, 1955; Lipman et al., 2024). This constraint creates an *extrapolation hazard*, formalized in Proposition 4.1.

Proposition 4.1 (Lipschitz Continuity of Transport Energy)**.** *Let $v_t^\theta(x, c)$ be a continuous neural vector field, uniformly bounded by M, that is locally Lipschitz continuous with respect to the conditioning variable c with constant K, and let $\|u_t^{OT}\|_2 \leq U$. Then the Transport Energy Excess $\mathcal{E}(c) \coloneqq \mathcal{E}(v^\theta; c)$ of Eq. 2 is locally Lipschitz continuous with constant $L \leq 2K(M + U)$ (Proof in Appendix D.1).*

Spectral bias favors vector fields with a small Lipschitz constant K (Rahaman et al., 2019), and the FM loss does not penalize a flat energy landscape. By Proposition 4.1, a small K then keeps the energy of any proximate off-manifold condition $\tilde{c} \in \mathcal{M}^c$ close to its minimal value on the manifold, since $|\mathcal{E}(v^\theta; \tilde{c}) - \mathcal{E}(v^\theta; c)| \leq L\|\tilde{c} - c\|_2$. The FM baselines in Section 5 confirm this empirically. This is the structural root of the extrapolation problem. Without active geometric regularization, the model defaults to the most efficient path regardless of validity, producing silent extrapolations that mimic valid flows.

4.2 ACTIVE GEOMETRIC DIVERGENCE VIA VECTOR FIELD MARGINS

To counteract the implicit smoothness bias described in Section 4.1, we must enforce inefficient transport for off-manifold inputs. We achieve this by projecting contrastive boundaries directly into the dynamic space of vector fields.

We treat the OT reference velocity u_t^{OT} as a *geometric anchor*. For a valid data-condition pair (x_1, c), standard FM drives the predicted velocity $v_t^\theta(x_t, c)$ toward this anchor. However, for an off-manifold condition $\tilde{c} \in \mathcal{M}^c$, we apply a targeted boundary objective that induces a locally larger Lipschitz constant around $\mathcal{M}$.

We decompose this geometric deviation into two orthogonal components. First, we define a kinetic energy margin ($\mathcal{L}_{\texttt{repel}}$) that bounds the magnitude of the off-manifold velocity:

$$\mathcal{L}_{\texttt{repel}} = \Phi\Big(\|u_t^{OT} - v_t^\theta(x_t, c)\|_2 - \|u_t^{OT} - v_t^\theta(x_t, \tilde{c})\|_2 + m_r\Big) \tag{3}$$

This term drives the off-manifold velocity to deviate in magnitude from the optimal transport speed. By breaking the constant-velocity property, we ensure the transport becomes energetically inefficient.

Second, we define a geodesic alignment margin ($\mathcal{L}_{\texttt{curve}}$) to penalize the directional similarity of the flows:

$$\mathcal{L}_{\texttt{curve}} = \Phi\Big(d_{\cos}(u_t^{OT}, v_t^\theta(x_t, c)) - d_{\cos}(u_t^{OT}, v_t^\theta(x_t, \tilde{c})) + m_c\Big) \tag{4}$$

where $d_{\cos}(a, b) = 1 - \frac{a \cdot b}{\|a\|\|b\|}$ measures the angular divergence. This penalty actively steers the vector field away from the reference direction, so that a field that keeps the correct speed still deviates

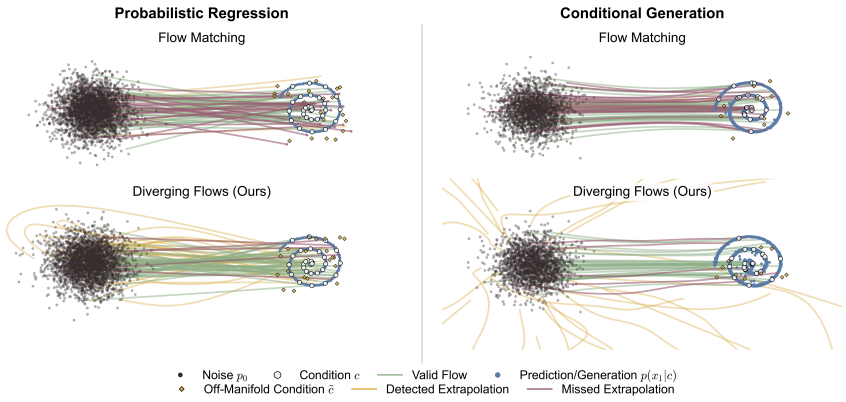


Figure 2: **Vector Field Dynamics on Synthetic Manifolds.** Probabilistic Regression (left, 10-step horizon) and Conditional Generation (right). (Top) Standard FM yields smooth convergence for all inputs, producing silent extrapolations. (Bottom) Diverging Flows: valid inputs (green) follow near-straight paths, while off-manifold queries (orange) follow detectable divergent flows.

in direction. $\mathcal{L}_{\texttt{curve}}$ is not redundant with $\mathcal{L}_{\texttt{repel}}$; penalizing speed alone admits a degenerate scalar-rescaling solution $v_t^\theta(x_t, \tilde{c}) = \gamma\, u_t^{OT}$ that satisfies the margin without changing direction, an edge case our energy-based formulation alone does not rule out (Remark D.3).

The two hyperparameters m_r and m_c establish energetic and directional constraints around the valid data manifold. Because m_r bounds a norm in the ambient data space, its scale must grow with dimensionality; Appendix D.4 gives a concentration-of-measure argument for $m_r \propto \sqrt{d}\,\sigma_v$, consistent with the observed behavior in our experiments. The function $\Phi : \mathbb{R} \to \mathbb{R}^+$ is a monotonically increasing penalty, which we instantiate as a hinge projection $\Phi(z) = \max(0, z)$. This hard-margin formulation is essential for optimization stability; once the off-manifold velocity deviates sufficiently to satisfy the margin, the contrastive gradient vanishes. This bounded penalty promotes geometric divergence without triggering unbounded vector field explosion, thereby preserving the structural integrity of on-manifold flows. The final Diverging Flows (DiFlo) objective integrates the baseline regression with these dynamic regularizers:

$$\mathcal{L}_{\texttt{DiFlo}} = \mathbb{E}_{t,x_t}\left[\|v_t^\theta(x_t, c) - u_t^{OT}\|_2^2\right] + \lambda\mathcal{L}_{\texttt{repel}} + \beta\mathcal{L}_{\texttt{curve}} \tag{5}$$

Theorem 4.2 relates the energy excess along a realized trajectory to its curvature.

Theorem 4.2 (Energy–Curvature Bound)**.** *Let $\psi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ be a flow map governed by the conditional velocity field $\frac{d\psi_t(x)}{dt} = v_t^\theta(\psi_t(x), \tilde{c})$, mapping an initial state $x_0 \sim p_0$ to a realized prediction $\hat{x}_1 = \psi_1(x_0)$ under a condition $\tilde{c}$. Let $u^{local} = \hat{x}_1 - x_0$ be the constant-velocity path connecting these specific endpoints. If the energy excess of the trajectory relative to this path is at least $\Delta_{local} > 0$, i.e., $\int_0^1 \|v_t^\theta(\psi_t(x_0), \tilde{c}) - u^{local}\|_2^2 dt \geq \Delta_{local}$, then its integrated squared acceleration $\kappa = \int_0^1 \|\frac{d}{dt} v_t^\theta(\psi_t(x_0), \tilde{c})\|_2^2 dt$ satisfies $\kappa \geq \pi^2 \Delta_{local}$ (Proof in Appendix D.3).*

The bound holds for any trajectory, so a large energy excess can only arise through curvature. The margins of Eq. 5 create this excess at boundary samples during training, and our experiments confirm that it carries over to inference, where the median Δ_{local} of off-manifold conditions is on average $4.2\times$ that of valid ones under Diverging Flows, against $1.08\times$ under standard FM (Appendix D.6). Valid conditions thus keep a small energy excess, while off-manifold ones follow inefficient, curved trajectories, as illustrated in Figure 1. The training algorithm is presented in Appendix I.

4.3 Shaping the Manifold Boundary

Diverging Flows does not require a dataset of negative examples. The contrastive margins only need to be enforced at conditions lying just outside the support of valid conditions, which we refer to as *boundary samples* and generate by perturbing the training conditions. The objective therefore raises

a barrier of high transport energy around $\mathcal{M}$, and by Proposition 4.1 the margin enforced at each boundary sample extends to its neighborhood.

As a consequence, conditions close to the boundary are the most difficult to detect. The FM objective supervises the transport energy only on $\mathcal{M}$, and the energy elsewhere is determined by how the network generalizes from it (Proposition 4.1). The barrier interrupts this generalization at the boundary, and recovering low energy beyond it would require the vector field to vary even more sharply, which the smoothness bias of neural networks disfavors (Rahaman et al., 2019). In line with this, a model trained exclusively with boundary samples perfectly separates far off-manifold conditions, such as noise or unrelated image datasets (Appendix F), whereas training on a fixed set of outliers performs considerably worse (Appendix J.1).

The choice of sampling strategy is flexible. Semantic perturbations can be used when domain knowledge is available, and feature-space Mixup or VOS (Du et al., 2022) are further options. In this work, we use Projected Gradient Descent (PGD) (Madry et al., 2018), which perturbs a valid condition within an ϵ-ball in the direction of steepest increase of the transport energy, using $\mathcal{K}$ steps of size $\eta = \epsilon/\mathcal{K}$,

$$\tilde{c}_{k+1} = \Pi_{\mathcal{B}^{\infty}_{\tilde{\epsilon}}(c)} \left(\tilde{c}_k + \eta \cdot \operatorname{sgn}\big(\nabla_{\tilde{c}} \mathcal{L}_{\text{FM}}(\tilde{c}_k)\big) \right). \tag{6}$$

An emergent property of this design is that PGD and the contrastive margins complement each other. Under standard FM, PGD produces weak boundary samples because the energy landscape around $\mathcal{M}$ remains flat, as predicted by Proposition 4.1. The margins of Diverging Flows steepen this landscape, which guides PGD toward the boundary and, in turn, sharpens the barrier (Appendix E).

4.4 Detecting Extrapolations via Trajectory Divergence

Each training path of FM is straight, but many paths pass through the same point x_t, so the network learns their average direction, toward $\mathbb{E}[x_1 \mid x_t]$. Early in integration this average covers many endpoints, and near $t = 1$ only one remains. The direction therefore changes along the way, and even valid trajectories bend slightly.

We define a constant-velocity geometric anchor to serve as a fixed reference metric. For a generated sequence of states $\{\hat{x}_{t_i}\}_{i=1}^{N}$ solved over N discrete time steps from x_0 to the final prediction $\hat{x}_1$, this linear interpolant is given by $x^{\text{ref}}_{t_i} = (1 - t_i)x_0 + t_i\hat{x}_1$. We propose the *Divergence from Optimal Trajectory* (DOT) score, a cheap state-space proxy for the deviation field $\tilde{\epsilon}_t = v^{\theta}_t(\psi_t(x_0), \tilde{c}) - u^{\text{local}}$ of Theorem 4.2. Computing κ or Δ_{local} requires the velocities along the trajectory; instead, since $x^{\text{ref}}_{t_i} = x_0 + t_i u^{\text{local}}$, the state-space gap at each step is $\hat{x}_{t_i} - x^{\text{ref}}_{t_i} = \int_0^{t_i} \tilde{\epsilon}_s \, ds$. We take its L_1 norm, averaged over dimensions and summed over the whole trajectory, as the practical score:

$$S_{DOT}(\hat{x}_{0:1}) = \sum_{i=1}^{N} \left(\frac{1}{D} \sum_{d=1}^{D} |\hat{x}^{(d)}_{t_i} - x^{\text{ref}^{(d)}}_{t_i}| \right) \tag{7}$$

where D is the spatial dimensionality (L_1 rather than κ's L_2, for cost and localized sensitivity; Appendix D.5 gives the full derivation). DOT accumulates $\tilde{\epsilon}_t$ over time instead of measuring it directly, so it is not bounded by Δ_{local}, but Appendix D.6 shows the two are closely correlated (Spearman $\rho = 0.97$). While on-manifold flows naturally exhibit minor curvature ($S_{DOT} > 0$), extrapolation flows exhibit a pronounced geometric divergence, resulting in significantly larger DOT scores ($S_{\text{DOT}}(c) \ll S_{\text{DOT}}(\tilde{c})$).

Since valid scores are not exactly zero, we calibrate a decision interval $[q_{lo}, q_{hi}]$ via Split Conformal Prediction (Angelopoulos et al., 2024) on a calibration set of valid conditions rather than fixing a heuristic threshold. This framework provides a distribution-free, finite-sample guarantee that the false rejection rate for valid inputs is at most a user-specified error rate α under exchangeability. The lower bound rejects unusually straight trajectories, which an FM model can produce, but was redundant in our experiments (Appendix G). Because the DOT metric relies solely on lightweight L_1 distance computations against a geometric anchor, it introduces near-zero computational overhead ($\sim 0.8\%$) to the standard generative inference pass. Further details on the conformal calibration and inference algorithms are provided in Appendices G and I. When an extrapolation is flagged, a fallback mechanism with hard guarantees should intervene (e.g., a physics-based simulator) rather than DiFlo prescribing the downstream action itself.

5 EXPERIMENTS

We evaluate Diverging Flows on (i) synthetic manifolds, (ii) weather temperature forecasting, and (iii) cross-domain style transfer. We construct the invalid conditions to obtain ground-truth labels, covering physical, semantic, noise, and dataset shifts (Appendix F). Our scope is native detection, i.e., a single predictive model that flags its own extrapolations. To the best of our knowledge, no existing predictor has this property, so we benchmark against standard FM under two scores, namely our DOT score, verifying that detection stems from the contrastive reshaping of the vector field rather than the metric alone, and the model likelihood, estimated with the instantaneous change of variables formula and Hutchinson's trace estimator (Lipman et al., 2024). For regression tasks only, we also compare against a three-member FM Ensemble scored by its variance; in conditional generation, a single valid condition maps to diverse outputs, making ensemble variance unreliable. As a reference outside this scope, DiffPath (Heng et al., 2024), a post-hoc detector that requires a second diffusion model, reaches comparable detection (Appendix L). Input-space detectors such as Mahalanobis (Lee et al., 2018) or deep nearest neighbors (Sun et al., 2022) likewise rely on an additional model of the conditions. They are complementary to Diverging Flows, which uses no model besides the predictor.

We quantify *Safety* via two metrics: (1) the AUROC (Fawcett, 2006), the probability that an off-manifold condition scores higher than a valid one (0.5 is chance); and (2) the FPR governed by split conformal prediction constraints. We calibrate a decision region C_α to guarantee a marginal coverage of $1-\alpha$ for on-manifold inputs. The FPR reports the percentage of off-manifold anomalies $\tilde{c}$ that incorrectly satisfy this condition, defined formally as $\text{FPR} = \mathbb{P}(S(\tilde{c}) \in C_\alpha \mid \tilde{c} \notin \mathcal{M})$. For *Fidelity*, we ensure our contrastive objective does not degrade predictive performance using task-specific metrics. In regression, we monitor physical consistency via Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR) (Hore & Ziou, 2010), and Structural Similarity (SSIM) (Wang et al., 2003; 2004). In generative tasks, we assess distribution quality via the Fréchet Inception Distance (FID) (Heusel et al., 2017) and Perceptual Similarity (LPIPS) (Zhang et al., 2018b). Detailed hyperparameters and architecture specifications are provided in Appendix M.

5.1 SYNTHETIC MANIFOLD EXPERIMENTS

Let $\mathcal{M}^2$ be a spiral manifold parameterized by $\theta \sim \mathcal{U}[0, 5\pi]$. A valid sample is given by $x(\theta) = r(\theta)(\cos\theta \ \ \sin\theta)^\intercal + \xi$, where $r(\theta) = \frac{\theta}{5\pi}$ is the normalized radius and $\xi \sim \mathcal{N}(0, \sigma^2 I)$ represents intrinsic noise ($\sigma_{\text{gen}} = 5 \cdot 10^{-3}$ and $\sigma_{\text{reg}} = 2 \cdot 10^{-3}$). The off-manifold set $\mathcal{M}_\epsilon^c = \{c \in [-1,1]^2 \mid \min_{c' \in \mathcal{M}} \|c - c'\|_2 > \epsilon\}$ is the ambient space separated from the manifold support by a margin ϵ.

We evaluate two conditional tasks: (i) **Probabilistic Regression:** Given a current state $c = x(\theta_k)$, the model predicts the future trajectory $y = [x(\theta_{k+1}), \ldots, x(\theta_{k+10})]$. This tests the preservation of local dynamics alongside the ability to reject off-manifold inputs. (ii) **Conditional Generation:** The model learns $p_\theta(x|c) \approx p_1(x)$ for any valid condition $c \in \mathcal{M}$, subject to $x \neq c$. Here, c acts as a validity token; off-manifold conditions must trigger vector field divergence rather than sampling valid data. For the contrastive training, we employ PGD to generate boundary samples. The results are illustrated in Figure 2, where we show how the flow ODE of FM and DiFlo behaves when an off-manifold input is encountered.

Table 1: **Results on Synthetic Manifolds**. FM fails to detect off-manifold inputs effectively, while Diverging Flows achieves consistently high separation with low FPR after the conformal calibration. We report the mean and standard deviation across three random seeds.

	Algorithm	AUROC ↑	FPR (%) ↓	MSE (10^{-5}) ↓
Gen.	FM-Likelihood	0.493 ±0.004	97.0 ±0.4	–
	FM-DOT	0.495 ±0.006	98.0 ±0.2	–
	DiFlo (Ours)	**0.981 ±0.004**	**4.6 ±1.0**	–
Regr.	FM-Likelihood	0.566 ±0.02	98.9 ±0.2	8.43 ±0.09
	FM-DOT	0.594 ±0.02	94.5 ±0.7	8.43 ±0.09
	FM Ensemble	0.688 ±0.01	88.8 ±2.0	**8.40 ±0.08**
	DiFlo (Ours)	**0.998 ±0.00**	**1.00 ±0.0**	8.48 ±0.10

Results on synthetic manifolds (Table 1) empirically confirm the smoothness hazard inherent to FM. The baseline yields AUROC scores near 0.50 for both tasks, effectively reducing off-manifold detection to random guessing. While the ensemble performs slightly better, it remains an unreliable detector. This indicates that without explicit constraints, the model extrapolates smooth paths even for off-manifold inputs. In contrast, Diverging Flows induces a sharp geometric transition, achieving consistently high separation while maintaining predictive fidelity comparable to the baseline.

Detection landscapes over the full ambient space, an ablation of each loss component and of the boundary sampling, and a sensitivity analysis are given in Appendices K and J.2.

5.2 Weather Temperature Forecasting

We evaluate the deployment viability of Diverging Flows by forecasting surface temperature heatmaps (64×64) six hours into the future, on the ERA5 dataset (Hersbach et al., 2020). The model $p_\theta(x_{t+6h}|x_t)$ is realized via the learned vector field $v_t^\theta(x_t, c)$, with the current atmospheric state as the conditioning input $c = x_t$; full architecture details are provided in Appendix M.4. For this task, we utilize PGD for generating boundary samples.

This domain demands strict adherence to physical laws; a reliable model must reject inputs that violate thermodynamic principles. To test the physical consistency, we adopt the hotspot protocol introduced by Chan et al. (2024), which injects thermal anomalies into historically cold regions and here alters only 0.1% of the input pixels. These perturbations are semantically plausible as images but violate local thermodynamic laws. A conservative forecaster must identify these inputs as physically infeasible rather than hallucinating. We compare our method against standard FM and an FM Ensemble.

Table 2: **Weather Temperature Forecasting Performance (ERA5)**. DiFlo reliably detects physical anomalies while maintaining competitive predictive fidelity on valid forecasts. We report the mean and standard deviation across three random seeds. Bold marks the best result.

Algorithm	AUROC ↑	MSE ↓	SSIM ↑	PSNR ↑
FM-Likelihood	0.512 ±0.02	$(8.1 \pm 1.1) \cdot 10^{-4}$	0.970 ±0.001	36.97 ±0.61
FM-DOT	0.667 ±0.02	$(8.1 \pm 1.1) \cdot 10^{-4}$	0.970 ±0.001	36.97 ±0.61
FM Ensemble	0.863 ±0.08	$\mathbf{(6.7 \pm 0.00) \cdot 10^{-4}}$	**0.975 ±0.001**	**37.73 ±0.65**
DiFlo (Ours)	**0.995 ±0.005**	$(8.22 \pm 0.3) \cdot 10^{-4}$	0.967 ±0.002	36.36 ±0.21

The results on ERA5 (Table 2) show that standard models fail to capture physical constraints. The FM baseline yields a near-random (likelihood) or weak (DOT) AUROC because it ignores the violation, treating the hotspot as valid high-frequency noise, as illustrated in Figure 3. The ensemble exhibits high performance across all predictive metrics and sufficient detection capability; however, this requires a threefold training and inference cost.

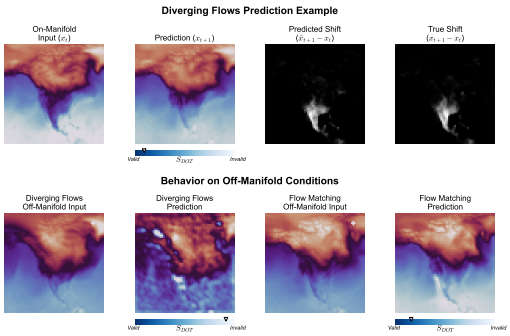


Figure 3: **High-Fidelity Forecasts with Diverging Flows.** (Top) Diverging Flows preserves the physical dynamics for valid inputs. (Bottom) Off-manifold thermal anomalies induce trajectory divergence, breaking the flow, while standard FM produces plausible but unwarranted forecasts.

In contrast, Diverging Flows achieves high detection performance. By altering the transport for off-manifold inputs, the network learns to separate physically inconsistent inputs. We observe that Diverging Flows causes negligible degradation in the predictive performance compared to the standard FM. This confirms that geometric regularization effectively filters unphysical dynamics without hindering the learning of valid temperature patterns. A classification head trained jointly on the same boundary samples reaches an AUROC of only 0.53 (Appendix F.1), so the gain stems from the flow geometry and not merely from having boundary samples. The learned boundary also generalizes to far off-manifold conditions such as noise and unrelated image datasets (Appendix F). Lastly, conformal prediction ($\alpha = 0.05$) yields an FPR of 7% on average (standard FM: 91%) while rejecting at most 5% of valid conditions (Appendix H.1). We also observe that the deviation often appears early in integration, which opens the door to flagging extrapolations before the prediction is complete (Appendix H.2). A sensitivity analysis (Appendix J.3) identifies the hyperparameters that matter in high dimension and gives simple rules for setting them, with fidelity unaffected across all configurations. Detection remains high (AUROC 0.974) at $4\times$ the pixel count (128×128, Appendix J.4).

5.3 Cross-Domain Style Transfer

Lastly, we investigate the model's ability to handle structurally disjoint conditioning and training manifolds by mapping grayscale MNIST digits (c) (LeCun & Cortes, 1998) to RGB Street View House Numbers (SVHN) (x) (Netzer et al., 2011). We introduce Fashion-MNIST (FMNIST) (Xiao et al., 2017) and Kuzushiji-MNIST (KMNIST) (Clanuwat et al., 2018) to probe whether Diverging Flows implicitly captures the semantic support of valid conditions. FMNIST represents a clear *structural shift* (clothing vs. digits), whereas KMNIST poses a harder *semantic shift*, sharing significant low-level statistics (e.g., stroke density) with MNIST. To evaluate the flexibility in boundary sampling for this visual domain, we compare both PGD and semantic transformations. We describe the transformations we used along with an ablation study on additional mining techniques in Appendix J.1.

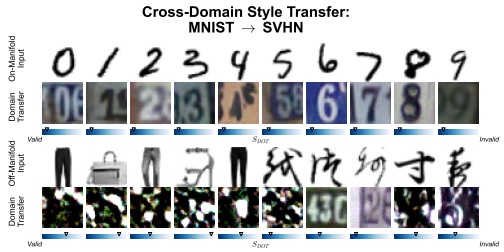


Figure 4: **Generative Quality in Cross-Domain Transfer.** Samples mapping MNIST → SVHN. Diverging Flows preserves the semantic identity of the digit.

Table 3: **Cross-Domain Style Transfer Results**. DiFlo learns a semantic boundary, rejecting off-manifold queries without degrading generative quality (FID) compared to the baseline.

Algorithm	AUROC ↑ (FMNIST/KMNIST)	FID-50K ↓	LPIPS
FM-Likelihood	0.529 / 0.516	**4.102**	0.2172
FM-DOT	0.527 / 0.513	**4.102**	0.2172
DiFlo PGD	**0.955 / 0.860**	4.104	0.2202
DiFlo Transformations	**0.987 / 0.892**	4.109	0.2213

The style transfer results (Table 3) highlight that FM ignores semantic boundaries, since it extrapolates confident paths for KMNIST (near-random AUROC). In contrast, Diverging Flows imposes a clear boundary via PGD or Transformations, learning to reject non-digit concepts. This awareness comes with negligible cost to generation quality; as shown in Figure 4, the model retains expressivity for valid inputs, confirming that geometric regularization targets off-manifold semantics without constraining the valid generative distribution. KMNIST is a near-boundary shift that shares stroke statistics with MNIST, and it is the hardest case for all methods. Qualitatively, the KMNIST characters mapped to plausible outputs are digit-like, while dissimilar ones are flagged. PGD also trails semantic transformations here, since a small pixel-space perturbation rarely makes a digit semantically invalid, whereas in ERA5 physical laws directly define which inputs are valid.

6 Conclusion and Limitations

Motivated by the widespread use of flow matching for prediction, we showed that standard FM models extrapolate silently, which prevents their use where a trustworthy predictor is required. We traced this to the smoothness of the learned vector field (Proposition 4.1) and showed, with Diverging Flows, that the flow can encode the missing information. Keeping transport near-optimal for valid conditions and inefficient for off-manifold ones lets a single FM model both predict and detect its own extrapolations, with negligible inference overhead.

The main limitation is the training cost of PGD, our fully domain-agnostic boundary sampler (2.3–$5\times$ a standard FM step, Appendix B), paid once; at inference, the DOT score adds only 0.8%. PGD is the only interchangeable component of the framework, so faster samplers can replace it without changing the formulation (Appendix B). The margins only involve velocities, so they extend directly to Riemannian FM (Chen & Lipman, 2024).

More broadly, this work paves the way for deploying probabilistic predictors where the training data cannot cover the deployment distribution. In robotics, covering every real-world situation is unrealistic (Sünderhauf et al., 2018; O'Neill et al., 2024), so a policy that flags its own extrapolations can hand control to a safe fallback instead of failing silently. The same applies to forecasting under unprecedented extremes (Sun et al., 2025) and to medical and molecular prediction.

AI USE STATEMENT

In preparing this manuscript, LLMs were only used as writing assistants to polish the wording, correct grammar, and revise sentences of the text for clarity and consistency. All suggestions were reviewed and edited by the authors, who take full responsibility for the content of the paper.

ETHICS STATEMENT

This work does not involve human subjects, and we do not release any new datasets. We use only existing, publicly available benchmarks (ERA5, MNIST, SVHN, Fashion-MNIST, Kuzushiji-MNIST) in accordance with their intended use. Our method is intended to improve the safety and reliability of generative models deployed in high-stakes prediction settings by exposing silent extrapolation failures rather than concealing them; we discuss the broader societal motivation and impact of this goal in Appendix A. We do not foresee ethical concerns specific to this work beyond the general considerations that apply to conditional generative modeling research.

REPRODUCIBILITY STATEMENT

Complete proofs of Proposition 4.1 and Theorem 4.2, along with the additional derivations connecting the DOT score to the curvature bound and the margin-scaling rule, are provided in Appendix D. Pseudocode for training and inference is given in Appendix I. Full architecture, hyperparameter, and data-processing details for all three experimental settings (synthetic manifolds, ERA5 weather forecasting, and cross-domain style transfer) are provided in Appendix M, and hyperparameter sensitivity sweeps are reported in Appendix J.

REFERENCES

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving Rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Martin Andrae, Tomas Landelius, Joel Oskarsson, and Fredrik Lindsten. Continuous ensemble weather forecasting with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Lucas Berry, Axel Brando, and David Meger. Shedding light on large generative networks: Estimating epistemic uncertainty in diffusion models. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael M. Bronstein, and Alexander Tong. SE(3)-stochastic flow matching for protein backbone generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kJFIH23hXb.

Xuenan Cao and Roozbeh Yousefzadeh. Extrapolation and AI transparency: Why machine learning models should reveal when they make decisions beyond their training. *Big Data & Society*, 10(1): 20539517231169731, 2023. doi: 10.1177/20539517231169731.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

Matthew Chan, Maria Molina, and Chris Metzler. Estimating epistemic and aleatoric uncertainty with a single model. *Advances in Neural Information Processing Systems*, 37:109845–109870, 2024.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks. *arXiv preprint arXiv:2003.09711*, 2020.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. ATOM: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 430–445. Springer, 2021.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023.

Hyunsun Choi, Eric Jang, and Alexander A Alemi. WAIC, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical Japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Earl A. Coddington and Norman Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1955.

Yajie Cui, Zhaoxiang Liu, and Shiguo Lian. A survey on unsupervised anomaly detection algorithms for industrial images. *IEEE Access*, 11:55297–55315, 2023. doi: 10.1109/ACCESS.2023.3282993.

Lars Doorenbos, Raphael Sznitman, and Pablo Márquez-Neila. Non-linear outlier synthesis for out-of-distribution detection. *arXiv preprint arXiv:2411.13619*, 2024.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36:60878–60901, 2023.

Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pp. 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2948–2957, 2023.

Julian Jorge Andrade Guerreiro, Naoto Inoue, Kento Masui, Mayu Otani, and Hideki Nakayama. LayoutFlow: Flow matching for layout generation. In *European Conference on Computer Vision*, pp. 56–72. Springer, 2024.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxCxhRcY7.

Alvin Heng, Harold Soh, et al. Out-of-distribution detection with a single unconditional diffusion model. *Advances in Neural Information Processing Systems*, 37:43952–43974, 2024.

Andrés Herrera-Poyatos, Javier Del Ser, Marcos López de Prado, Fei-Yue Wang, Enrique Herrera-Viedma, and Francisco Herrera. A framework for responsible ai systems: Building societal trust through domain definition, trustworthy ai design, auditability, accountability, and governance. *arXiv preprint arXiv:2503.04739*, 2025.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369. IEEE, 2010.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4 (26):eaau5872, 2019. doi: 10.1126/scirobotics.aau5872. URL https://www.science.org/doi/abs/10.1126/scirobotics.aau5872.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620 (7976):982–987, 2023.

Taekyung Ki, Dongchan Min, and Gyeongsu Chae. FLOAT: Generative motion latent flow matching for audio-driven talking portrait. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14699–14710, 2025.

Beomsu Kim, Jaemin Kim, Jeongsol Kim, and Jong Chul Ye. Generalized consistency trajectory models for image manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Zjv38dg1Hb.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Bernie Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. *Advances in Neural Information Processing Systems*, 36:28341–28364, 2023.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with Gaussian process priors for probabilistic time series forecasting. *arXiv preprint arXiv:2410.03024*, 2024.

Siqi Kou, Lei Gan, Dequan Wang, Chongxuan Li, and Zhijie Deng. BayesDiff: Estimating pixel-wise uncertainty in diffusion via Bayesian inference. In *International Conference on Learning Representations*, volume 2024, pp. 17046–17063, 2024.

Preethi Lahoti, Krishna Gummadi, and Gerhard Weikum. Responsible model deployment via model-agnostic uncertainty learning. *Machine Learning*, 112(3):939–970, 2023.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

Yann LeCun and Corinna Cortes. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31, 2018.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Hengzhuang Li and Teng Zhang. Outlier synthesis via Hamiltonian Monte Carlo for out-of-distribution detection. In *International Conference on Learning Representations*, volume 2025, pp. 57543–57582, 2025.

Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (OOD) detection: Masked image modeling is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11578–11589, June 2023.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023a.

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *International Conference on Machine Learning*, pp. 22528–22538. PMLR, 2023b.

Shane Lubold and Clark N Taylor. Formal definitions of conservative probability distribution functions (PDFs). *Information Fusion*, 88:175–183, 2022.

Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 34:19235–19247, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Giorgio Mariani, Irene Tallini, Emilian Postolache, Michele Mancusi, Luca Cosmo, and Emanuele Rodolà. Multi-source diffusion models for simultaneous music generation and separation. In *The Twelfth International Conference on Learning Representations*, 2024.

Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62): eabk2822, 2022. doi: 10.1126/scirobotics.abk2822. URL https://www.science.org/doi/abs/10.1126/scirobotics.abk2822.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1xwNhCcYm.

National Transportation Safety Board. Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018. Technical Report NTSB/HAR-19/03, National Transportation Safety Board, Washington, D.C., 2019. URL https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, volume 2011, pp. 7. Granada, 2011.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models: Open X-Embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Olivier C Pasche, Jonathan Wider, Zhongwei Zhang, Jakob Zscheischler, and Sebastian Engelke. Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, 4(1):e240033, 2025.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

Ziyang Rao, Yiren Zhao, Weiyu Guo, Ben Fei, Yandong Guo, and Hui Xiong. The geometry of flow-matching uncertainty: A cost-free uncertainty proxy and its application in flow-based vla failure detection. *arXiv preprint arXiv:2607.27933*, 2026.

Quentin Rouxel, Andrea Ferrari, Serena Ivaldi, and Jean-Baptiste Mouret. Flow matching imitation learning for multi-support manipulation. In *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, pp. 528–535. IEEE, 2024.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SyxIWpVYvr.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Zhengdao Shao, Liansheng Zhuang, Jie Yan, and Liting Chen. Conservative in-distribution Q-learning for offline reinforcement learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024. doi: 10.1109/IJCNN60899.2024.10650768.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Y. Qiang Sun, Pedram Hassanzadeh, Mohsen Zand, Ashesh Chattopadhyay, Jonathan Weare, and Dorian S. Abbot. Can AI weather models predict out-of-distribution gray swan tropical cyclones? *Proceedings of the National Academy of Sciences*, 122(21):e2420914122, 2025. doi: 10.1073/pnas.2420914122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2420914122.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5):405–420, 2018.

Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JHklpEZqduQ.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Matias Valdenegro-Toro and Radina Stoykova. The dilemma of uncertainty estimation for general purpose AI in the EU AI Act. *arXiv preprint arXiv:2408.11249*, 2024.

Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, Jianye Hao, and Bo Han. Out-of-distribution detection with implicit outlier transformation. *arXiv preprint arXiv:2303.05033*, 2023.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. IEEE, 2003.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 2024. arXiv:2110.11334, revised 2023.

Suhee Yoon, Sanghyu Yoon, Ye Seul Sim, Sungik Choi, Kyungeun Lee, Hye-Seung Cho, Hankook Lee, and Woohyung Lim. Diffusion-based semantic outlier generation via nuisance awareness for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13078–13086, 2025.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018b.

Xi Nicole Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. *Advances in Neural Information Processing Systems*, 37:107198–107224, 2024.

Appendix Contents

A Broader Impact **17**

B Limitations and Scope **17**

C Extended Related Work **18**

D Remarks and Proofs **19**

D.1 Proof of Proposition 4.1 19

D.2 Margin Propagation Around Boundary Samples 19

D.3 Proof of Theorem 4.2 20

D.4 A remark on the margin losses 20

D.5 Connecting the DOT Score to the Curvature Bound 21

D.6 Empirical Check of the Energy Excess at Inference 22

E Synergy of PGD and Contrastive Objectives **23**

F Generalization to Far Off-Manifold Conditions **23**

F.1 Comparison to a Jointly-Trained Classification-Head Baseline 25

G Split Conformal Prediction Calibration **26**

G.1 Interpretation of the FPR Metric 27

H Additional Detection Analyses on ERA5 **27**

H.1 Detection Metrics and Conformal Coverage 27

H.2 Early Detection During Integration 27

I Algorithmic Formulation of Diverging Flows **28**

J Ablations and Sensitivity Analysis **29**

J.1 Efficacy of Boundary Sampling Strategies 29

J.2 Ablation and Sensitivity Analysis on Synthetic Manifolds 30

J.3 Sensitivity Analysis on High Dimensional Weather Temperature Forecasting 32

J.4 Scaling to Higher-Resolution ERA5 33

K Visualization of Detection Landscapes **33**

L Comparison with Passive Trajectory Detectors **34**

L.1 DiffPath 34

L.2 Concurrent Work: Acceleration Proxy of Standard FM 35

M Experiment Details **35**

M.1 Theoretical Formulation of Detection Baselines 35

M.2 General Experimental Setup 36

M.3 Synthetic Manifold Experiment Details 36

M.4 Weather Temperature Forecasting 37

M.5 Cross-Domain Style Transfer Details 38

N Additional Qualitative Results **38**

A Broader Impact

The goal of this work is to advance the reliability of generative models in scientific and industrial tasks. By enabling models to detect and reject extrapolations, our method mitigates the risks of silent hallucinations in safety-critical applications such as weather forecasting, autonomous systems, and scientific discovery. We believe this work contributes positively to the responsible deployment of machine learning in high-stakes domains by transforming invisible failure modes into detectable signals. We do not foresee specific negative societal consequences, as the primary focus of this work is to advance machine learning for reliability and transparency.

B Limitations and Scope

Training cost of domain-agnostic boundary sampling. Diverging Flows adds no inference cost beyond the DOT score ($\sim 0.8\%$), but it increases training time when boundary samples are generated with a fully domain-agnostic method. With $\mathcal{K}$-step PGD, each training step requires $\mathcal{K}$ additional forward and backward passes to perturb the condition, plus one pass to evaluate the margins at the boundary samples. A naive implementation therefore costs up to $(\mathcal{K}+2)\times$ a standard FM step. This cost is paid once during training and does not grow with the number of queries at deployment.

Table 4 reports the measured cost on ERA5. Relative to an FM step, the margins at the boundary samples add about $0.9\times$ and each PGD step about $0.65\times$, so the measured cost stays below the $(\mathcal{K}+2)\times$ estimate. Taking the first PGD gradient from the graph of the FM loss, which has the same objective, saves one forward pass and $0.3\times$ for every $\mathcal{K}$ without changing the boundary samples.

Table 4: **Training cost on ERA5.** Wall-clock time per training step (U-Net, batch size 64, one RTX 4090, median of three interleaved runs of 50 steps each, data loading excluded) and ratio to standard FM. `Naive` runs PGD from scratch, and `Reuse` takes its first gradient from the FM loss. Bold marks the lowest cost of DiFlo. Inference cost is unchanged apart from the DOT score ($\sim 0.8\%$).

	Naive		Reuse	
Method	ms/step ↓	Ratio to FM ↓	ms/step ↓	Ratio to FM ↓
Standard FM	131.6	1.00	–	–
DiFlo ($\mathcal{K}=1$)	336.2	2.55	**296.3**	**2.25**
DiFlo ($\mathcal{K}=3$)	508.1	3.86	468.1	3.56
DiFlo ($\mathcal{K}=5$)	679.7	5.17	639.9	4.86

Reducing the cost of PGD. Several options can reduce this overhead without changing the sampler. A single step ($\mathcal{K}=1$) performs on par with $\mathcal{K}=5$ on the synthetic tasks (Table 12). Taking the first PGD gradient from the graph of the FM loss saves $0.3\times$ (Table 4), and reusing the gradient of the previous training step, in the spirit of "free" and fast adversarial training (Shafahi et al., 2019; Wong et al., 2020), would remove that step entirely. Finally, although we have not tested it, boundary samples could possibly be refreshed every few training steps instead of at every step, or the margins applied while fine-tuning a pretrained FM model instead of training from scratch, so that the overhead is limited to a short phase.

Moving away from iterative boundary samplers. Because Diverging Flows accepts boundary samples from any source, PGD can also be replaced altogether. When domain knowledge is available, cheap semantic transformations match or exceed PGD on images (Table 10). Without it, the OOD detection literature offers several non-iterative alternatives, beyond the classifier-based outlier synthesis methods reviewed in Appendix C. Mixup (Zhang et al., 2018a) and HamOS (Li & Zhang, 2025) sample outliers from the in-distribution data or features without gradients with respect to the input. Learned generators amortize the search into a single forward pass, as in AdvGAN (Xiao et al., 2018), which replaces PGD by a network that outputs the perturbation. More recently, diffusion models have been used to synthesize outliers close to the data, e.g., Dream-OOD (Du et al., 2023), SONA (Yoon et al., 2025) and NCIS (Doorenbos et al., 2024), and DOE (Wang et al., 2023) perturbs the model weights instead of the data. These methods were designed for classifiers on natural images, and adapting them to the conditioning spaces of scientific predictors is a promising direction. We

leave it to future work, since finding the best boundary sampler is outside the scope of this paper, whose aim is to show that extrapolation detection can be embedded in the flow geometry of a single model.

Types of off-manifold conditions. Off-manifold conditions include both inputs that are physically or semantically invalid and real states that are absent from the training data, such as the unprecedented extremes of Section 1. Our benchmarks test the first kind (synthetic hotspots, unrelated image domains). Real regimes held out from training, e.g., a season or a class of extreme events, are a natural next benchmark.

Valid trajectories. The margins act on off-manifold conditions, but valid trajectories of Diverging Flows can also deviate slightly from the straight path. Their energy excess remains small next to that of off-manifold conditions (Table 5), and fidelity is unaffected at the 50 integration steps we use. The effect in the few-step regime remains to be evaluated.

Hyperparameters in high dimension. On the spiral, detection is stable over wide hyperparameter ranges (Table 12). In high dimension, detection depends on the scale of the margins and of the perturbation radius (Table 13), which we address with simple rules: m_r grows with $\sqrt{d}$ (Appendix D.4), ϵ is set to the natural noise scale of the normalized domain, and the curvature term is kept gentle (Appendix J.3). With these rules, the ERA5 configuration transfers to 128×128 without retuning (Appendix J.4), and fidelity is unaffected across all tested configurations.

C Extended Related Work

Boundary samples and outlier exposure in classifiers. Training with samples near or beyond the data boundary is an established idea for discriminative OOD detection. Lee et al. (2017) generate boundary samples with a GAN and train the classifier to be uncertain on them, Outlier Exposure (Hendrycks et al., 2019) uses an auxiliary outlier dataset, ATOM (Chen et al., 2021) mines informative outliers, and ALOE (Chen et al., 2020) additionally uses adversarially perturbed inliers and outliers. VOS (Du et al., 2022) and NPOS (Tao et al., 2023) synthesize outliers in feature space, which removes the need for an auxiliary dataset (Appendix B discusses such samplers as replacements for PGD). Diverging Flows shares the principle that the model must see the boundary to learn it, and our PGD boundary samples play the role of the adversarial outliers of ALOE and ATOM. The difference is where the signal is placed. These methods shape the output distribution of a classifier, while Diverging Flows shapes the transport geometry of a conditional generative model, so the detection signal comes from the prediction itself and no classification head is required. A classification head trained on the same PGD boundary samples reaches an AUROC of only 0.53 on ERA5 (Appendix F.1).

Energy-based OOD scores. Liu et al. (2020) score inputs by the free energy of the classifier logits. Despite the shared term, the transport energy used here is a different quantity: the expected squared deviation of the learned velocity from the straight reference velocity (Eq. 2), which is defined for any conditional FM model and does not require logits.

Uncertainty estimation in diffusion models. BayesDiff (Kou et al., 2024) estimates pixel-wise uncertainty with a last-layer Laplace approximation, DECU (Berry et al., 2024) with an ensemble of conditional networks, and HyperDM (Chan et al., 2024) with a hypernetwork that samples model weights. These methods estimate epistemic uncertainty from the spread of multiple samples or weight draws. Diverging Flows instead obtains a detection score from a single deterministic trajectory of a single model, which keeps the inference cost of standard FM.

Straight flows. Rectified flow (Liu et al., 2023a) and minibatch OT couplings (Tong et al., 2023) straighten FM trajectories to reduce the number of integration steps. Diverging Flows keeps this efficiency for valid conditions and uses the loss of straightness as a detection signal for off-manifold ones.

Conformal outlier detection and selective prediction. Our threshold follows split conformal prediction (Gammerman et al., 1998; Angelopoulos et al., 2024), which is closely related to conformal p-values for outlier testing (Bates et al., 2023), and the resulting reject option connects Diverging Flows to selective prediction (Geifman & El-Yaniv, 2017), where a model abstains on inputs it cannot handle reliably.

D Remarks and Proofs

D.1 Proof of Proposition 4.1

Proof. To determine the Lipschitz constant L, we must bound the absolute difference $|\mathcal{E}(\tilde{c}) - \mathcal{E}(c)|$ for a perturbed off-manifold condition $\tilde{c}$ relative to a valid on-manifold condition c. By the triangle inequality for expectations:

$$|\mathcal{E}(\tilde{c}) - \mathcal{E}(c)| \leq \mathbb{E}_{t,x_t}\left[\left|||v_t^\theta(x_t,\tilde{c}) - u^{OT}||_2^2 - ||v_t^\theta(x_t,c) - u^{OT}||_2^2\right|\right] \tag{8}$$

For any fixed t and x_t, we can rewrite the difference of squared L_2 norms using the properties of inner products. Applying the Cauchy-Schwarz inequality followed by the triangle inequality yields:

$$\begin{aligned}
&\left|||v_t^\theta(x_t,\tilde{c}) - u^{OT}||_2^2 - ||v_t^\theta(x_t,c) - u^{OT}||_2^2\right| \\
&= \left|\langle v_t^\theta(x_t,\tilde{c}) - v_t^\theta(x_t,c), v_t^\theta(x_t,\tilde{c}) + v_t^\theta(x_t,c) - 2u^{OT}\rangle\right| \\
&\leq ||v_t^\theta(x_t,\tilde{c}) - v_t^\theta(x_t,c)||_2 \cdot ||v_t^\theta(x_t,\tilde{c}) + v_t^\theta(x_t,c) - 2u^{OT}||_2 \\
&\leq ||v_t^\theta(x_t,\tilde{c}) - v_t^\theta(x_t,c)||_2 \left(||v_t^\theta(x_t,\tilde{c})||_2 + ||v_t^\theta(x_t,c)||_2 + 2||u^{OT}||_2\right)
\end{aligned} \tag{9}$$

By hypothesis, the vector field is locally Lipschitz continuous with constant K, such that $||v_t^\theta(x_t,\tilde{c}) - v_t^\theta(x_t,c)||_2 \leq K||\tilde{c} - c||_2$. Furthermore, the vector field is uniformly bounded, such that $||v_t^\theta(\cdot)||_2 \leq M$, and $\|u^{OT}\|_2 \leq U$ by assumption. Substituting these upper bounds and factoring the deterministic constants out of the expectation gives:

$$\begin{aligned}
|\mathcal{E}(\tilde{c}) - \mathcal{E}(c)| &\leq \mathbb{E}_{t,x_t}\left[K||\tilde{c} - c||_2(2M + 2U)\right] \\
&= 2K(M+U)||\tilde{c} - c||_2
\end{aligned} \tag{10}$$

Thus, the Transport Energy Excess $\mathcal{E}(c)$ is Lipschitz continuous with constant $L \leq 2K(M+U)$. □

Remark D.1 (The Mechanics of Silent Extrapolation via Spectral Bias). The implicit regularization of deep neural networks, specifically their spectral bias (Rahaman et al., 2019), favors smooth, low-frequency functions. Standard Flow Matching fits the vector field on training conditions concentrated on $\mathcal{M}$ and places no constraint on its behavior nearby, so this bias favors extending the on-manifold field smoothly into the neighborhood of $\mathcal{M}$, where Proposition 4.1 then keeps the energy excess close to its on-manifold value. We do not claim a quantitative bound on the local Lipschitz constant, which can be large for trained networks. Whether the landscape is flat near $\mathcal{M}$ is therefore an empirical question, which the FM baselines in Section 5 and Appendix D.6 answer positively.

D.2 Margin Propagation Around Boundary Samples

Proposition 4.1 bounds how fast the transport energy can change around the valid manifold, which is the source of silent extrapolation. The same bound also limits how fast a margin enforced at a boundary sample can decay around it. Suppose that training raises the energy excess at a boundary sample $\tilde{c}$ to $\mathcal{E}(\tilde{c}) \geq \Delta > 0$. Applying Proposition 4.1 between $\tilde{c}$ and a nearby condition c' gives

$$\mathcal{E}(c') \geq \mathcal{E}(\tilde{c}) - L\|c' - \tilde{c}\|_2 \geq \Delta - L\|c' - \tilde{c}\|_2, \tag{11}$$

so the excess stays above $\Delta/2$ within a distance $\Delta/(2L)$ of the boundary sample. The size of Δ is set by the contrastive margins. For example, when the hinge in Eq. 3 is inactive on the training pairs, the velocity at $\tilde{c}$ deviates from u_t^{OT} by at least m_r, so the excess at $\tilde{c}$ is at least m_r^2.

Consequently, boundary samples that are sufficiently dense keep the energy excess high over a neighborhood of the boundary, and not only at the sampled points. The Lipschitz regularity that hides extrapolation near $\mathcal{M}$ therefore also allows a finite set of boundary samples to shape the energy landscape around it. A steeper landscape (larger L) shrinks this radius, so the samples must then be denser. This argument is local and depends on constants that are hard to estimate for deep networks, so we use it as an intuition for the barrier rather than as a quantitative guarantee, and verify its consequences empirically (Appendices D.6 and F).

D.3 Proof of Theorem 4.2

Proof. The generative flow must continuously connect the initial state x_0 to its final predicted state $\hat{x}_1$. We can decompose the learned conditional velocity field, evaluated along this generated trajectory, into the constant-velocity baseline of the realized path and a time-dependent geometric deviation:

$$v_t^\theta(\psi_t(x_0), \tilde{c}) = u^{\text{local}} + \tilde{\epsilon}_t \tag{12}$$

Because the integral of the total velocity over time must exactly equal the linear spatial displacement between the endpoints, we have:

$$\int_0^1 (u^{\text{local}} + \tilde{\epsilon}_t)dt = \hat{x}_1 - x_0 = u^{\text{local}} \implies \int_0^1 \tilde{\epsilon}_t dt = 0 \tag{13}$$

Thus, the geometric deviation $\tilde{\epsilon}_t$ has a temporal mean of zero.

Regularity of $\tilde{\epsilon}_t$. To rigorously apply the Poincaré–Wirtinger inequality, we must first establish that the error trajectory $\tilde{\epsilon}_t \in H^1([0,1];\mathbb{R}^d)$. By construction, our parameterized vector field $v_t^\theta(x, c)$ utilizes smooth activation functions (e.g., SiLU, GELU) and differentiable time embeddings, rendering it continuously differentiable with respect to both time and state. Consequently, the flow $v_t^\theta(\psi_t(x_0), \tilde{c})$ is continuously differentiable in t along the generated trajectory $\psi_t(x_0)$. Since the baseline transport vector u^{local} is constant with respect to time, the deviation $\tilde{\epsilon}_t$ inherently preserves this continuous differentiability. Furthermore, because the vector field is bounded by M, the trajectory stays within the ball of radius M around x_0 for $t \in [0,1]$. On this compact set, the continuously differentiable field has bounded derivatives, so the total time derivative $\dot{\tilde{\epsilon}}_t$ is square-integrable on $t \in [0,1]$. Therefore, the H^1 regularity requirement is inherently satisfied.

By hypothesis, the transport energy excess relative to the linear path is bounded below:

$$\int_0^1 \|\tilde{\epsilon}_t\|_2^2 dt \geq \Delta_{\text{local}} \tag{14}$$

Since u^{local} is a constant vector field, the acceleration of the trajectory is exactly the temporal derivative of the deviation: $\frac{d}{dt}v_t^\theta = \dot{\tilde{\epsilon}}_t$. Applying the Poincaré–Wirtinger inequality to the zero-mean function $\tilde{\epsilon}_t$, and noting that π^2 is exactly the first non-zero eigenvalue of the Neumann Laplacian on the interval $[0,1]$, we obtain the sharp bound:

$$\int_0^1 \|\tilde{\epsilon}_t\|_2^2 dt \leq \frac{1}{\pi^2} \int_0^1 \|\dot{\tilde{\epsilon}}_t\|_2^2 dt \tag{15}$$

Substituting the transport energy bound yields $\kappa \geq \pi^2 \Delta_{\text{local}}$. Any energy excess relative to the constant-velocity path is therefore accompanied by a proportional integrated squared acceleration. □

Remark D.2 (Dependence of u^{local} on $\tilde{c}$). We note that $u^{\text{local}} = \hat{x}_1 - x_0$ depends implicitly on the off-manifold condition $\tilde{c}$ through the generated endpoint $\hat{x}_1$. Consequently, the curvature lower bound $\kappa \geq \pi^2 \Delta_{\text{local}}$ is trajectory-specific rather than universal. This does not weaken the detection mechanism, since the bound is stated relative to each trajectory's own linear baseline. Whenever the contrastive objective induces an energy excess $\Delta_{\text{local}} > 0$ along a realized trajectory conditioned on $\tilde{c} \in \mathcal{M}^c$, its curvature is bounded below regardless of where the endpoint $\hat{x}_1$ lands in state space. The DOT score measures deviation relative to the same realized endpoint $\hat{x}_1$, so the score and the bound refer to the same baseline. Appendix D.6 verifies on ERA5 that this excess is present at inference.

D.4 A remark on the margin losses

Penalizing the vector field at an off-manifold condition $\tilde{c}$ risks inadvertently "bleeding" into the flow of the nearby valid condition c due to network continuity. This is mitigated by two mechanisms. First, our margin-based repulsion loss $\mathcal{L}_{\texttt{repel}}$ zero-outs its gradient once the minimum required transport divergence is met, preventing unbounded local distortion. Second, assuming the network is locally Lipschitz, the spatial bleed scales proportionally with the distance to the boundary sample. For instance, if an adversarial strategy like PGD is used to generate boundary samples within an ϵ-ball, the flow deviation at the valid condition is bounded by $\mathcal{O}(\epsilon)$. By enforcing a tight search radius (Appendix J.2), the valid predictive flow remains safely isolated, a safeguard empirically verified by our fidelity metrics across all tasks.

D.4.1 SCALING OF THE KINETIC MARGIN m_r WITH DIMENSIONALITY

The kinetic margin m_r bounds the difference of two L_2 norms in the ambient data space $\mathbb{R}^d$ (Eq. 3). Consider a valid-condition velocity $v_t^\theta(x_t, c)$ whose d independent coordinates each deviate from u_t^{OT} by an approximately i.i.d. random variable with variance σ_v^2 under the training distribution. By concentration of measure for sums of independent bounded-variance terms, the norm of this deviation vector concentrates around $\|v_t^\theta(x_t, c) - u_t^{OT}\|_2 \approx \sqrt{d}\,\sigma_v$, with fluctuations of order $\mathcal{O}(1)$ around this value (i.e. the relative fluctuation shrinks as d grows). For the margin to separate a genuinely off-manifold velocity from this concentrating baseline with a fixed, dimensionality-independent *safety factor*, m_r must therefore scale as

$$m_r \propto \sqrt{d}\,\sigma_v, \tag{16}$$

rather than remaining fixed. A margin held constant across dimensions would either vanish relative to the ambient norm scale for large d (too weak to separate) or dominate it for small d (destabilizing training), consistent with the observed behavior in our experiments, where the fitted m_r grows by roughly two orders of magnitude between the 2D spiral task ($d = 2$) and the 64×64 weather task ($d = 4096$; Table 13). As with $\epsilon \in [0.07, 0.2]$ (Appendix J.2), Eq. 16 is intended as a starting point that gets a practitioner within the right order of magnitude, not an exact target. This is consistent with the hinge formulation of Φ, which only requires clearing a threshold rather than matching a precise value. The rule fixes how m_r scales with d at a fixed deviation scale σ_v, which itself depends on the dataset. For instance, style transfer uses $m_r = 10$ at $d = 3072$ and ERA5 uses $m_r = 100$ at $d = 4096$; the dimensions differ by a factor of only 1.15 in $\sqrt{d}$, so the difference in m_r reflects a different σ_v rather than a different d.

Remark D.3 (Necessity of Both Margin Losses). $\mathcal{L}_{\texttt{repel}}$ alone does not rule out a degenerate *scalar-rescaling* solution $v_t^\theta(x_t, \tilde{c}) = \gamma\, u_t^{OT}$ for some $\gamma \neq 1$: this satisfies the kinetic margin (it changes speed) while leaving the trajectory's geodesic direction, and hence the deviation field $\tilde{\epsilon}_t$'s useful signal, unchanged. $\mathcal{L}_{\texttt{curve}}$ rules this out directly, since it penalizes angular similarity to u_t^{OT} independently of speed. This matches the ablation in Table 11: removing $\mathcal{L}_{\texttt{curve}}$ degrades AUROC to 0.931 and removing $\mathcal{L}_{\texttt{repel}}$ degrades it further to 0.711, confirming that repulsion alone carries most of the raw detection signal but curvature is what prevents this degenerate edge case.

D.5 CONNECTING THE DOT SCORE TO THE CURVATURE BOUND

We show here that the DOT score of Eq. 7 is an exact discretization of the deviation field $\tilde{\epsilon}_t$ used in the proof of Theorem 4.2 (Appendix D.3), rather than an independently chosen heuristic.

The identity. Recall from the proof of Theorem 4.2 that the realized velocity along a generated trajectory decomposes as $v_t^\theta(\psi_t(x_0), \tilde{c}) = u^{\text{local}} + \tilde{\epsilon}_t$, where $u^{\text{local}} = \hat{x}_1 - x_0$ is the constant-velocity path between the realized endpoints and $\int_0^1 \tilde{\epsilon}_t\, dt = 0$. The realized state at time t_i is

$$\hat{x}_{t_i} = x_0 + \int_0^{t_i} v_s^\theta(\psi_s(x_0), \tilde{c})\, ds = x_0 + t_i u^{\text{local}} + \int_0^{t_i} \tilde{\epsilon}_s\, ds. \tag{17}$$

The DOT reference anchor is, by definition, the linear interpolant $x_{t_i}^{\text{ref}} = (1 - t_i)x_0 + t_i\hat{x}_1 = x_0 + t_i u^{\text{local}}$, which is the same constant-velocity baseline that appears in the theorem's proof. Subtracting,

$$\hat{x}_{t_i} - x_{t_i}^{\text{ref}} = \int_0^{t_i} \tilde{\epsilon}_s\, ds, \tag{18}$$

so the per-step gap DOT measures is exactly the running (partial) integral of $\tilde{\epsilon}_t$, and

$$S_{DOT}(\hat{x}_{0:1}) = \sum_{i=1}^{N} \frac{1}{D} \left\| \int_0^{t_i} \tilde{\epsilon}_s\, ds \right\|_1. \tag{19}$$

Why this connects to κ. Theorem 4.2 shows that whenever $\int_0^1 \|\tilde{\epsilon}_t\|_2^2\, dt \geq \Delta_{\text{local}} > 0$, the curvature $\kappa = \int_0^1 \|\dot{\tilde{\epsilon}}_t\|_2^2\, dt$ is lower-bounded by $\pi^2\Delta_{\text{local}}$. Because $\tilde{\epsilon}_t$ is zero-mean (Appendix D.3) but carries non-trivial energy whenever $\Delta_{\text{local}} > 0$, it cannot be identically zero on $[0, 1]$; by continuity of $\tilde{\epsilon}_t$, this forces the running integral $\int_0^{t_i} \tilde{\epsilon}_s\, ds$ to be bounded away from zero over a non-trivial range of

t_i, which is precisely the quantity summed by S_{DOT}. In this sense S_{DOT} and κ are two different norms/aggregations (a summed running L_1 integral vs. an integrated squared L_2 derivative) applied to the same underlying deviation field $\tilde{\epsilon}_t$. The running integral is, however, a weaker norm: a deviation that oscillates quickly can carry a large Δ_{local} while its running integral stays small, so S_{DOT} is not lower-bounded by Δ_{local} in general. The link between the two is therefore empirical, and on ERA5 they are closely correlated (Spearman $\rho = 0.97$, Appendix D.6).

Why L_1 rather than the theorem's L_2. We deliberately depart from the theorem's L_2 norm when computing S_{DOT} in practice. The L_2 norm on $\dot{\tilde{\epsilon}}_t$ was required for the Poincaré–Wirtinger step in the proof, but at inference we have no need to differentiate the trajectory again; the L_1 norm on the running integral is (i) cheaper, requiring only a running sum and an absolute difference rather than a squared norm, and (ii) more sensitive to spatially localized deviation, since a single large-magnitude coordinate is not down-weighted relative to many small ones the way it would be by an averaged squared norm. In our experiments, we observed no noticeable difference in detection performance between the two norms, and we keep L_1 because it is cheaper and also covers this edge case of spatially localized deviations. See Remark D.3 for why both $\mathcal{L}_{\texttt{repel}}$ and $\mathcal{L}_{\texttt{curve}}$ are required for this identity to be non-degenerate.

D.6 Empirical Check of the Energy Excess at Inference

Theorem 4.2 applies to any trajectory whose energy excess reaches Δ_{local}, whereas the contrastive margins are only enforced during training, along interpolation paths at boundary samples. We therefore check whether this excess is present at inference, for unseen conditions and along the trajectories actually produced by the ODE solver. We use the ERA5 setting of Section 5.2 and evaluate a standard FM model and Diverging Flows on the same test set and the same hotspot conditions, for three training seeds. For each trajectory, we decompose the Euler velocities as $v_{t_i} = u^{\text{local}} + \tilde{\epsilon}_{t_i}$ and compute the per-dimension quantities $\Delta_{\text{local}} = \frac{1}{D}\int_0^1 \|\tilde{\epsilon}_t\|_2^2\, dt$ and $\kappa = \frac{1}{D}\int_0^1 \|\dot{\tilde{\epsilon}}_t\|_2^2\, dt$ with the step size of the solver, starting from the source sample x_0 and using all 50 steps.

Table 5: **Energy excess and curvature along realized ERA5 trajectories.** Medians over test (on) and hotspot (off) conditions, mean and standard deviation over three seeds. AUROC columns use the corresponding quantity as a detection score, and ρ is the Spearman correlation between DOT and Δ_{local} over both groups. Absolute values depend on the coupling learned by each model and are not comparable across rows, so the informative quantity is the contrast between on and off within a model.

	Δ_{local} ($\times 10^{-3}$)			κ		DOT	
Model	On	Off	AUROC	On	Off	AUROC	ρ
Standard FM	0.31 $\pm$0.02	0.34 $\pm$0.02	0.615 $\pm$0.040	0.22 $\pm$0.004	0.24 $\pm$0.01	0.667 $\pm$0.023	0.79 $\pm$0.04
Diverging Flows	0.91 $\pm$0.32	3.76 $\pm$2.00	0.975 $\pm$0.024	0.75 $\pm$0.20	3.78 $\pm$3.17	0.995 $\pm$0.005	0.97 $\pm$0.01

Table 5 and Figure 5 show that standard FM produces the same trajectories for valid and hotspot conditions. Both groups have nearly the same energy excess (median ratio 1.08 ± 0.03), and neither Δ_{local} nor the DOT score separates them, which is the silent extrapolation described by Proposition 4.1. Under Diverging Flows, valid conditions keep a small energy excess, while hotspot conditions reach a median excess $4.2\times$ larger on average (between $2.3\times$ and $5.3\times$ across seeds). Because both models are evaluated on identical conditions, this separation comes from the training objective and not from the inputs themselves. The DOT score closely follows Δ_{local} ($\rho = 0.97 \pm 0.01$), which supports its use as a proxy for the deviation field. The contrast in energy excess varies across seeds ($2.3\times$ to $5.3\times$), but the DOT score separates the two groups in every seed (AUROC 0.990 to 0.999), since it only requires the off-manifold excess to exceed the valid one, not a fixed ratio.

The hotspot conditions are not uniform. Some deviate strongly from the start of integration, while those closer to the manifold remain near the linear path early on and diverge mainly as t approaches 1, which explains the wide band in Figure 5. Since DOT accumulates the state gap over the whole trajectory, it also captures these late deviations and reaches a higher AUROC than Δ_{local} alone. Finally, the inequality $\kappa \geq \pi^2 \Delta_{\text{local}}$ holds for every trajectory of both models, as it must for any zero-mean deviation. The informative quantity is therefore whether Δ_{local} is large, and not the bound itself.

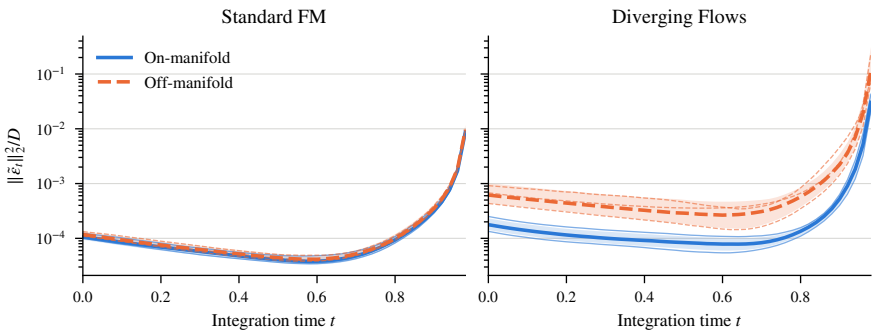


Figure 5: **Deviation from the constant-velocity path along ERA5 trajectories.** Median (thick line) and interquartile range (band) of $\|\tilde{\epsilon}_t\|_2^2/D$ over integration time for test (on-manifold) and hotspot (off-manifold) conditions, pooled over three seeds. Thin lines show the median of each seed. Standard FM follows the same path for both groups, while Diverging Flows separates them over the whole trajectory.

E Synergy of PGD and Contrastive Objectives

We visualize the dynamic synergy between our contrastive margins and PGD to demonstrate how the model explicitly learns the manifold boundary. Figures 6 and 7 track the spatial distribution of off-manifold conditions generated by PGD on the 2D Spiral manifold across training steps for both standard FM and Diverging Flows.

Under standard FM training (Figure 6), the network's transport energy landscape remains inherently flat due to the smoothness bias of the baseline objective. Consequently, when PGD perturbations optimize to maximize this baseline transport energy, the process lacks strong directional guidance. The resulting boundary samples remain diffusely scattered around the valid manifold $\mathcal{M}$. Even after 100K training steps, PGD fails to outline a distinct boundary, yielding weak samples that provide no meaningful safety constraints.

In contrast, under Diverging Flows (Figure 7), the model actively shapes the energy landscape. At initialization (Step 0), the boundary samples are similarly diffuse. However, as training progresses, the structural penalties ($\mathcal{L}_{\texttt{repel}}$ and $\mathcal{L}_{\texttt{curve}}$) penalize these initial points, which systematically steepens the local gradients around the manifold support. This dynamic steepening guides PGD to discover progressively harder, highly localized boundary samples. By 50K steps, these adversarial samples tightly trace the contour of the valid conditions, establishing the sharp geometric transition required for robust extrapolation awareness.

F Generalization to Far Off-Manifold Conditions

A standard critique of targeted boundary sampling strategies, such as adversarial PGD, is the risk of manifold boundary overfitting. By explicitly optimizing the contrastive objective (Eq. 5) on perturbations that lie in the local neighborhood (ϵ-ball) of the valid data support $\mathcal{M}$, one might hypothesize that the model only learns a localized threshold of extrapolation awareness. If true, this would leave the deep ambient space (conditions residing far off-manifold) vulnerable to the same silent extrapolation hazards observed in unregularized Flow Matching models.

However, the sharp geometric transition induced by the objective of Section 4 mitigates this failure mode. By inducing transport inefficiency at the boundary $\partial\mathcal{M}$, our contrastive objective builds a high-energy barrier that encloses the support of the valid conditioning distribution.

Optimization Argument. By the argument of Appendix D.2, the margin enforced at each boundary sample persists over a neighborhood of radius $\Delta/(2L)$, so boundary samples that cover $\partial\mathcal{M}$ form a high-energy shell around $\mathcal{M}$. The synergy with PGD (Appendix E) places these samples along the whole contour of the valid support. Lipschitz continuity alone does not prevent the energy

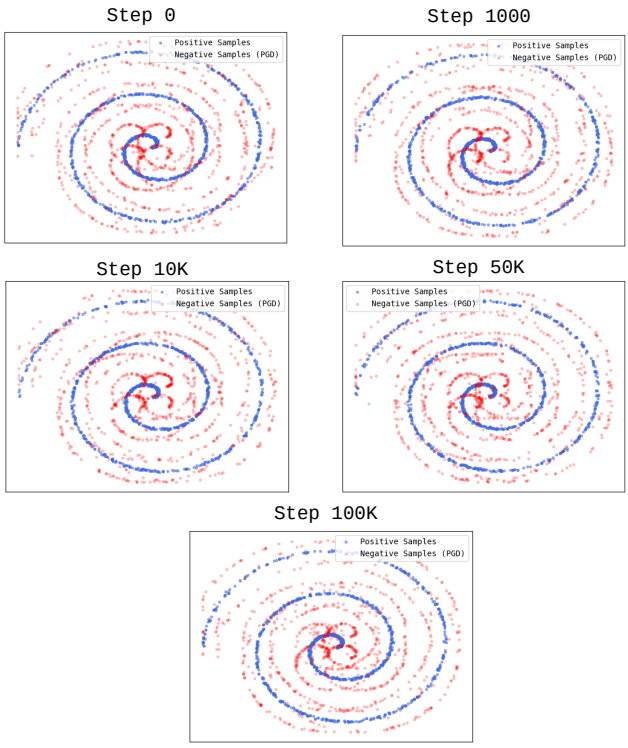


Figure 6: **Evolution of PGD boundary samples during standard FM training on the 2D Spiral.** Blue points represent valid on-manifold conditions ($c \in \mathcal{M}$), while red points denote adversarial off-manifold queries ($\tilde{c} \notin \mathcal{M}$). Because standard FM flattens the energy landscape, local gradients are weak; PGD fails to find a sharp boundary, leaving the boundary samples diffusely scattered even at late training stages.

from decreasing again far from this shell, so the behavior beyond it follows from how the network generalizes. The FM objective supervises the transport energy only on $\mathcal{M}$, and low energy at off-manifold conditions arises from the generalization of the network from $\mathcal{M}$ (Proposition 4.1), which is precisely the source of silent extrapolation. The shell interrupts this generalization at the boundary. Recovering low energy beyond it would require the vector field to vary even more sharply, a high-frequency behavior that the spectral bias of neural networks disfavors (Rahaman et al., 2019). Moreover, conditions are normalized to $[-1, 1]$, which bounds how far any input can lie from the shell. We therefore expect far off-manifold trajectories to remain inefficient and detectable via the DOT score, and verify this empirically below.

Empirical Validation. To empirically validate this global geometric constraint, we evaluate our Diverging Flows model, trained on the ERA5 Weather Temperature Forecasting dataset using local PGD mining ($\epsilon = 0.1$), against severe far off-manifold conditioning inputs. We subject the physical weather model to conditions drawn from entirely disjoint semantic manifolds: MNIST (LeCun & Cortes, 1998), Fashion-MNIST (FMNIST) (Xiao et al., 2017), and Kuzushiji-MNIST (KMNIST) (Clanuwat et al., 2018), as well as pure structural noise (Uniform and Gaussian). All far off-manifold inputs are upsampled to the 64×64 grid resolution and normalized to the $[-1, 1]$ interval to ensure they trigger the same magnitude-based activations within the U-Net backbone.

As demonstrated in Table 6, the model achieves perfect extrapolation detection across all evaluated regimes. Although these inputs differ strongly from weather fields, the likelihood of a standard FM model trained on the same data separates them only partially (AUROC 0.65 to 0.73), consistent with the known failure of likelihood-based detection (Nalisnick et al., 2019). These results indicate that local boundary sampling under the Diverging Flows objective is sufficient to constrain the ambient space in practice, in line with the optimization argument above. The resulting barrier around the conditioning manifold may also open opportunities for manifold exploration in FM models.

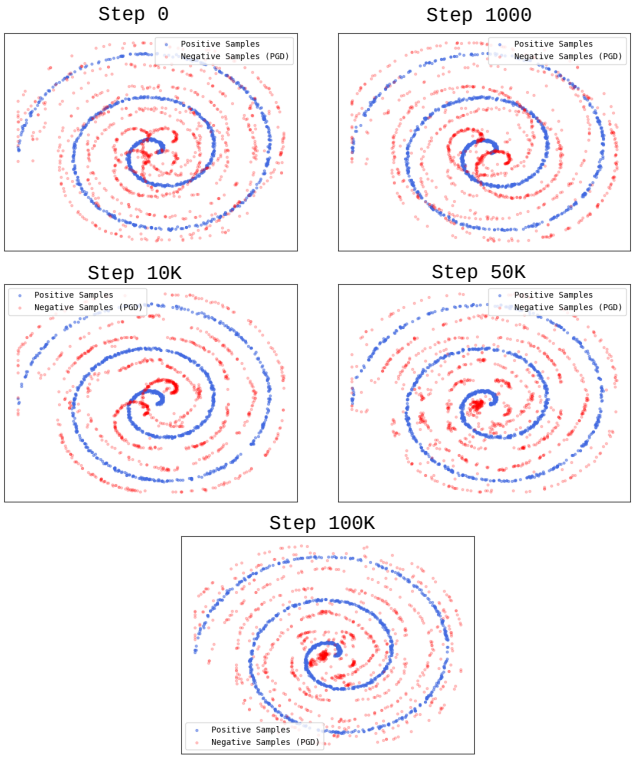


Figure 7: **Evolution of PGD boundary samples during Diverging Flows training on the 2D Spiral.** Blue points represent valid on-manifold conditions ($c \in \mathcal{M}$), while red points denote adversarial queries ($\tilde{c} \notin \mathcal{M}$). Driven by the contrastive margins, the local energy landscape steepens over time, actively guiding PGD to discover boundary samples that tightly trace the precise boundary of the valid data support.

Table 6: **Far Off-Manifold Detection Performance (ERA5 Model).** The DiFlo model, trained exclusively with local boundary perturbations, generalizes its detection to noise and unrelated image datasets, with an AUROC of 1.000 on all of them, whereas the likelihood of standard FM separates them only partially.

	AUROC ↑	
Far Off-Manifold Condition (Dataset)	FM-Likelihood	DiFlo (Ours)
Gaussian Noise ($\mathcal{N}(0, I)$)	0.73	**1.000**
Uniform Noise ($\mathcal{U}[-1, 1]$)	0.70	**1.000**
MNIST (Semantic Shift)	0.65	**1.000**
FMNIST (Semantic Shift)	0.69	**1.000**
KMNIST (Semantic Shift)	0.68	**1.000**

F.1 Comparison to a Jointly-Trained Classification-Head Baseline

The results above show that DiFlo generalizes detection far beyond its training boundary samples. A natural question is whether this generalization stems from the flow-geometry mechanism itself, or simply from the availability of off-manifold data during training, i.e., whether a conventional discriminative detector trained on the exact same PGD boundary samples would generalize just as well. To isolate this, we train a binary classification head jointly with the same FM backbone (the network outputs twice the number of channels, split into two halves: the first is the predicted vector field and the second is linearly projected and passed through a softmax to classify the condition), optimized with binary cross-entropy (BCE) on the identical PGD boundary samples used to train DiFlo, and compare its detection AUROC against ours on both the synthetic regression task and ERA5.

Table 7: **Classification-Head Baseline on Identical Boundary Samples.** A discriminative head trained via BCE on the same PGD boundary samples as DiFlo fits its own training distribution but collapses on unseen off-manifold conditions (near-chance on ERA5), while DiFlo generalizes without any auxiliary detector.

	Synthetic Regression		ERA5	
Algorithm	AUROC ↑	MSE ↓	AUROC ↑	MSE ↓
FM + Cls. Head	0.71	$8.45 \cdot 10^{-5}$	0.53	$9.04 \cdot 10^{-4}$
DiFlo (Ours)	**0.998**	$8.48 \cdot 10^{-5}$	**0.995**	$8.22 \cdot 10^{-4}$

As shown in Table 7, the classification-head baseline fits its own training boundary samples reasonably on the synthetic task (AUROC 0.71) but collapses to near-chance on ERA5 (AUROC 0.53, true positive rate of only 5.3% at a false positive rate of 0.8%), despite matching DiFlo's fidelity (MSE) on both tasks. We attribute this gap to the objective itself: the BCE gradient never references the optimal-transport target u_t^{OT}, so it provides no incentive to remove the smoothness-driven flat-energy solution described in Proposition 4.1 anywhere other than at the specific samples it was trained on; it learns a decision boundary local to those points rather than a geometric property of the vector field itself. Unlike $\mathcal{L}_{\texttt{repel}}$ and $\mathcal{L}_{\texttt{curve}}$, which operate in the same coordinate system as the fidelity loss $\mathcal{L}_{\texttt{FM}}$, a separate BCE head optimizes an objective that is unrelated to the transport geometry.

G Split Conformal Prediction Calibration

To ensure safety-critical reliability during deployment, we employ Split Conformal Prediction (Gammerman et al., 1998; Angelopoulos et al., 2024) to construct a statistically valid decision region. Because our objective is to detect geometric turbulence, we specifically frame this as a Conformal Anomaly Detection problem. Given a user-specified error rate $\alpha \in (0, 1)$, our goal is to identify a validity interval $[q_{lo}, q_{hi}]$ such that the probability of a valid, on-manifold input falling outside it is at most α.

We reserve a calibration dataset of M_{cal} exchangeable on-manifold conditions. We first compute the divergence score S_{DOT} for each sample and sort them in ascending order, denoting the sorted set as $\mathcal{S}_{cal} = \{s_{(1)}, s_{(2)}, \ldots, s_{(M_{\text{cal}})}\}$. We split the error budget equally between the two tails and set

$$q_{lo} = s_{(k_{lo})}, \quad k_{lo} = \lfloor (M_{\text{cal}} + 1)\, \alpha/2 \rfloor, \qquad q_{hi} = s_{(k_{hi})}, \quad k_{hi} = \lceil (M_{\text{cal}} + 1)(1 - \alpha/2) \rceil. \tag{20}$$

At inference time, a query condition c_{test} is accepted as valid if and only if its score lies within the calibrated interval:

$$q_{lo} \leq S_{DOT}(c_{test}) \leq q_{hi} \tag{21}$$

Otherwise, the generative process is flagged as an unsafe extrapolation. Based on the exchangeability of the calibration and test data, this procedure provides a guarantee on the marginal coverage. The probability that a valid on-manifold input is correctly accepted satisfies:

$$\mathbb{P}\left(q_{lo} \leq S_{DOT}(c_{test}) \leq q_{hi}\right) \geq 1 - \alpha \tag{22}$$

This ensures that the False Rejection Rate (FRR), the rate at which valid inputs are incorrectly flagged as unsafe, is at most α. The guarantee is marginal and assumes that valid conditions at deployment are exchangeable with the calibration set; it does not hold under a shift of the valid conditions themselves.

Role of the lower bound. The upper bound q_{hi} flags trajectories that deviate more from the constant-velocity path than valid ones do, which is the behavior induced by Diverging Flows. The lower bound q_{lo} additionally flags trajectories that are unusually straight. This covers degenerate fields that satisfy the margins without curving the trajectory, e.g., a large constant velocity, whose DOT score is close to zero. In our experiments, the lower bound is redundant. On ERA5 (three seeds, $\alpha = 0.05$), no hotspot condition scores below q_{lo} in any seed, so the interval and the upper bound alone flag the same hotspot conditions (88.7% on average). The lower tail only costs part of the error budget on valid conditions, which remains within α. Over random calibration splits, this costs Diverging Flows about four points of detection at $\alpha = 0.05$ and about two points from $\alpha = 0.10$ (Table 8).

G.1 Interpretation of the FPR Metric

In our main text experiments, we report the False Positive Rate (FPR) at a fixed calibration error rate of $\alpha = 0.05$ (which corresponds to a Target Coverage of 95% for valid data). It is critical to distinguish this evaluation metric from the FRR discussed above.

Because split conformal prediction keeps the coverage of valid data at least $1-\alpha$ under exchangeability (Appendix G), the quality of the geometric detection metric is determined entirely by its False Positive Rate, defined here as the fraction of *off-manifold* samples that falsely fall within the acceptable $[q_{lo}, q_{hi}]$ interval. Since the interval is two-sided, the FPR and the AUROC, which only ranks scores in one direction, can differ: off-manifold scores that fall below q_{lo} are rejected but lower the AUROC. This explains rows such as Diverging Flows w/o $\mathcal{L}_{\text{repel}}$ in Table 11 (AUROC 0.711, FPR 20.45%).

The FPR is best read against two reference cases. An ideal detector perfectly separates the score distributions of on- and off-manifold inputs, so even when the interval is widened to cover 99% of valid data, the FPR (extrapolation leakage) remains near 0%. At the other extreme, if the two score distributions overlap perfectly, as for the overly smooth standard FM models, accepting $X\%$ of the valid data implies accepting approximately $X\%$ of the invalid data. The FPR then approaches the coverage, as with the $\sim 95\%$ FPR of unregularized models in Table 1 at 95% coverage.

H Additional Detection Analyses on ERA5

We analyze further the ERA5 scores of Section 5.2, using the same three training seeds as Appendix D.6 for standard FM and Diverging Flows, and report means over seeds.

H.1 Detection Metrics and Conformal Coverage

Table 8 reports threshold-free metrics of the DOT score and the behavior of split conformal calibration. For the conformal columns, we split the valid conditions at random into equal calibration and test halves 2000 times, calibrate either the two-sided interval of Appendix G or an upper bound alone, and report the mean rate of rejected valid conditions and of flagged hotspot conditions, averaged over seeds.

Table 8: **Detection metrics and conformal calibration on ERA5 (three seeds).** FPR@95 is the fraction of hotspot conditions accepted when 95% of valid conditions are accepted. Conformal columns report valid conditions rejected / hotspot conditions flagged (%), averaged over 2000 random calibration splits per seed and over seeds. The standard deviation over seeds of the flagged rate of DiFlo is 8.4 points for the interval at $\alpha = 0.05$ and at most 3.4 points otherwise.

					Rejected / flagged (%)		
Model	Threshold	AUROC ↑	AUPR ↑	FPR@95 ↓	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$
Standard FM	Interval	0.667 ±0.023	0.657 ±0.040	83.3 ±5.3%	4.4 / 8.7	9.9 / 17.3	19.8 / 30.1
	Upper only				5.0 / 16.0	9.9 / 26.7	19.9 / 40.5
DiFlo	Interval	**0.995 ±0.005**	**0.995 ±0.005**	**2.6 ±3.3%**	4.4 / 93.0	9.9 / 97.2	19.8 / 99.4
	Upper only				4.9 / 97.2	9.9 / 99.4	19.9 / 99.9

The rate of rejected valid conditions stays within α in every setting, as guaranteed by split conformal prediction, and Diverging Flows flags 93% of hotspot conditions at $\alpha = 0.05$ and at least 97% from $\alpha = 0.10$. With the interval calibrated on the held-out calibration set at $\alpha = 0.05$, no hotspot condition of Diverging Flows scores below q_{lo} in any seed, and the interval and the upper bound alone flag the same conditions. Over random splits, spending half of the error budget on the lower tail costs about four points of detection at $\alpha = 0.05$ and about two points from $\alpha = 0.10$, which we accept in exchange for rejecting degenerate straight trajectories (Appendix G).

H.2 Early Detection During Integration

As an additional analysis, we examine whether extrapolations can be detected before integration ends. The DOT score compares each state with the straight path to the final prediction $\hat{x}_1$, so it is available only at the end of integration. Since off-manifold trajectories of Diverging Flows deviate from the

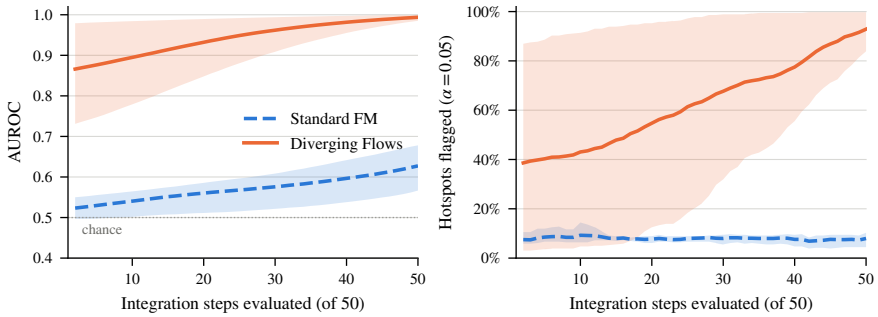


Figure 8: **Early detection on ERA5 (three seeds).** Detection with the causal score S_T as a function of the number T of integration steps evaluated, mean over seeds (line) and range across seeds (band). Left: AUROC. Right: fraction of hotspot conditions flagged by a two-sided conformal interval at $\alpha = 0.05$ calibrated on the held-out calibration set with the same number of steps.

start of integration (Figure 5), we also evaluate a causal score that replaces the unknown endpoint with the initial velocity. After T steps,

$$S_T = \sum_{i=1}^{T} \frac{1}{D} \left\| \hat{x}_{t_i} - \left(x_0 + t_i \, v_0^{\theta}(x_0, c) \right) \right\|_1 , \tag{23}$$

which only uses the states computed so far. Figure 8 shows its detection performance as a function of T. For Diverging Flows, the AUROC after 2 steps (4% of the integration) is 0.87 on average and ranges from 0.73 to 0.98 across seeds, whereas it stays close to chance for standard FM (0.50 to 0.55). It increases with T and reaches at least 0.985 at the end of integration for every seed. The rate of flagged hotspot conditions at $\alpha = 0.05$ varies more across seeds: after 2 steps it is 87%, 26% and 3%, and after 25 steps 97%, 69% and 18%. The early deviation is therefore present in every seed, but its strength depends on the training run, and the full DOT score remains the reliable detector.

A two-stage scheme can therefore stop part of the extrapolations after a few network evaluations and hand them to a fallback, and apply the full DOT score to the remaining ones at the end of integration. Since the early detection rate depends on the training run, it should be measured on the calibration set before relying on it in time-critical settings such as robotics.

I Algorithmic Formulation of Diverging Flows

Standard Flow Matching inference involves numerically integrating the learned velocity field v_t^{θ} from a prior noise sample x_0 to materialize a data sample $\hat{x}_1$. Algorithm 2 details this procedure within the Diverging Flows framework. First, we generate the full trajectory $\{\hat{x}_{t_i}\}_{i=1}^{N}$ using a numerical ODE solver. Second, we define our constant-velocity geometric anchor x_t^{ref} as the linear interpolation between the fixed source x_0 and the generated prediction $\hat{x}_1$. As defined in Eq. 7, the divergence score S_{DOT} is computed as the L_1 deviation between the realized neural trajectory and this ideal anchor, spatiotemporally averaged over all dimensions and summed over time steps.

While the exact S_{DOT} value fundamentally depends on the discrete steps produced by the numerical solver, we observe that the relative separation between on- and off-manifold inputs remains highly robust. As demonstrated in Table 9, detection performance (AUROC) and generative quality (LPIPS) remain stable across different solver families (Euler, Heun, RK4) and step counts, indicating that valid and invalid trajectories are affected similarly by discretization.

To formulate a rigorous binary decision during deployment, we compare the calculated divergence score S_{DOT} against the conformal interval $[q_{lo}, q_{hi}]$, which is pre-calibrated on a hold-out set of valid conditions at a user-specified significance level α. The input condition c is accepted as valid if and only if $q_{lo} \leq S_{DOT} \leq q_{hi}$; otherwise, the generative process is flagged as an unsafe extrapolation.

Algorithm 1 Training Algorithm of DiFlo

Require: Dataset $\mathcal{D}$, Loss weights λ, β, Margins m_r, m_c, learning rate α_{lr}
Initialize flow model parameters θ
while not converged **do**
1. Coupling
Sample batch $(x_1, c) \sim \mathcal{D}$, prior noise $x_0 \sim p_0$, time $t \sim \mathcal{U}[0, 1]$
Construct linear interpolant: $x_t = (1-t)x_0 + tx_1$
Compute target velocity: $u_t = x_1 - x_0$
2. Boundary Sampling
$\tilde{c} \leftarrow \textsc{BoundarySampling}(c, x_t, u_t, \theta)$ ▷ *Agnostic sampler (e.g., PGD, Mixup)*
3. Velocity Predictions
$\hat{v}_c \leftarrow v_t^\theta(x_t, c)$ ▷ *On-manifold field*
$\hat{v}_{\tilde{c}} \leftarrow v_t^\theta(x_t, \tilde{c})$ ▷ *Off-manifold field*
4. Objective Computation
$\mathcal{L}_{\texttt{FM}} = \|\hat{v}_c - u_t\|_2^2$
$\mathcal{L}_{\texttt{repel}} = \Phi\big(\|u_t - \hat{v}_c\|_2 - \|u_t - \hat{v}_{\tilde{c}}\|_2 + m_r\big)$
$\mathcal{L}_{\texttt{curve}} = \Phi\big(d_{cos}(u_t, \hat{v}_c) - d_{cos}(u_t, \hat{v}_{\tilde{c}}) + m_c\big)$
$\mathcal{L}_{\texttt{DiFlo}} = \mathcal{L}_{\texttt{FM}} + \lambda\mathcal{L}_{\texttt{repel}} + \beta\mathcal{L}_{\texttt{curve}}$
5. Optimization Step
$\theta \leftarrow \theta - \alpha_{\text{lr}}\nabla_\theta\mathcal{L}_{\texttt{DiFlo}}$ ▷ *Update via optimizer*
end while

Algorithm 2 Inference and DOT Scoring (Euler Integrator)

Require: Condition c, Initial noise $x_0 \sim p_0$, Integration steps N
Ensure: Prediction $\hat{x}_1$, Divergence score S_{DOT}
1. Initialization
$\Delta t \leftarrow 1/N$
$x \leftarrow x_0$
$\mathcal{X}_{traj} \leftarrow [x_0]$ ▷ *Store trajectory states*
2. Generative Flow Integration
for $i = 0$ **to** $N-1$ **do**
$t \leftarrow i/N$
$v \leftarrow v_t^\theta(x, c)$ ▷ *Evaluate marginal vector field*
$x \leftarrow x + v \cdot \Delta t$ ▷ *Euler step*
Append x to $\mathcal{X}_{traj}$
end for
$\hat{x}_1 \leftarrow x$ ▷ *Materialize final prediction*
3. Trajectory Divergence Computation
$S_{DOT} \leftarrow 0$
for $i = 1$ **to** N **do** ▷ *Evaluate N realized steps*
$t \leftarrow i/N$
$\hat{x}_t \leftarrow \mathcal{X}_{traj}[i]$
$x_t^{\text{ref}} \leftarrow (1-t)x_0 + t\hat{x}_1$ ▷ *Compute geometric anchor*
$S_{DOT} \leftarrow S_{DOT} + \frac{1}{D}\|\hat{x}_t - x_t^{\text{ref}}\|_1$ ▷ *Accumulate spatial deviation*
end for
return $\hat{x}_1, S_{DOT}$

J Ablations and Sensitivity Analysis

J.1 Efficacy of Boundary Sampling Strategies

We evaluate DiFlo trained with PGD and three alternative boundary sampling strategies:

1. **Heuristics**: Domain-specific spatial augmentations (random block erasing, elastic deformations, affine shearing, and perspective distortions) intended to create near-manifold perturbations;

Table 9: **Evaluation of detection across numerical ODE solvers.** The detection performance (AUROC), mean velocity, and sample quality (LPIPS) remain highly stable across various integration schemes and step counts on the style transfer task using the PGD mining.

Integrator	Steps	AUROC ↑		Mean Velocity		LPIPS ↓
		FMNIST	KMNIST	On	Off	
Euler	25	0.955	0.859	0.100	0.256	0.2169
Euler	50	0.954	0.855	0.097	0.278	0.2175
Heun2	25	0.956	0.858	0.094	0.269	0.2198
Heun2	50	0.955	0.855	0.093	0.279	0.2196
RK4	25	0.955	0.860	0.093	0.268	0.2202
RK4	50	0.955	0.857	0.096	0.274	0.2197

2. **Masking**: Random and checkerboard masking strategies that destroy local spatial information; and
3. **Outlier Exposure**: Using the Omniglot (Lake et al., 2015) dataset as a source of explicit, fixed negative samples.

Table 10: **Boundary Sampling Ablation.** Comparison of boundary-definition strategies. While domain-specific heuristics naturally excel on known image manifolds, they fail to generalize to complex physical systems (Weather). In the table, we report the detection performance (AUROC↑).

Sampling Method	Temp. Forecast	Style Transfer (FMNIST / KMNIST)
Masking	0.729	0.654 / 0.541
Outlier Exposure	—	0.710 / 0.572
Domain Transforms	0.733	**0.987 / 0.892**
PGD	**0.995**	0.955 / 0.860

The results in Table 10 highlight a fundamental limitation of static sampling, which requires *a priori* knowledge of the manifold boundary. **Heuristic Transforms** perform exceptionally well on the image tasks (outperforming PGD on FMNIST and KMNIST) because spatial shifts perfectly align with the known invariances of character recognition. However, this strategy fails completely on the physical Weather forecasting task. In an atmospheric system, a heuristically transformed temperature field is not simply an "off-manifold" sample; it may represent an entirely different valid state or inherently violate fundamental thermodynamic laws. Without explicit domain knowledge, naive augmentations are unsuitable for defining the boundary of physical validity.

Similarly, **Outlier Exposure** and **Masking** struggle to generalize. Outlier exposure overfits to the explicit geometry of the chosen outlier dataset (Omniglot) and fails to detect semantic shifts (KMNIST) that do not structurally resemble those fixed training samples. Since outlier exposure uses genuine negative data yet underperforms samples placed at the boundary, detection in Diverging Flows is driven by proximity to the boundary rather than by the availability of negative data.

In contrast, **PGD** serves as a versatile, modality-agnostic baseline. By dynamically optimizing for the most sensitive perturbations in the local neighborhood of the input, PGD automatically probes the relevant boundary constraints. While domain-specific heuristics trivially exploit known symmetries when they are available, PGD provides a reliable fallback that performs well across entirely disparate modalities without requiring manual, domain-specific engineering.

J.2 Ablation and Sensitivity Analysis on Synthetic Manifolds

Ablation Study. To isolate the contribution of each objective component, we compare DiFlo against three variants: (i) w/o $\mathcal{L}_{\texttt{repel}}$ ($\lambda = 0$) to test directional divergence alone; (ii) w/o $\mathcal{L}_{\texttt{curve}}$ ($\beta = 0$) to test magnitude penalties alone; and (iii) Random Noise, replacing PGD boundary sampling with $\tilde{c} \sim \mathcal{U}[-1, 1]$ to validate the necessity of a more sophisticated mining strategy.

The ablation study highlights the critical role of PGD. Replacing PGD with random noise samples degrades performance significantly, confirming that samples far from the boundary fail to tighten the

Table 11: **Ablation on Synthetic Manifold Regression.** We isolate the impact of each DiFlo component. Replacing PGD with uniform noise ($\tilde{c} \sim \mathcal{U}$) or removing the kinetic repulsion margin ($L_{\texttt{repel}}$) severely degrades detection. Both structural penalties and targeted boundary sampling are required to minimize leakage (FPR) while maintaining predictive fidelity (MSE).

	Algorithm	AUROC ↑	FPR(%) ↓	MSE ↓
Reg.	DiFlo w/o $\mathcal{L}_{\texttt{repel}}$	0.711	20.45	$8.51 \cdot 10^{-5}$
	DiFlo w/o $\mathcal{L}_{\texttt{curve}}$	0.931	10.69	$8.48 \cdot 10^{-5}$
	DiFlo w/ $\tilde{c} \sim \mathcal{U}$	0.662	90.45	$8.50 \cdot 10^{-5}$

vector field around the manifold support. Furthermore, we observe that the repulsion term $\mathcal{L}_{\texttt{repel}}$ is the primary driver of the detection signal; removing it causes a sharper drop in performance compared to removing the curvature penalty $\mathcal{L}_{\texttt{curve}}$, though both are required for optimal separation.

Sensitivity Analysis. To rigorously evaluate the stability of Diverging Flows, we conduct a comprehensive hyperparameter sensitivity analysis. We ablate the PGD adversarial radius ϵ, the number of PGD steps $\mathcal{K}$, and the contrastive margins (m_r, m_c) on the 2D Spiral dataset across both the Conditional Generation and Probabilistic Regression tasks. The robust operating ranges and their corresponding performance metrics are summarized in Table 12.

Table 12: **Sensitivity Analysis on the 2D Spiral Manifold.** We report the robust ranges for PGD and contrastive hyperparameters. Diverging Flows exhibits stability across a wide spectrum of configurations, maintaining near-perfect detection without degrading baseline predictive accuracy (MSE).

Hyperparameter	Tested Range	AUROC ↑	FPR (%) ↓	MSE (10^{-5}) ↓
		Conditional Generation Task		
Adversarial Radius ϵ	0.01 (Fails)	~ 0.493	95.5	—
Adversarial Radius ϵ	[0.05, 0.50]	> 0.980	< 4.9	—
PGD Steps $\mathcal{K}$	$\{1, 5\}$	> 0.980	< 4.8	—
Repulsion Margin m_r	[0.30, 2.00]	> 0.971	< 5.5	—
Curvature Margin m_c	[0.01, 1.70]	> 0.979	< 5.2	—
		Probabilistic Regression Task		
Adversarial Radius ϵ	[0.01, 0.50]	> 0.983	< 4.8	≤ 8.52
PGD Steps $\mathcal{K}$	$\{1, 5\}$	> 0.996	< 2.7	≤ 8.50
Repulsion Margin m_r	[0.50, 2.00]	> 0.993	< 2.1	≤ 8.53
Curvature Margin m_c	[0.10, 1.70]	> 0.996	< 3.7	≤ 8.53

Task-Dependent Geometric Sensitivity. The optimal PGD projection radius ϵ depends directly on the intrinsic thickness (variance) of the underlying data manifold. In the Generation task, the spiral dataset exhibits higher spatial variance. Consequently, a microscopic perturbation ($\epsilon = 0.01$) is insufficient to escape this wider distribution; the perturbed samples physically overlap with valid data, causing the detection mechanism to fail (AUROC ~ 0.49). Conversely, the dataset utilized for the Regression task possesses much lower variance, yielding a significantly "thinner" manifold. Because of this tighter distribution, that exact same $\epsilon = 0.01$ shift easily pushes samples completely off-manifold, resulting in near-perfect separation (AUROC > 0.98). In practice, establishing ϵ based on the natural scale of the normalized domain (e.g., defaulting to $\epsilon = 0.10$) safely clears the boundary across varying manifold thicknesses.

Computational Efficiency and Margin Stability. Notably, a single-step PGD attack ($\mathcal{K} = 1$) performs almost identically to a 5-step attack across both tasks. This suggests that the gradient of the standard Flow Matching loss, when the model is trained with the contrastive objectives, is a strong local signal, allowing Diverging Flows to identify boundary samples with minimal computational training overhead. Furthermore, as long as a positive contrastive margin is enforced, the exact values of m_r and m_c do not require fragile tuning. The detection mechanism remains highly stable (maintaining AUROC > 0.97) across broadly swept margin ranges without any major degradation of the baseline MSE.

J.3 Sensitivity Analysis on High Dimensional Weather Temperature Forecasting

Dimensionality and Geometric Sensitivity. The sensitivity analysis on the high-dimensional ERA5 dataset (Table 13) reveals distinct geometric behaviors compared to the low-dimensional synthetic manifolds. First, we observe that the repulsion margin (m_r) must scale significantly. Small margins ($m_r \leq 10$) fail to separate from the natural transport energy variance of a 64×64 system, reducing detection to near-random performance. Reliable separation requires a large margin ($m_r \geq 100$), consistent with the $\sqrt{d}$ scaling of Appendix D.4, so the margin scale is the main quantity to adapt when moving to high dimension.

Conversely, directional divergence requires a soft penalty. Imposing aggressive curvature bounds ($\beta = 5$ or $m_c = 1.8$) degrades the AUROC signal (~ 0.75), whereas moderate angular regularization keeps detection high (AUROC 0.995 for the base values $\beta = 0.9$, $m_c = 0.9$, and 0.992 and 0.977 for $\beta = 0.1$ and $m_c = 0.3$). In high-dimensional ODE integration, minor initial angular deviations cascade into massive spatial divergence; over-penalizing this geometry likely disrupts the local Lipschitz continuity required for stable solving. Finally, across all configurations, the predictive fidelity remains remarkably stable, confirming that DiFlo's dynamic boundaries govern off-manifold behavior without bleeding into the valid data distribution.

Table 13: **Weather Temperature Forecasting Sensitivity Analysis.** Evaluation of DiFlo's hyperparameters on the high-dimensional ERA5 dataset. Each block changes one hyperparameter from the base configuration of Table 18 ($\lambda = 0.7$, $m_r = 100$, $\beta = 0.9$, $m_c = 0.9$, $\epsilon = 0.1$), which is the model of Table 2 (mean of three seeds; sweep rows are single runs). The sweep identifies the operating range in high dimension: a large kinetic margin ($m_r \geq 100$), a moderate adversarial radius ($\epsilon = 0.1$), and a directional penalty that is not too strong ($\beta \leq 0.9$, $m_c \leq 0.9$). Fidelity is stable across all configurations.

Hyperparameter	MSE (10^{-4}) ↓	SSIM ↑	PSNR ↑	AUROC ↑
Base configuration	8.2	0.967	36.36	0.995
Repulsion Weight				
$\lambda = 0.1$	7.7	0.967	37.11	0.991
$\lambda = 1.5$	7.6	0.968	37.18	0.913
$\lambda = 5.0$	7.4	0.968	37.31	0.941
Repulsion Margin				
$m_\mathrm{r} = 1$	7.3	0.969	37.33	0.718
$m_\mathrm{r} = 10$	7.5	0.969	37.25	0.668
$m_\mathrm{r} = 200$	7.4	0.968	37.27	0.949
Curvature Weight				
$\beta = 0.1$	7.7	0.967	37.11	0.992
$\beta = 1.5$	7.8	0.968	37.06	0.898
$\beta = 5.0$	8.8	0.966	36.93	0.765
Curvature Margin				
$m_\mathrm{c} = 0.3$	7.3	0.969	37.34	0.977
$m_\mathrm{c} = 1.8$	7.7	0.966	37.11	0.752
Adversarial Radius				
PGD $\epsilon = 0.05$	8.0	0.965	36.96	0.914
PGD $\epsilon = 0.8$	7.0	0.969	37.37	0.640

Practical Guideline for ϵ Selection. The adversarial radius ϵ governs the distance between valid conditions and boundary samples in the normalized input space. We observe a consistent principle across all tasks: ϵ should be set to the natural noise scale of the normalized domain. Concretely, since all inputs are normalized to $[-1, 1]^d$, we recommend an $\epsilon \in [0.07, 0.2]$ as a default. This range safely clears the manifold boundary across varying manifold thicknesses (as demonstrated in Table 12 for synthetic data) while remaining in the local neighborhood required for PGD to find informative boundary samples. The failure of $\epsilon = 0.8$ on ERA5 is instructive: a perturbation of 0.8 in the normalized space corresponds to a shift of 40% of the full temperature range of the data,

which is physically implausible, which pushes samples so far off-manifold that they provide no useful boundary signal. This is analogous to using too large a learning rate; the optimization overshoots the relevant region.

J.4 Scaling to Higher-Resolution ERA5

ERA5 (Section 5.2) is a large-scale, global reanalysis product, and the 64×64 crop used in our main experiments is only one of many resolutions at which it is standardly used. To test whether DiFlo's detection capability holds as resolution increases, we retrain on ERA5 at 128×128 ($4\times$ the pixel count of our main experiments) using the same U-Net family and ϵ-normalization scheme, without retuning any hyperparameter from the 64×64 configuration in Table 13.

Table 14: **Scaling to 128×128 ERA5.** Detection quality is preserved when quadrupling the pixel count with no hyperparameter retuning, indicating DiFlo is not specifically tuned to the 64×64 resolution used in the main experiments.

Resolution	MSE ↓	SSIM ↑	PSNR ↑	AUROC ↑
64×64 (main)	$8.22 \cdot 10^{-4}$	0.967	36.36	0.995
128×128	$4 \cdot 10^{-3}$	0.9516	33.6	0.974

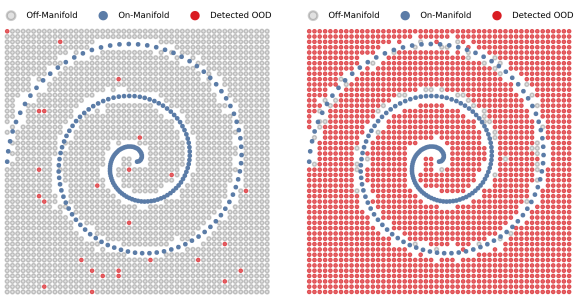
AUROC drops only modestly, from 0.995 to 0.974, despite the substantially harder generative task at higher resolution (reflected in the higher MSE and lower PSNR/SSIM). We did not train a standard FM baseline at this resolution, so this experiment provides evidence on detection only. It indicates that the margins, which act on the transport geometry of the vector field (Eq. 5), do not depend on a specific resolution.

K Visualization of Detection Landscapes

In this section, we visualize the decision boundaries learned by the models within the 2D ambient conditioning space. Figures 9 and 10 depict the landscape of accepted versus rejected conditions.

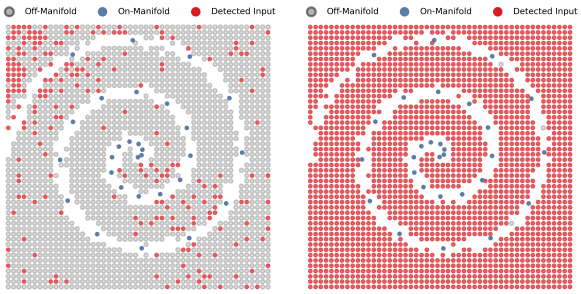
In these plots, the underlying data manifold is the 2D spiral. The background represents the full ambient space $[-1, 1]^2$. We evaluate the model on a dense grid of conditions; regions highlighted in **red** indicate conditions flagged as extrapolations (i.e., inputs that trigger a high S_{DOT} score).

The results visually confirm the silent failure hypothesis: Standard Flow Matching (left) accepts almost the entire ambient space as valid, driven by its inherent smoothness bias. In contrast, DiFlo (right) effectively creates a tight validity boundary that follows the data support, while rejecting the surrounding off-manifold regions (red).



(a) Flow Matching (b) Diverging Flows (Ours)

Figure 9: **Detection Landscape:** Conditional Generation. The standard model (a) fails to distinguish the manifold from the background. Diverging Flows (b) successfully identifies the ambient space as invalid (red), accepting only the spiral support.



(a) Flow Matching (b) Diverging Flows (Ours)

Figure 10: **Detection Landscape:** Probabilistic Regression. Even in the regression task, where the condition represents a dynamic state x_t, Diverging Flows (b) maintains a precise decision boundary around the valid trajectory path.

L Comparison with Passive Trajectory Detectors

L.1 DiffPath

For completeness, we provide a reference comparison against DiffPath (Heng et al., 2024), a method that analyzes reverse generative trajectories for OOD detection. Unlike Diverging Flows, which actively regularizes the underlying vector field during training, DiffPath operates as a passive, post-hoc detector on a frozen model.

Because DiffPath requires a pre-trained generative model, and no off-the-shelf models exist for our specific scientific datasets, we trained an unconditional diffusion model from scratch on the valid conditions of each task, i.e., on the inputs whose validity is tested, since pretrained models do not transfer to our data. DiffPath utilizes this model to perform the reverse generation process on a hold-out set of valid calibration samples. It extracts various discrete trajectory statistics and fits a statistical density estimator (e.g., Gaussian Mixture Model) to these features. At inference time, the validity of a query input is determined by the likelihood of its trajectory statistics under this fitted distribution.

Table 15 presents the comparative detection performance. DiffPath-6D is on par with or better than Diverging Flows on all benchmarks, most notably on KMNIST. The two approaches differ in what they require. For each query, DiffPath runs 20 DDIM passes through a second diffusion model, which is trained separately on the valid conditions, and evaluates a density model fitted on the resulting trajectory statistics. This comes in addition to the predictive model itself. Diverging Flows requires no model besides the predictor and computes its score from the trajectory that the prediction already produces, at $\sim 0.8\%$ overhead. The contribution of Diverging Flows is therefore not higher detection accuracy than dedicated post-hoc detectors, but comparable detection within a single predictive model.

Table 15: **Comparison with DiffPath.** We report detection performance (AUROC) against DiffPath, a post-hoc trajectory auditor that requires a second diffusion model. DiFlo achieves comparable detection within the predictive model itself.

Method	Temp. Forecast ↑	Style Transfer (FMNIST / KMNIST) ↑
Passive Post-Hoc Statistics		
DiffPath-1D	0.973	0.981 / 0.863
DiffPath-6D	0.993	0.991 / 0.929
Active Generative Regularization		
DiFlo(Transforms)	0.733	0.987 / 0.892
DiFlo(PGD)	0.995	0.955 / 0.860

Remark on Flow vs. Diffusion Dynamics. DiffPath was designed for diffusion models, whose probability-flow trajectories are strongly curved, so that trajectory statistics of a frozen model can

differ between in- and out-of-distribution inputs. Standard FM trajectories are much closer to straight lines, and in our settings passive trajectory statistics of a frozen FM model were insufficient: the FM-DOT baselines in Tables 1–3 and the FM rows of Table 5 remain far below Diverging Flows, and so does the acceleration proxy of concurrent work (Appendix L.2). This motivates regularizing the vector field during training rather than auditing it afterwards.

L.2 Concurrent Work: Acceleration Proxy of Standard FM

Concurrent work (Rao et al., 2026) proposes the denoising acceleration of a standard FM model as a training-free uncertainty proxy and applies it to failure detection in vision-language-action policies. With velocities v_{t_i} of the sampling trajectory and a prefix of p steps, the score is

$$\text{accel}_p = p \, \frac{\sum_{i=1}^{p-1} \|v_{t_i} - v_{t_{i-1}}\|_2}{\sum_{i=0}^{p-1} \|v_{t_i}\|_2}, \tag{24}$$

computed from the velocities that the solver already produces. We evaluate it on the ERA5 trajectories of Appendix D.6 (three seeds, 50 Euler steps) over the full path ($p = 50$) and over the first half ($p = 25$), which the authors report as the most informative range, together with their straightness baseline (the ratio of chord to arc length) and the DOT score.

Table 16: **Passive trajectory statistics on ERA5 hotspots.** AUROC, mean and standard deviation over three seeds, for scores computed on the trajectories of standard FM and of DiFlo.

Score	Standard FM	DiFlo
Straightness	0.628 ±0.037	0.978 ±0.023
Acceleration ($p = 25$)	0.571 ±0.053	0.985 ±0.020
Acceleration ($p = 50$)	0.594 ±0.039	0.966 ±0.031
DOT	0.667 ±0.023	**0.995 ±0.005**

Table 16 shows that on standard FM every passive statistic stays between 0.57 and 0.67. Even the best prefix, chosen with knowledge of the hotspot labels, reaches only 0.65, so the weak signal is not a matter of tuning. This is the silent extrapolation of Proposition 4.1: for off-manifold conditions, standard FM trajectories do not bend more than valid ones. On Diverging Flows, the same acceleration proxy reaches 0.97 to 0.99, since the training objective creates the bending that any curvature statistic can read. The two approaches are therefore complementary: Diverging Flows shapes the vector field, and passive statistics, including the acceleration proxy, can serve as readouts of it. DOT remains the most accurate and the most stable score across seeds. These results concern off-manifold conditions on ERA5 and do not evaluate the proxy on the failure detection task for which it was proposed.

M Experiment Details

M.1 Theoretical Formulation of Detection Baselines

We compare Diverging Flows against two intrinsic baselines: (i) a likelihood proxy derived from the learned vector field, and (ii) ensemble predictive variance.

M.1.1 Flow-Matching Likelihood

A natural extrapolation signal is whether the model assigns low marginal probability to the queried condition. Although Flow Matching optimizes a transport-based objective rather than a direct likelihood bound, the learned vector field v_t^θ governs a probability flow ODE that induces a tractable density over the state space.

Let the state x_t evolve according to $\frac{dx_t}{dt} = v_t^\theta(x_t, c)$. By the continuity equation, the log-density evolves along the trajectory as (Lipman et al., 2024):

$$\frac{d}{dt} \log p_t(x_t|c) = -\nabla_x \cdot v_t^\theta(x_t, c) \tag{25}$$

Since the divergence of the vector field equals the trace of its Jacobian, integrating from $t = 0$ to $t = 1$ yields the implied log-likelihood:

$$\log p_1(x_1|c) = \log p_0(x_0) - \int_0^1 \mathrm{Tr}\left(\nabla_x v_t^\theta(x_t, c)\right) dt \tag{26}$$

Computing the exact Jacobian trace requires $\mathcal{O}(d)$ backward passes, so we instead approximate it with the Hutchinson estimator. Drawing a probe vector $z \sim \mathcal{N}(0, I)$ gives an unbiased stochastic estimate:

$$\log p_1(x_1|c) \approx \log p_0(x_0) - \int_0^1 z^\top \nabla_x v_t^\theta(x_t, c)\, z\, dt \tag{27}$$

The Jacobian-vector product $z^\top \nabla_x v_t^\theta$ is computed via reverse-mode automatic differentiation, integrated jointly with the state ODE. The extrapolation score is then defined as the negative proxy log-density, $S_{\text{div}}(c) = -\log p_1(x_1|c)$, where higher values indicate low compatibility under the model's learned marginal dynamics.

M.1.2 Ensemble Predictive Variance

We additionally evaluate Deep Ensembles as a standard epistemic uncertainty baseline. We independently train $K_{\text{ens}} = 3$ Flow Matching models $\{\theta_k\}_{k=1}^{K_{\text{ens}}}$ under identical hyperparameters but distinct random initializations.

To decouple epistemic uncertainty from aleatoric stochasticity, we evaluate each model's deterministic conditional expectation $\hat{x}_1^{(k)}$ under condition c, and define the extrapolation score as the empirical spatial variance across predictions:

$$S_{\text{Ens}}(c) = \frac{1}{K_{\text{ens}}} \sum_{k=1}^{K_{\text{ens}}} \|\hat{x}_1^{(k)} - \mu_c\|_2^2, \quad \mu_c = \frac{1}{K_{\text{ens}}} \sum_{k=1}^{K_{\text{ens}}} \hat{x}_1^{(k)} \tag{28}$$

For valid conditions, the independently trained models converge to consistent predictive flows, yielding low variance. For off-manifold queries, the models disagree, inflating S_{Ens}. The critical limitation of this approach is computational: detection requires K_{ens} independent ODE solves at inference time, imposing an $\mathcal{O}(K_{\text{ens}})$ overhead that is prohibitive in real-time deployments.

M.2 General Experimental Setup

Across all evaluated tasks, experiments were conducted using a single NVIDIA RTX 4090 GPU. During optimization, we applied a learning rate schedule consisting of a 500-step linear warmup from 10^{-8} to the base learning rate, followed by a cosine decay. No off-manifold evaluation data (hotspot conditions, FMNIST, KMNIST, or the far off-manifold sets) was used to select hyperparameters; in pixel space, it is not known in advance which perturbations matter. In all detection experiments, the score of each condition is computed from a single sampling trajectory, i.e., one draw of $x_0 \sim p_0$ per condition. Prior to training, all continuous input data across tasks was normalized to the $[-1, 1]$ range to ensure stable vector field regression and consistent application of the L_∞ adversarial bounds. To ensure statistically rigorous evaluation and properly calibrated detection boundaries, all datasets were partitioned using a fixed, unified split ratio: 70% utilized for model training, 10% reserved as a hold-out calibration set for Split Conformal Prediction, and 20% utilized for final testing, if a test set is not provided.

Evaluation Metrics. For generative tasks, the Fréchet Inception Distance (Heusel et al., 2017) compares the Inception feature statistics of real (μ_r, Σ_r) and generated (μ_g, Σ_g) samples as $\text{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$.

M.3 Synthetic Manifold Experiment Details

Data Generation. We utilize the 2D spiral and ambient off-manifold setup as formally defined in Section 5.1. We enforce a distance of 0.025 between the manifold support and the off-manifold sampling region to ensure the detection task is unambiguous.

Model Architecture. We employ a time-conditioned Multi-Layer Perceptron (MLP) that processes inputs as flat vectors. Specifically, the **input** is a direct concatenation of the noisy state ($x_t \in \mathbb{R}^2$ in the conditional generation task or $x_t \in \mathbb{R}^{20}$ in the regression task), the condition $c \in \mathbb{R}^2$, and the time scalar $t \in [0, 1]$. The **backbone** consists of three hidden layers of width 512 using SiLU activations, followed by a linear **output** projection to the dimension of x_t ($\mathbb{R}^2$ or $\mathbb{R}^{20}$) predicting the continuous velocity field u_t.

Training Setup. The model is trained using the AdamW optimizer for 400,000 training steps in the generation task and 200,000 training steps in the regression task. Detailed hyperparameters for the architecture and the contrastive loss functions are listed in Table 17.

Table 17: **Hyperparameters for Synthetic Manifold Experiments.**

Hyperparameter	Conditional Generation	Probabilistic Regression
Hidden Layers	3	3
Hidden Dimension	512	512
Activation	SiLU	SiLU
Optimizer	AdamW	AdamW
Learning Rate	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$
Weight Decay	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$
Batch Size	256	256
Training Steps	400,000	200,000
PGD Steps ($\mathcal{K}$)	3	3
Perturbation (ϵ)	0.1	0.1
Repulsion Weight (λ)	0.1	0.1
Curvature Weight (β)	0.05	0.1
Margins (m_r / m_c)	1.0 / 0.9	1.0 / 0.9
Integrator	Euler	Euler
Integration Steps	50	50
Significance Level (α)	0.05	0.05
Target Coverage	95%	95%

Note on the Spiral Benchmark. The 2D spiral is a canonical benchmark in the study of Neural ODEs and continuous-time dynamical systems (Chen et al., 2018). Unlike simple shapes, a spiral represents a stiff geometric structure where the optimal transport path (a straight line) often intersects regions that are off-manifold. In the context of probabilistic regression, this experiment is critical because it validates that the learned vector field captures the true non-linear differential equation governing the system's evolution, rather than memorizing discrete point-to-point mappings. Successfully modeling this dynamics while simultaneously detecting off-manifold perturbations demonstrates that Diverging Flows can enforce safety constraints without over-simplifying the complex curvature of the underlying data manifold.

M.4 Weather Temperature Forecasting

Dataset. We demonstrate the applicability of Diverging Flows to climate science via two-meter surface temperature forecasting using the ERA5 reanalysis dataset (Hersbach et al., 2020). To formulate the predictive task, we construct conditional pairs (c, x_1), where the condition c represents the historical temperature map at time t and the target x_1 is the realized state at $t + 6$ hours. The data is sampled at standard six-hour synoptic intervals (00, 06, 12, 18 UTC). Physically, the global temperature limits (spanning roughly 210K to 313K) are normalized to the $[-1, 1]$ interval. Following Chan et al. (2024), off-manifold hotspot conditions are created by setting a historically cold region of a valid condition to the maximum normalized value of 1 (about 313K). For comparison, the PGD radius $\epsilon = 0.1$ used during training corresponds to about 5K per pixel, so the hotspots lie far outside the perturbations seen during training, and detecting them relies on the generalization of the learned barrier.

Model Architecture. We employ a U-Net backbone modified for conditional generation. For **time conditioning**, the time step $t \in [0, 1]$ is embedded using Fourier features followed by an MLP to produce a 128-dimensional embedding, which is injected into every residual block via FiLM layers. The **condition encoder** processes the conditioning input c (current state) through a separate learnable

network consisting of a convolution, two residual blocks, and one downsampling operation. Finally, for **cross-attention injection**, the latent representation from the condition encoder is projected to match the time embedding dimension and injected into the U-Net backbone via cross-attention layers at the 16×16 resolution (lowest level).

Training Setup. We train the model for 50,000 training steps with a batch size of 64. To ensure stability, we use gradient clipping (norm 1.0) and an Exponential Moving Average (EMA) of model weights with decay 0.9999.

Table 18: **Hyperparameters used in the Weather Temperature Forecasting (ERA5) experiment.**

Hyperparameter	Value
Input Resolution	64×64
Channels	[64, 64, 128]
ResBlocks per Scale	2
Attention Resolution	16×16
Conditioning	Cross-Attention
Time Embedding Dim	128
Optimizer	AdamW
Learning Rate	$3 \cdot 10^{-4}$
Weight Decay	$1 \cdot 10^{-2}$
AdamW Betas	[0.9, 0.95]
Batch Size	64
Training Steps	50,000
EMA Decay	0.9999
Gradient Clip	1.0
PGD Steps ($\mathcal{K}$)	5
Perturbation (ϵ)	0.1
Repulsion Weight (λ)	0.7
Curvature Weight (β)	0.9
Margin m_r	100.0
Margin m_c	0.9
Integrator	Euler
Integration Steps (N)	50
Significance Level (α)	0.05
Target Coverage	95%

M.5 Cross-Domain Style Transfer Details

Dataset. We perform a cross-domain translation task mapping grayscale MNIST digits ($1 \times 32 \times 32$) to RGB SVHN digits ($3 \times 32 \times 32$).

Model Architecture. We use a U-Net backbone similar to the weather temperature forecasting experiment, but with a stronger dual-conditioning mechanism to preserve high-frequency structural details. For **input concatenation**, we explicitly concatenate the 1-channel conditioning image c with the 3-channel noisy state x_t at the model input, providing strong spatial guidance for the digit structure. Additionally, c is processed by a shallow condition encoder (1 ResBlock) and injected via **cross-attention** at the 16×16 and 8×8 resolutions. The **backbone** utilizes channel widths of [64, 64, 128] with attention layers at the two lower resolutions.

Training & Inference. The model is trained for 500,000 training steps using AdamW with a learning rate decay schedule. For inference, we perform $N = 50$ steps.

N Additional Qualitative Results

In this section, we provide weather temperature forecasting (Figure 12) and style transfer (Figure 11) samples from Diverging Flows.

Table 19: **Hyperparameters used in the Style Transfer (MNIST $\rightarrow$ SVHN) experiment.**

Hyperparameter	Value
Input Resolution	32×32
Channels	[64, 64, 128]
ResBlocks per Scale	2
Attention Resolutions	$16 \times 16, 8 \times 8$
Conditioning	Concat + Cross-Attention
Optimizer	AdamW
Learning Rate	$1 \cdot 10^{-4}$
Weight Decay	$1 \cdot 10^{-3}$
AdamW Betas	[0.9, 0.95]
Accumulation Steps	4
Training Steps	500,000
Batch Size	128
EMA Decay	0.9999
PGD Steps ($\mathcal{K}$)	5
Perturbation (ϵ)	0.2
Repulsion Weight (λ)	0.1
Curvature Weight (β)	0.7
Margin m_r	10.0
Margin m_c	1.3
Integrator	Euler
Integration Steps (N)	50
Significance Level (α)	0.05
Target Coverage	95%



Figure 11: **Extended Style Transfer Samples.** Pairs of Conditioning Inputs (MNIST) and Generated Outputs (SVHN). The model successfully maps the sparse structural guidance of the grayscale digit to the rich, multi-modal distribution of street view imagery.

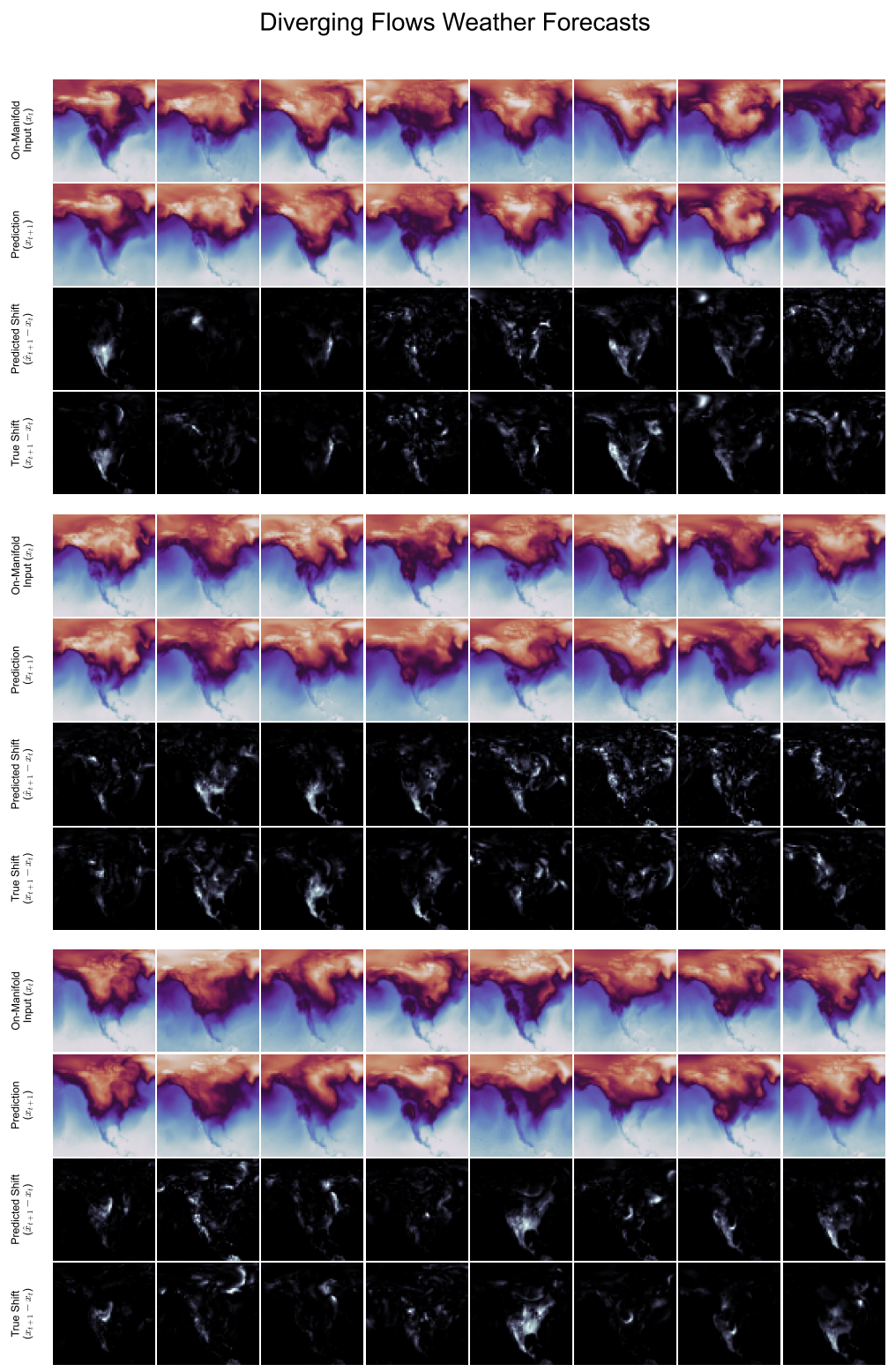


Figure 12: **Extended Weather Temperature Forecasting Samples.** Randomly selected 6-hour temperature forecasts generated by Diverging Flows on the ERA5 validation set. The model generates physically consistent heatmaps that align with the ground truth dynamics.