FORECASTS ARE PATHS: CURVATURE-MATCHED HYVÄRINEN RESIDUALS AND OVERLAP-CONSTRAINED ORDINAL FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon forecasters produce trajectories, yet they are usually trained as if the horizon were a bag of independent coordinates. Pointwise squared error measures amplitude, but it neither controls how forecast errors evolve nor distinguishes trajectories whose local peaks, valleys, and turning order differ. We introduce HYOT, a single-pass loss that scores a forecast in two complementary geometries. In the cardinal geometry, a curvature-matched Gaussian–Student Hyvärinen score acts on both residual levels and residual innovations, combining the local efficiency of squared error with controlled tail influence. In the ordinal geometry, each five-point window induces a differentiable law over all $5! = 120$ rankings. Because adjacent windows share four observations, every hard transition has an exact source–insertion representation over a 120×5 transition alphabet. A factorized differentiable soft descriptor and entropy-gated Hellinger score align the predicted and observed ordinal flows. The resulting objective directly optimizes trajectory geometry during training while preserving the original inference procedure. Experiments across five forecasting architectures, six datasets, and four horizons show that optimizing the path, rather than isolated points, improves aggregate MSE and MAE. A representative trajectory-structure evaluation further shows that the MSE gain can be accompanied by lower Innovation MAE and Ordinal Hellinger and higher Local Rank Agreement.

1 INTRODUCTION

Long-horizon forecasting has advanced through increasingly specialized representations. DLinear separates trend and seasonal structure (Zeng et al., 2023), while PatchTST organizes history into temporal patches (Nie et al., 2023). TimesNet models temporal variation in a two-dimensional space (Wu et al., 2023); TimeMixer develops multiscale mixing (Wang et al., 2024); TimeBridge couples local dependency modeling with long-term cointegration under non-stationarity (Liu et al., 2025); and iTransformer inverts the conventional tokenization of multivariate series (Liu et al., 2024). Despite these architectural differences, training is still dominated by coordinate-wise squared error. This objective treats the horizon as a collection of targets, whereas the predicted object is a *path*. Such a path has cardinal geometry—numerical deviation and its temporal propagation—and ordinal geometry—local ordering structure and its evolution. Forecasts with similar MSE can differ sharply in both geometries.

Path-aware objectives have approached this structure from complementary directions. Soft-DTW introduces differentiable sequence alignment (Cuturi & Blondel, 2017). DILATE separates shape and temporal distortion (Le Guen & Thome, 2019), while patch-wise structural supervision compares local forecast segments (Kudrat et al., 2025). We focus on the geometry of forecasting residuals. The Hyvärinen rule scores a distribution through local derivatives of its log density (Hyvärinen, 2005), with proper local scoring theory providing its statistical foundation (Parry et al., 2012). Building on this rule, the cardinal branch of HYOT combines Gaussian and Student-induced geometries through exact curvature matching. A closed-form admissible weight range preserves the MSE descent orientation in every scalar residual coordinate, and applying the score to both residual levels and residual innovations captures numerical error together with its short-range evolution.

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.

The ordinal branch models how local path configurations evolve across overlapping windows. Permutation patterns provide a compact description of temporal order (Bandt & Pompe, 2002), and ordinal partition networks demonstrate their value for dynamical structure (McCullough et al., 2015). We make this representation differentiable using the choice model of Luce (Luce, 2005) and its Plackett–Luce ranking formulation (Plackett, 1975). Each five-point window induces a distribution over $5! = 120$ rankings. Because adjacent windows share four observations, every hard transition has an exact source–insertion representation over a 120×5 ordinal-flow alphabet. A factorized differentiable soft descriptor, Hellinger score, and target ordinal-concentration gate align this flow, and the cardinal and ordinal branches jointly optimize the same forecast trajectory. Our contributions are fourfold:

- We derive a curvature-matched Gaussian–Student Hyvärinen score with local MSE geometry and a closed-form residual-wise direction condition.
- We construct a dual residual field over levels and innovations, with reference-scale adjustment and a normalization that exactly preserves aggregate trace curvature.
- We introduce an overlap-constrained ordinal flow with an exact 120×5 hard-transition alphabet and a factorized differentiable soft descriptor.
- We unite cardinal residual geometry and ordinal transition dynamics in an architecture-agnostic objective that directly optimizes forecast trajectories.

2 Related Work

Long-horizon forecasting has progressed through architectural biases for temporal structure. DLinear separates trend and seasonal components (Zeng et al., 2023), PatchTST learns from temporal patches (Nie et al., 2023), and TimesNet represents temporal variation in a two-dimensional space (Wu et al., 2023). More recent designs develop multiscale mixing (Wang et al., 2024) or invert the tokenization of multivariate series (Liu et al., 2024). A complementary line improves the training objective itself. Soft-DTW makes sequence alignment differentiable (Cuturi & Blondel, 2017); DILATE decomposes shape and temporal distortion (Le Guen & Thome, 2019); and patch-wise supervision measures local structural agreement (Kudrat et al., 2025). HyOT follows this path-level view while organizing supervision in two intrinsic coordinates: cardinal residual geometry and ordinal transition dynamics.

The cardinal construction is grounded in local proper scoring. The Hyvärinen rule evaluates a density through derivatives of its log density (Hyvärinen, 2005), and proper local scoring theory characterizes the statistical basis of this family (Parry et al., 2012). The ordinal construction connects to permutation patterns, which encode local temporal order without requiring absolute amplitudes (Bandt & Pompe, 2002), and to ordinal partition networks that describe transitions among such patterns (McCullough et al., 2015). Differentiability is supplied by the choice model of Luce (Luce, 2005) and its Plackett–Luce ranking law (Plackett, 1975). HyOT brings these ideas into a single forecasting objective: curvature matching couples Gaussian and Student-induced residual scores, while the four-point overlap of successive five-point windows yields an exact source–insertion hard-transition alphabet and a factorized differentiable soft descriptor.

3 Method

3.1 Forecast paths and design overview

Let L, H, and C denote the input length, forecast horizon, and number of variables. We consider $H \geq 6$, which provides at least two overlapping five-point windows and one adjacent ordinal transition. Given a history $\boldsymbol{X} \in \mathbb{R}^{L \times C}$ and future target $\boldsymbol{Y} \in \mathbb{R}^{H \times C}$, a forecasting backbone produces

$$\widehat{\boldsymbol{Y}} = f_\theta(\boldsymbol{X}), \qquad \widehat{\boldsymbol{Y}} \in \mathbb{R}^{H \times C}. \tag{1}$$

Standard MSE treats the HC outputs as pointwise coordinates and aggregates their squared deviations. This is effective for amplitude accuracy, but it does not specify how the errors are distributed along the horizon or whether the predicted path preserves local rises, falls, and reversals. Two forecasts can therefore have similar MSE while exhibiting substantially different temporal organization. HyOT

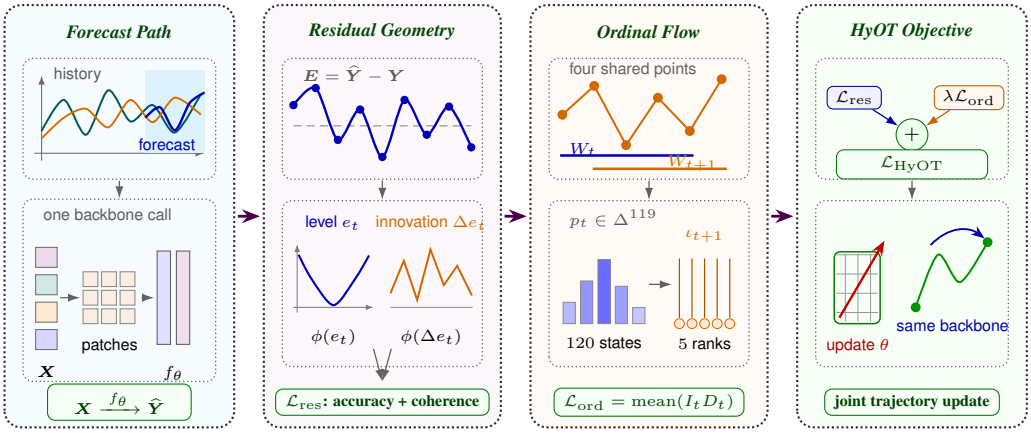


Figure 1: HyOT architecture with residual and ordinal-flow branches.

instead scores the forecast as a path through two complementary geometries. Its cardinal branch acts on the residual field and its first temporal difference,

$$e_{t,c} = \hat{y}_{t,c} - y_{t,c}, \qquad \Delta e_{t,c} = e_{t,c} - e_{t-1,c}, \quad 2 \le t \le H, \tag{2}$$

while its ordinal branch compares overlapping windows $W_{t,c} = (y_{t,c}, \dots, y_{t+4,c})$ and $\widehat{W}_{t,c} = (\hat{y}_{t,c}, \dots, \hat{y}_{t+4,c})$. The cardinal geometry asks how far the prediction is from the target and how that error evolves; the ordinal geometry asks whether the predicted trajectory preserves local ordering and turning structure as the window advances. Thus the two branches retain, respectively, metric information and relational shape information. HyOT is therefore not an alternative forecasting architecture, but a trajectory-level supervision mechanism imposed on the output of an arbitrary forecasting backbone. Both branches are computed from one forecast and jointly update the original backbone, as shown in Figure 1.

3.2 Curvature-matched cardinal residuals

For a twice differentiable strictly positive probability density $q : \mathbb{R} \to \mathbb{R}_{>0}$ and a scalar residual $e \in \mathbb{R}$, the Hyvärinen score is

$$S_{\text{H}}(q, e) = \frac{\partial^2}{\partial e^2} \log q(e) + \frac{1}{2} \left(\frac{\partial}{\partial e} \log q(e) \right)^2 . \tag{3}$$

It is a proper local scoring rule and is invariant to the normalization constant of q (Hyvärinen, 2005; Parry et al., 2012). For a Gaussian location family, equation 3 is equivalent, up to positive scaling and an additive constant, to e^2. This makes MSE the Gaussian member of the construction.

The Gaussian member supplies a coercive tail and the optimization geometry expected by standard forecasting backbones. To control the effect of rare residuals, we pair its score with the score induced by the variance-parameterized Student density

$$q_{\text{St}}(e; \sigma, \nu) \propto \left(1 + \frac{e^2}{a} \right)^{-(\nu+1)/2}, \qquad a = (\nu - 2)\sigma^2, \qquad \sigma > 0, \quad \nu > 2, \tag{4}$$

where $\sigma > 0$ specifies a fixed reference residual scale, $\nu > 2$ controls the tail geometry, and $a = (\nu - 2)\sigma^2$ parameterizes the Student reference distribution to have variance σ^2. The Student score and its curvature factor are

$$H_{\text{St}}(e; \sigma, \nu) = \frac{(\nu + 1) \left[\frac{1}{2}(\nu + 3)e^2 - a \right]}{(a + e^2)^2}, \qquad \kappa_{\nu,\sigma} = \frac{2a^2}{(\nu + 1)(\nu + 7)}. \tag{5}$$

We combine the coercive Gaussian member with the Student correction through

$$\phi_{\sigma,\nu,w}(e) = \frac{e^2 + w\kappa_{\nu,\sigma} H_{\text{St}}(e; \sigma, \nu)}{1 + w}, \qquad w \ge 0. \tag{6}$$

Proposition 1 (Local MSE geometry). *For every $\nu > 2$, $\sigma > 0$, and $w \geq 0$, the score in equation 6 satisfies*

$$\phi'_{\sigma,\nu,w}(0) = 0, \qquad \phi''_{\sigma,\nu,w}(0) = 2. \tag{7}$$

Equivalently, by evenness, $\phi_{\sigma,\nu,w}(e) = \phi_{\sigma,\nu,w}(0) + e^2 + \mathcal{O}(e^4)$ as $e \to 0$; up to an additive constant, the score matches squared-error geometry with fourth-order corrections.

The proof is given in Appendix B.1.

Theorem 2 (Residual-wise direction and gradient attenuation). *Let*

$$w_{\max}(\nu) = \frac{27(\nu+7)(\nu+5)^2}{(\nu+3)^3}. \tag{8}$$

If $0 \leq w < w_{\max}(\nu)$, then

$$0 < \frac{\phi'_{\sigma,\nu,w}(e)}{2e} \leq 1 \quad \textit{for every } e \neq 0. \tag{9}$$

Consequently, $e = 0$ is the unique global minimizer. For every scalar residual coordinate, the gradient of $\phi_{\sigma,\nu,w}$ has the same sign as the corresponding squared-error gradient, while its magnitude is no larger than that of squared error. Moreover, $\phi'_{\sigma,\nu,w}(e)/(2e) \to 1/(1+w)$ as $|e| \to \infty$.

The proof in Appendix B.2 derives both envelopes of the Student correction and the tail limit. This guarantee is coordinate-wise in residual space and does not imply collinearity with the parameter-space gradient of MSE.

We apply the same scalar score to residual levels and residual innovations:

$$\mathcal{L}_{\text{level}} = \frac{1}{HC} \sum_{t=1}^{H} \sum_{c=1}^{C} \phi_{\sigma,\nu,w}(e_{t,c}), \tag{10}$$

$$\mathcal{L}_{\text{innov}} = \frac{1}{(H-1)C} \sum_{t=2}^{H} \sum_{c=1}^{C} \phi_{\sqrt{2}\sigma,\nu,w}(\Delta e_{t,c}), \tag{11}$$

$$\mathcal{L}_{\text{res}} = \frac{\mathcal{L}_{\text{level}} + \rho \mathcal{L}_{\text{innov}}}{1+2\rho}, \qquad \rho \geq 0. \tag{12}$$

Under an independent equal-scale reference in which adjacent residuals each have variance σ^2, their difference has reference variance $2\sigma^2$. We therefore apply the same score family to residual innovations using the adjusted reference scale $\sqrt{2}\sigma$. Expanding the definition gives $\Delta e_{t,c} = (\hat{y}_{t,c} - \hat{y}_{t-1,c}) - (y_{t,c} - y_{t-1,c})$: the branch penalizes disagreement between predicted and observed increments, not temporal variation itself. A sharp increase or decrease is therefore not suppressed when the target exhibits the same movement. This distinguishes two forecasts that have similar pointwise MSE but different temporal behavior; for example, a path alternating above and below the target produces larger residual innovations than a comparably biased but smoothly evolving path.

Proposition 3 (Aggregate trace-curvature preservation). *For one variable, let $\mathbf{e} = (e_1, \ldots, e_H)^\top \in \mathbb{R}^H$ be its residual path and let $\mathbf{G} \in \mathbb{R}^{(H-1)\times H}$ be the first-difference matrix satisfying $(\mathbf{G}\mathbf{e})_t = e_{t+1} - e_t$ for $1 \leq t \leq H-1$. Let ℓ_{level}, ℓ_{innov}, and ℓ_{res} denote the corresponding channelwise losses before averaging over variables. At the zero-residual path $\mathbf{e} = \mathbf{0}$,*

$$\operatorname{tr} \nabla^2_{\mathbf{e}} \ell_{\text{level}} = 2, \qquad \operatorname{tr} \nabla^2_{\mathbf{e}} \ell_{\text{innov}} = 4. \tag{13}$$

Therefore $\operatorname{tr} \nabla^2_{\mathbf{e}} \ell_{\text{res}} = 2$, exactly matching the trace curvature of mean squared error for every $\rho \geq 0$.

The proof is given in Appendix B.3. The innovation term still introduces cross-time coupling through the first-difference operator.

3.3 Overlap-constrained ordinal flow

Soft five-point order states. For variable c, we define the target-derived reference scale

$$\bar{y}_c = \frac{1}{H} \sum_{t=1}^{H} y_{t,c}, \qquad s_c = \operatorname{sg}\left(\max\left\{ \sqrt{\frac{1}{H} \sum_{t=1}^{H} (y_{t,c} - \bar{y}_c)^2}, \varepsilon \right\} \right), \tag{14}$$

where $\varepsilon > 0$ is a numerical floor and $\mathrm{sg}(\cdot)$ denotes stop-gradient. The same detached s_c is used for target and prediction. For a five-point window $V = (v_1, \ldots, v_5)$ with mean $\bar{v}$, define utilities $z_i = (v_i - \bar{v})/(\tau s_c)$, where $\tau > 0$ controls ranking softness. For $\pi = (\pi_1, \ldots, \pi_5) \in \mathfrak{S}_5$, we use a descending-value convention: π_1 is the local index assigned the highest rank. These utilities induce a Plackett–Luce distribution over the $5! = 120$ permutations in $\mathfrak{S}_5$:

$$p_\tau(\pi \mid V; s_c) = \prod_{\ell=1}^{4} \frac{\exp z_{\pi_\ell}}{\sum_{i \in \mathcal{A}_\ell(\pi)} \exp z_i}, \qquad \mathcal{A}_\ell(\pi) = \{1, \ldots, 5\} \setminus \{\pi_1, \ldots, \pi_{\ell-1}\}. \tag{15}$$

Applying Equation 15 to target and predicted windows gives source-state laws $p_t, \hat{p}_t \in \Delta^{119}$. The scale is detached so target and prediction share one ordinal coordinate system without allowing the prediction to manipulate its normalization.

Exact hard transitions and differentiable descriptors. Adjacent five-point windows share four observations.

Theorem 4 (Overlap transition representation)**.** *For six distinct underlying values, or ties resolved consistently across the two windows, every hard transition is uniquely represented by*

$$(\pi_t, j_{t+1}) \in \mathfrak{S}_5 \times \{1, \ldots, 5\}, \tag{16}$$

where $\pi_t \in \mathfrak{S}_5$ is the source-window ordering and $j_{t+1} \in \{1, \ldots, 5\}$ is the descending rank at which the entering observation is inserted among the retained observations. Hence each source has exactly five legal successors and the transition alphabet contains $5! \times 5 = 600$ states.

The construction deletes the departing local index from π_t, preserves the induced order of the four retained indices, and inserts the entering index at rank j_{t+1}. The inverse recovers the source state and insertion rank, proving uniqueness. The proof is given in Appendix B.4.

To construct its soft counterpart, recall that the entering observation is locally indexed as 5 inside every successor window. Enumerating $\mathfrak{S}_5$ as $\{\pi^{(k)}\}_{k=1}^{120}$, define the fixed rank-marginalization matrix $\mathbf{M} \in \{0,1\}^{120 \times 5}$ by

$$\mathbf{M}_{k,j} = \mathbb{I}[\mathrm{rank}_{\pi^{(k)}}(5) = j], \qquad k \in \{1, \ldots, 120\}, \quad j \in \{1, \ldots, 5\}, \tag{17}$$

where $\mathbb{I}[\cdot]$ is the indicator function and $\mathrm{rank}_\pi(i)$ is the position occupied by local index i in permutation π. The target and predicted insertion laws are

$$\iota_{t+1} = p_{t+1}^\top \mathbf{M}, \qquad \hat{\iota}_{t+1} = \hat{p}_{t+1}^\top \mathbf{M}. \tag{18}$$

We then form factorized soft transition descriptors

$$T_t(k, j) = p_t(k)\iota_{t+1}(j), \qquad \widehat{T}_t(k, j) = \hat{p}_t(k)\hat{\iota}_{t+1}(j). \tag{19}$$

The hard alphabet in Equation 16 is an exact combinatorial representation; Equation 19 is its factorized differentiable descriptor. This distinction allows the method to retain the legal transition structure while avoiding a discrete operation in the predicted branch.

For windows with unique rankings, the soft laws concentrate as $\tau \to 0$, so $\mathrm{vec}(T_t) \to \delta_{(\pi_t, j_{t+1})}$ with the analogous limit for $\widehat{T}_t$. The differentiable descriptor therefore approaches the exact source–insertion coordinate in the low-temperature limit while remaining smooth at finite temperature.

For probability vectors $\mathbf{r} = (r_1, \ldots, r_K)$ and $\mathbf{s} = (s_1, \ldots, s_K)$ on the same K-state alphabet, define the Bhattacharyya coefficient BC and squared Hellinger distance d_{H}^2 as

$$\mathrm{BC}(\mathbf{r}, \mathbf{s}) = \sum_{i=1}^{K} \sqrt{r_i s_i}, \qquad d_{\mathrm{H}}^2(\mathbf{r}, \mathbf{s}) = 1 - \mathrm{BC}(\mathbf{r}, \mathbf{s}). \tag{20}$$

Proposition 5 (Factorized Hellinger decomposition)**.** *For transition position t, let $D_t = d_{\mathrm{H}}^2\Big(\mathrm{vec}(T_t), \mathrm{vec}(\widehat{T}_t)\Big)$. Then*

$$D_t = 1 - \underbrace{\sum_{k=1}^{120} \sqrt{p_t(k)\hat{p}_t(k)}}_{A_{\mathrm{src},t}} \underbrace{\sum_{j=1}^{5} \sqrt{\iota_{t+1}(j)\hat{\iota}_{t+1}(j)}}_{A_{\mathrm{ins},t}}. \tag{21}$$

where $A_{\mathrm{src},t}$ measures target–prediction agreement between source-state laws and $A_{\mathrm{ins},t}$ measures agreement between insertion-rank laws.

Equivalently, equation 21 reveals the multiplicative decomposition $1-D_t = A_{\mathrm{src},t}A_{\mathrm{ins},t}$. A transition achieves high affinity only when the predicted path simultaneously recovers the source configuration and the insertion behavior of the entering observation. The factorization evaluates this transition affinity directly from the two marginal agreements. The proof is given in Appendix B.5.

Near-ties should not receive the same weight as clearly resolved target orders. We therefore use the target-only concentration gate

$$\mathrm{Ent}_K(\mathbf{r}) = -\frac{\sum_{i=1}^{K} r_i \log r_i}{\log K}, \qquad 0 \log 0 := 0, \tag{22}$$

followed by

$$I_t = \left[1 - \tfrac{1}{2}\left(\mathrm{Ent}_{120}(p_t) + \mathrm{Ent}_5(\iota_{t+1})\right)\right]_{[0,1]}, \qquad \mathcal{L}_{\mathrm{ord}} = \frac{1}{(H-5)C}\sum_{t=1}^{H-5}\sum_{c=1}^{C} I_{t,c} D_{t,c}, \tag{23}$$

where Ent_K is entropy normalized by $\log K$ and $[x]_{[0,1]} = \min\{1, \max\{0, x\}\}$ denotes clipping. All target-derived quantities in the gate and target transition descriptor are detached. Appendix A provides supporting details.

Gradients flow through $(\hat{p}_t, \hat{\iota}_{t+1})$ to $\widehat{\boldsymbol{Y}}$, while $(s_c, p_t, \iota_{t+1}, I_t)$ remain fixed during each update.

3.4 Unified objective

The final training objective is

$$\mathcal{L}_{\mathrm{HYOT}} = \mathcal{L}_{\mathrm{res}} + \lambda \mathcal{L}_{\mathrm{ord}}, \qquad \lambda \geq 0. \tag{24}$$

The parameters (σ, ν, w) specify the reference residual scale, Student-tail geometry, and strength of robust gradient attenuation; ρ weights residual evolution; τ controls ranking softness; and λ combines cardinal and ordinal supervision. Training minimizes the empirical average of Equation 24.

HYOT requires one forward pass of the forecasting backbone. Both branches are computed directly from $(\widehat{\boldsymbol{Y}}, \boldsymbol{Y})$ and introduce no auxiliary encoder, decoder, graph, latent variable, or learned parameter. The fixed permutation enumeration and rank-marginalization matrix are shared across datasets and models. Consequently, training uses the original backbone plus the HYOT loss, whereas inference uses the original backbone alone with unchanged parameter count and prediction procedure. Because the ordinal window size and permutation table are constants, the additional training complexity per forecast path grows linearly as $\mathcal{O}(CH)$.

4 Experiments

Protocol. We evaluate HYOT across five forecasting backbones and six datasets under each model's official training configuration. The input length is fixed to 96, and the forecast horizon is $H \in \{96, 192, 336, 720\}$. Every paired experiment shares the backbone, data split, and optimization setup and differs in the training objective. Unless stated otherwise, the loss uses $\sigma = 0.3$, $\nu = 12$, $w = 5$, $\rho = 0.5$, $\tau = 0.25$, and $\lambda = 0.2$. Performance is measured by test MSE and MAE.

Main results. Table 1 pairs MSE-loss training with HYOT under matched experimental settings. Across the 120 paired settings, HYOT improves aggregate performance, with mean relative MSE and MAE reductions of 2.12% and 2.80%, respectively. The improvement is distributed across model families rather than being driven by a single backbone: DLinear delivers average reductions of 4.45% in MSE and 4.90% in MAE, while the Transformer- and mixing-based forecasters also benefit across multiple datasets and horizons. The effect is particularly pronounced on longer trajectories, where local deviations can accumulate along the forecast path. On ETTh2 at $H = 720$, DLinear MSE falls from 0.774 to 0.578 (25.3%) and MAE falls from 0.631 to 0.536 (15.1%); DLinear further averages an 11.2% MSE reduction across ETTh2 and 6.0% MSE/6.6% MAE reductions across ETTm2. The gains transfer across architectures, reaching 9.9% for TimeBridge on ETTh1 at $H = 720$, 9.1% for TimeMixer on ETTh1 at $H = 336$, and 8.6% for TimesNet on ETTm1 at $H = 720$. Taken together, these results show that supervising residual evolution and local ordinal flow complements distinct architectural inductive biases and improves aggregate forecasting accuracy across the evaluation matrix.

Table 1: Long-term forecasting results with input length 96; bold marks the better paired value.

Models		DLinear				TimesNet				TimeBridge				TimeMixer				NLinear			
		MSE loss		**HyOT**		MSE loss		**HyOT**		MSE loss		**HyOT**		MSE loss		**HyOT**		MSE loss		**HyOT**	
Dataset	H	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.383	0.396	**0.380**	**0.391**	0.389	0.412	**0.384**	**0.403**	**0.374**	0.395	0.378	**0.388**	0.385	0.401	**0.370**	**0.386**	0.393	0.400	**0.384**	**0.392**
	192	**0.433**	**0.426**	0.437	0.429	0.440	0.442	**0.425**	**0.428**	**0.426**	0.426	0.430	**0.418**	0.443	0.430	**0.431**	**0.420**	0.449	0.433	**0.445**	**0.428**
	336	0.491	0.467	**0.473**	**0.447**	0.494	0.471	**0.462**	**0.446**	0.481	0.461	**0.473**	**0.439**	0.513	0.470	**0.466**	**0.433**	0.485	0.449	**0.475**	**0.439**
	720	0.528	0.519	**0.491**	**0.489**	0.519	0.495	**0.490**	**0.477**	0.528	0.505	**0.476**	**0.461**	**0.493**	0.473	0.493	**0.465**	0.471	0.462	**0.466**	**0.457**
	Avg	0.459	0.452	**0.445**	**0.439**	0.460	0.455	**0.440**	**0.438**	0.452	0.447	**0.439**	**0.427**	0.458	0.444	**0.440**	**0.426**	0.449	0.436	**0.443**	**0.429**
ETTh2	96	0.329	0.380	**0.293**	**0.348**	0.337	0.371	**0.316**	**0.357**	0.293	0.345	**0.285**	**0.334**	**0.290**	0.343	0.293	**0.340**	0.289	0.338	**0.288**	**0.335**
	192	0.431	0.443	**0.400**	**0.418**	0.405	0.415	**0.390**	**0.398**	0.374	0.395	**0.366**	**0.385**	0.383	0.402	**0.369**	**0.389**	0.375	0.390	**0.375**	**0.388**
	336	0.459	0.462	**0.451**	**0.456**	0.455	0.451	**0.438**	**0.439**	0.425	0.432	**0.418**	**0.423**	**0.416**	**0.431**	0.430	0.432	**0.414**	0.426	0.415	**0.423**
	720	0.774	0.631	**0.578**	**0.536**	**0.434**	**0.448**	0.451	0.457	0.422	0.440	**0.420**	**0.437**	**0.419**	**0.441**	0.428	0.443	0.423	0.444	**0.418**	**0.437**
	Avg	0.498	0.479	**0.431**	**0.440**	0.408	0.421	**0.399**	**0.413**	0.379	0.403	**0.372**	**0.395**	**0.377**	0.404	0.380	**0.401**	0.375	0.399	**0.374**	**0.395**
ETTm1	96	0.346	0.374	**0.332**	**0.354**	0.334	0.375	**0.330**	**0.363**	0.327	0.365	**0.311**	**0.342**	0.326	0.363	**0.310**	**0.342**	0.356	0.376	**0.338**	**0.357**
	192	0.383	0.393	**0.374**	**0.376**	0.406	0.413	**0.373**	**0.384**	0.375	0.389	**0.364**	**0.371**	0.369	0.389	**0.357**	**0.368**	0.394	0.395	**0.382**	**0.379**
	336	0.417	0.418	**0.406**	**0.399**	0.416	0.423	**0.399**	**0.404**	0.403	0.410	**0.400**	**0.397**	0.399	0.408	**0.385**	**0.392**	0.427	0.416	**0.417**	**0.403**
	720	0.479	0.457	**0.469**	**0.438**	0.505	0.468	**0.461**	**0.437**	0.473	0.452	**0.464**	**0.436**	0.458	0.446	**0.448**	**0.429**	0.489	0.451	**0.484**	**0.439**
	Avg	0.406	0.410	**0.395**	**0.392**	0.415	0.420	**0.391**	**0.397**	0.394	0.404	**0.385**	**0.386**	0.388	0.401	**0.375**	**0.383**	0.416	0.409	**0.405**	**0.394**
ETTm2	96	0.187	0.281	**0.181**	**0.265**	0.190	0.266	**0.184**	**0.262**	0.179	0.261	**0.175**	**0.252**	**0.175**	0.258	0.178	**0.256**	0.182	0.265	**0.180**	**0.258**
	192	0.272	0.349	**0.249**	**0.318**	0.252	0.307	**0.247**	**0.301**	0.244	0.304	**0.243**	**0.298**	0.242	0.302	**0.237**	**0.296**	0.246	0.304	**0.245**	**0.299**
	336	0.343	0.395	**0.317**	**0.367**	0.322	0.349	**0.305**	**0.337**	**0.307**	0.344	0.310	**0.342**	0.295	0.340	**0.293**	**0.333**	**0.306**	0.341	0.306	**0.337**
	720	0.439	0.444	**0.418**	**0.422**	0.420	0.406	**0.418**	**0.402**	0.404	0.401	**0.402**	**0.398**	0.393	0.397	**0.384**	**0.389**	0.408	0.398	**0.406**	**0.394**
	Avg	0.310	0.367	**0.291**	**0.343**	0.296	0.332	**0.288**	**0.325**	0.283	0.327	**0.282**	**0.323**	0.276	0.324	**0.273**	**0.319**	0.286	0.327	**0.284**	**0.322**
Weather	96	0.197	0.255	**0.196**	**0.239**	0.169	0.219	**0.164**	**0.211**	0.176	0.217	**0.172**	**0.209**	0.164	0.212	**0.162**	**0.202**	**0.202**	0.240	0.204	**0.237**
	192	0.239	0.297	**0.236**	**0.281**	**0.225**	0.265	0.226	**0.262**	0.221	0.257	**0.220**	**0.251**	0.211	0.254	**0.209**	**0.246**	**0.248**	0.277	0.250	**0.274**
	336	0.283	0.332	**0.278**	**0.312**	0.282	0.304	**0.278**	**0.297**	0.277	0.298	**0.275**	**0.292**	**0.264**	0.293	0.265	**0.287**	**0.300**	0.313	0.301	**0.310**
	720	0.348	0.385	**0.338**	**0.362**	0.359	0.355	**0.356**	**0.350**	0.353	0.349	**0.351**	**0.342**	**0.339**	0.343	0.343	**0.338**	**0.373**	0.360	0.374	**0.358**
	Avg	0.267	0.317	**0.262**	**0.299**	0.259	0.286	**0.256**	**0.280**	0.257	0.280	**0.255**	**0.273**	**0.244**	0.275	0.245	**0.268**	**0.281**	0.298	0.282	**0.295**
Exchange	96	**0.078**	**0.199**	0.082	0.204	**0.105**	**0.235**	0.114	0.245	0.092	0.214	**0.083**	**0.201**	**0.083**	**0.201**	0.091	0.210	0.086	0.202	**0.082**	**0.198**
	192	0.156	0.292	**0.151**	**0.283**	**0.219**	0.340	0.219	**0.337**	0.191	0.315	**0.178**	**0.301**	**0.184**	**0.305**	0.194	0.313	0.176	0.296	**0.172**	**0.293**
	336	0.307	0.419	**0.274**	**0.393**	**0.363**	**0.439**	0.375	0.441	**0.341**	**0.427**	0.348	0.428	0.344	0.422	**0.337**	**0.419**	0.329	0.414	**0.317**	**0.406**
	720	0.782	0.669	**0.778**	**0.659**	**0.931**	**0.735**	0.970	0.748	0.867	0.703	**0.850**	**0.695**	**1.161**	0.798	1.162	**0.789**	0.886	0.711	**0.852**	**0.696**
	Avg	0.331	0.395	**0.321**	**0.385**	**0.405**	**0.437**	0.419	0.443	0.373	0.415	**0.365**	**0.406**	**0.443**	**0.431**	0.446	0.433	0.369	0.406	**0.356**	**0.398**

Table 2: Training-objective comparison with input length 96; bold marks the best value.

	DLinear / ETTh2				**DLinear / ETTm2**				**TimeMixer / ETTm2**			
	$H = 96$		$H = 720$		$H = 96$		$H = 720$		$H = 96$		$H = 720$	
Objective	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
MSE loss	0.329	0.380	0.774	0.631	0.187	0.281	0.439	0.444	**0.175**	0.258	0.393	0.397
Soft-DTW	0.384	0.425	0.750	0.619	0.209	0.300	0.609	0.542	0.192	0.273	0.485	0.459
DILATE	0.370	0.412	0.754	0.621	0.207	0.299	0.599	0.538	0.191	0.272	0.469	0.446
PS Loss	0.299	0.351	0.592	0.552	0.189	0.288	0.441	0.461	0.188	0.272	0.399	0.411
HyOT (ours)	**0.293**	**0.348**	**0.578**	**0.536**	**0.181**	**0.265**	**0.418**	**0.422**	0.178	**0.256**	**0.384**	**0.389**

Training-objective comparison. Table 2 compares HyOT with MSE, Soft-DTW (Cuturi & Blondel, 2017), DILATE (Le Guen & Thome, 2019), and the recent Patch-wise Structural (PS) Loss (Kudrat et al., 2025) under matched architecture, data, and optimization settings. HyOT achieves lower MSE and MAE in all $36/36$ comparisons with the three structural objectives. Relative to PS Loss at $H = 720$, HyOT reduces MSE/MAE by **2.4%/3.0%** on DLinear–ETTh2, **5.2%/8.5%** on DLinear–ETTm2, and **3.8%/5.2%** on TimeMixer–ETTm2. The gain is therefore not merely due to adding structural supervision: HyOT also outperforms existing alignment- and patch-structural objectives.

Component ablation. Table 3 isolates the contributions of the two geometric branches and their internal structure. Removing the Student correction weakens the robust cardinal geometry, increasing ETTh2-720 MSE/MAE from $0.578/0.536$ to $0.620/0.555$. Removing innovations produces a larger degradation to $0.642/0.566$, showing that matching residual levels alone does not recover how errors evolve along the horizon. The ordinal branch is also substantive: without ordinal flow, the same setting reaches $0.628/0.560$, while retaining local ranking but removing the overlap-constrained transition structure gives $0.603/0.548$. The latter remains better than discarding ordinal flow entirely, which indicates that local order and its transition dynamics provide complementary information. The same ordering is visible on ETTm2-720, and the full objective is best in all $6/6$ reported metrics, supporting the joint use of robust cardinal correction, innovations, ordinal geometry, and source–insertion transitions.

Table 3: DLinear component ablation with input length 96; bold marks the best value.

	ETTh2, $H = 96$		**ETTh2,** $H = 720$		**ETTm2,** $H = 720$	
Variant	MSE	MAE	MSE	MAE	MSE	MAE
MSE loss	0.329	0.380	0.774	0.631	0.439	0.444
w/o Student correction	0.304	0.357	0.620	0.555	0.426	0.430
w/o innovations	0.311	0.363	0.642	0.566	0.431	0.435
w/o ordinal flow	0.306	0.359	0.628	0.560	0.428	0.432
w/o transition structure	0.299	0.353	0.603	0.548	0.423	0.427
Full HYOT	**0.293**	**0.348**	**0.578**	**0.536**	**0.418**	**0.422**

Window-size sensitivity. Table 4 and Figure 2 show a shallow U-shaped response around $m = 5$, which attains the best value in all six reported metrics. Increasing m from 3 to 5 reduces MSE/MAE by $4.0\%/2.5\%$ on DLinear–ETTh2, $2.6\%/2.1\%$ on DLinear–ETTm2, and $2.0\%/1.8\%$ on TimeMixer–ETTm2. Shorter windows provide less context for distinguishing local trajectory configurations, whereas expanding beyond five points yields no additional gain: moving from $m = 5$ to $m = 6$ changes MSE by only 0.003, 0.002, and 0.001 in the three settings, and $m = 7$ degrades each metric further. The smooth variation on both sides of the minimum indicates that performance is not brittle to a single window size, while the repeated optimum identifies $m = 5$ as an effective empirical balance between local context and transition complexity under the evaluated settings.

Table 4: Window-size sweep at $H = 720$; bold marks the best value.

	DLinear / ETTh2		**DLinear / ETTm2**		**TimeMixer / ETTm2**	
Window size m	MSE	MAE	MSE	MAE	MSE	MAE
3	0.602	0.550	0.429	0.431	0.392	0.396
4	0.585	0.540	0.421	0.425	0.386	0.391
5	**0.578**	**0.536**	**0.418**	**0.422**	**0.384**	**0.389**
6	0.581	0.538	0.420	0.424	0.385	0.390
7	0.589	0.544	0.424	0.427	0.388	0.393

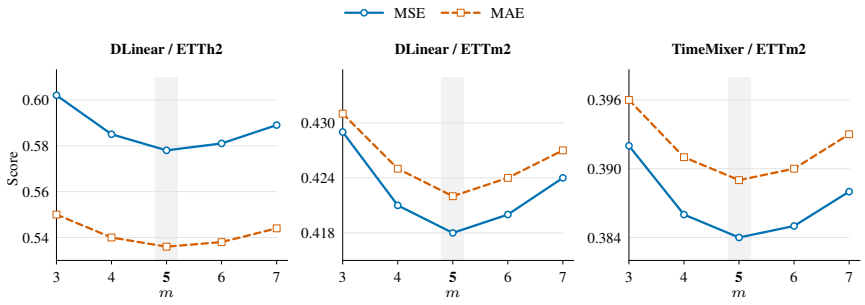


Figure 2: Metric trends across ordinal window sizes m at $H = 720$.

Trajectory-structure evaluation. Table 5 and Figure 3 compare pointwise accuracy and path structure. Innovation MAE measures cardinal evolution through first-difference error, Ordinal Hellinger measures source–insertion transition discrepancy, and Local Rank Agreement measures local ordinal structure through the ten pairwise orders in each five-point window. HYOT reduces Innovation MAE by 9.5% and Ordinal Hellinger by 8.2%, while raising Local Rank Agreement by 3.0 percentage points, demonstrating better local evolution and order structure alongside lower MSE.

Table 5: Trajectory-structure evaluation on TimeMixer–ETTm2 at $H = 720$.

Objective	MSE↓	Innovation MAE↓	Ordinal Hellinger↓	Local Rank Agreement↑
MSE loss	0.393	0.106	0.310	0.526
HyOT	**0.384**	**0.096**	**0.285**	**0.556**

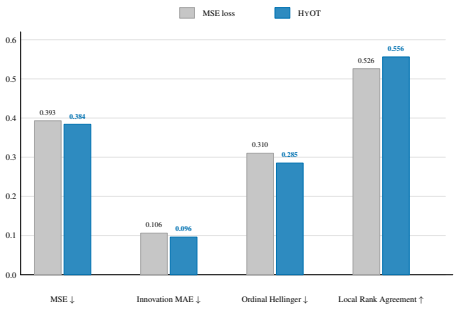


Figure 3: Trajectory metrics for TimeMixer–ETTm2 at $H = 720$.

Hyperparameter sensitivity. We conduct a one-at-a-time sensitivity study on DLinear–ETTh2 at $H = 720$, varying one loss parameter while fixing the remaining parameters to the defaults in Section 4. Table 6 reports test errors to three decimal places.

Table 6: Hyperparameter sensitivity on DLinear–ETTh2 ($H = 720$); group-wise best values in bold.

λ			τ			ρ			w		
Value	MSE↓	MAE↓	Value	MSE↓	MAE↓	Value	MSE↓	MAE↓	Value	MSE↓	MAE↓
0.00	0.628	0.560	0.10	0.591	0.544	0.00	0.642	0.566	0	0.620	0.555
0.05	0.603	0.550	0.15	0.583	0.539	0.25	0.594	0.545	2	0.590	0.543
0.10	0.587	0.541	**0.25**	**0.578**	**0.536**	**0.50**	**0.578**	**0.536**	**5**	**0.578**	**0.536**
0.20	**0.578**	**0.536**	0.35	0.581	0.538	0.75	0.582	0.538	10	0.582	0.539
0.30	0.581	0.538	0.50	0.589	0.543	1.00	0.590	0.543	20	0.592	0.545
0.40	0.588	0.542		–			–			–	

Across all four sweeps, the default values form interior minima. Setting $\lambda = 0$ removes ordinal supervision and raises MSE/MAE to $0.628/0.560$; the best result occurs at $\lambda = 0.20$, with only mild variation from 0.10 to 0.30. Likewise, $\tau = 0.25$ balances an overly sharp ranking law at small temperatures against diffuse ordinal distinctions at large temperatures. Removing innovations gives $0.642/0.566$, whereas $\rho = 0.50$ is best and $\rho = 0.75$ remains close. The moderate Student weight $w = 5$ is also best, while both $w = 0$ and $w = 20$ are worse. These smooth neighborhoods show that the selected configuration is not an isolated operating point.

5 Conclusion

We argued that long-horizon forecasting should be optimized as a structured path rather than as independent coordinates. HyOT implements this principle through two complementary forms of supervision: its cardinal branch scores residual levels and innovations with curvature-matched Gaussian–Student Hyvärinen geometry, while its ordinal branch models local ordering structure through overlapping ranking windows and factorized source–insertion descriptors. Across five forecasting architectures, six datasets, and four forecast horizons, HyOT improves aggregate MSE and MAE relative to conventional MSE training and outperforms the evaluated structural objectives in all reported comparisons. In the representative TimeMixer–ETTm2 evaluation of Table 5, the MSE reduction can be accompanied by 9.5% lower Innovation MAE, 8.2% lower Ordinal Hellinger, and 3.0 percentage points higher Local Rank Agreement, linking pointwise accuracy to cardinal evolution and local ordinal structure. Both branches use the same forecast and leave inference unchanged.

Reproducibility Statement

Section 4 specifies the paired evaluation protocol, input length, forecast horizons, metrics, and default loss parameters. Section 3 gives the objective, definitions, and theoretical statements; Appendix A provides supporting details, Appendix B provides the theoretical proofs, and Section 4 reports the hyperparameter sensitivity study.

AI Use Statement

Generative AI tools were used to assist with reasoning, language editing, and manuscript organization. The authors reviewed and verified all AI-assisted text, mathematical statements, citations, implementation details, and experimental claims, and remain responsible for the complete content of the submission.

References

Christoph Bandt and Bernd Pompe. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88(17):174102, 2002. doi: 10.1103/PhysRevLett.88.174102.

Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 894–903. PMLR, 2017. URL https://proceedings.mlr.press/v70/cuturi17a.html.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

Dilfira Kudrat, Zongxia Xie, Yanru Sun, Tianyu Jia, and Qinghua Hu. Patch-wise structural loss for time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31841–31859. PMLR, 2025. URL https://proceedings.mlr.press/v267/kudrat25a.html.

Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/466accbac9a66b805ba50e42ad715740-Paper.pdf.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. TimeBridge: Non-stationarity matters for long-term time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39815–39840. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25cb.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Dover Publications, 2005.

Michael McCullough, Michael Small, Thomas Stemler, and Herbert Ho-Ching Iu. Time lagged ordinal partition networks for capturing dynamics of continuous dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(5):053101, 2015. doi: 10.1063/1.4919075.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012. doi: 10.1214/12-AOS971.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975. doi: 10.2307/2346567.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

A Supporting method details

A.1 Problem formulation: scoring a forecast path

Let L, H, and C denote the input length, forecast horizon, and number of variables, respectively. We consider $H \geq 6$, which provides at least two overlapping five-point windows and one adjacent ordinal transition. A forecasting instance contains a history $\boldsymbol{X} \in \mathbb{R}^{L\times C}$ and its future target $\boldsymbol{Y} \in \mathbb{R}^{H\times C}$. A forecasting backbone f_θ with learnable parameters θ produces the point forecast

$$\widehat{\boldsymbol{Y}} = f_\theta(\boldsymbol{X}), \qquad \widehat{\boldsymbol{Y}} \in \mathbb{R}^{H\times C}. \tag{25}$$

HyOT treats a forecast horizon as a structured path with two complementary geometries: cardinal geometry describes numerical deviation and its temporal propagation, whereas ordinal geometry describes local ordering structure and its evolution across overlapping windows. The *cardinal branch* operates in residual coordinates and scores both pointwise deviations and their temporal changes. In parallel, the *ordinal branch* converts overlapping local windows into soft ranking states and scores how those states evolve along the horizon. Both branches are computed from the same $\widehat{\boldsymbol{Y}}$ and jointly optimize f_θ.

We use $t \in \{1, \ldots, H\}$ for a forecast position and $c \in \{1, \ldots, C\}$ for a variable. For paired target–prediction quantities, a hat denotes the prediction-derived counterpart; residual, transition, and loss quantities use their own dedicated symbols. The cardinal branch begins with the residual matrix and its first temporal difference,

$$\boldsymbol{E} = \widehat{\boldsymbol{Y}} - \boldsymbol{Y} \in \mathbb{R}^{H\times C}, \qquad e_{t,c} = \hat{y}_{t,c} - y_{t,c}, \tag{26}$$

$$\Delta\boldsymbol{E} \in \mathbb{R}^{(H-1)\times C}, \qquad \Delta e_{t,c} = e_{t,c} - e_{t-1,c}, \quad t \geq 2. \tag{27}$$

The ordinal branch forms target and predicted five-point windows

$$W_{t,c} = (y_{t,c}, \ldots, y_{t+4,c}), \qquad \widehat{W}_{t,c} = (\hat{y}_{t,c}, \ldots, \hat{y}_{t+4,c}) \in \mathbb{R}^5, \quad 1 \leq t \leq H-4. \tag{28}$$

Thus the cardinal path is $\widehat{\boldsymbol{Y}} \mapsto \{\boldsymbol{E}, \Delta\boldsymbol{E}\} \mapsto \mathcal{L}_{\text{res}}$, whereas the ordinal path is $\widehat{\boldsymbol{Y}} \mapsto \{\widehat{W}_{t,c}\} \mapsto$ soft ordinal transitions $\mapsto \mathcal{L}_{\text{ord}}$. The two scalar losses are combined with a nonnegative branch coefficient λ. For notational clarity, we suppress c while defining scalar or per-window transformations and restore it in loss aggregation. Figure 1 in the main paper shows the complete flow.

A.2 Curvature-matched Hyvärinen residual field

The cardinal branch assigns every scalar residual a curvature-matched Gaussian–Student Hyvärinen score. The same score family is applied to both the residual levels in $\boldsymbol{E}$ and the residual innovations in $\Delta\boldsymbol{E}$, with the reference scale adjusted for the innovation field. The resulting scalar loss jointly measures numerical error and short-range error evolution.

A.2.1 FROM LOCAL PROPER SCORING TO A RESIDUAL LOSS

The Hyvärinen score and Student reference density are defined in Equations 3–4 of the main text. Their closed-form Student score is

$$H_{\mathrm{St}}(e;\sigma,\nu) = \frac{(\nu+1)\left[\frac{1}{2}(\nu+3)e^2 - a\right]}{(a+e^2)^2}. \tag{29}$$

The resulting rational score remains finite for every residual and its contribution decays in the tails.

A.2.2 EXACT CURVATURE MATCHING

To align the Student-induced geometry with squared error at small residuals, we define the curvature factor $\kappa_{\nu,\sigma}$ and the scalar residual score $\phi_{\sigma,\nu,w} : \mathbb{R} \to \mathbb{R}$ by

$$\kappa_{\nu,\sigma} = \frac{2a^2}{(\nu+1)(\nu+7)}, \qquad \phi_{\sigma,\nu,w}(e) = \frac{e^2 + w\kappa_{\nu,\sigma} H_{\mathrm{St}}(e;\sigma,\nu)}{1+w}, \tag{30}$$

Here $w \geq 0$ controls the contribution of the Student-induced geometry. The curvature factor $\kappa_{\nu,\sigma}$ aligns its local learning geometry with squared error, while (σ, ν, w) jointly determine the reference scale, tail geometry, and degree of gradient attenuation.

Proposition 1 in the main text states this result.

The proof is given in Appendix B.1. The result follows from $H''_{\mathrm{St}}(0) = (\nu+1)(\nu+7)/a^2$; the factor $\kappa_{\nu,\sigma}$ is its exact reciprocal-curvature correction. The score therefore preserves the backbone's small-error learning scale while controlling tail influence.

A.2.3 RESIDUAL-WISE DIRECTION AND GRADIENT ATTENUATION

Beyond local curvature matching, the proposed residual field preserves the descent orientation of squared error in each scalar residual coordinate. This property holds over a closed-form admissible range of w.

Theorem 2 in the main text states this result.

The proof in Appendix B.2 derives both envelopes of the Student correction and the tail limit. Thus, admissible weights preserve the local learning geometry of MSE and retain its residual-wise descent orientation over the entire scalar residual domain, while progressively attenuating the relative influence of large residuals toward the tails. This guarantee is coordinate-wise in residual space and does not imply collinearity with the parameter-space gradient of MSE.

A.2.4 RESIDUAL LEVELS AND RESIDUAL INNOVATIONS

The residual-level loss averages equation 30 over all horizon positions and variables:

$$\mathcal{L}_{\mathrm{level}} = \frac{1}{HC}\sum_{t=1}^{H}\sum_{c=1}^{C}\phi_{\sigma,\nu,w}(e_{t,c}). \tag{31}$$

It measures how far each predicted value is from its target. The innovation field complements this pointwise quantity by recording how the error changes between adjacent forecast positions:

$$\begin{aligned}\Delta e_{t,c} &= e_{t,c} - e_{t-1,c} \\ &= (\hat{y}_{t,c} - \hat{y}_{t-1,c}) - (y_{t,c} - y_{t-1,c}), \qquad 2 \leq t \leq H.\end{aligned} \tag{32}$$

Thus $\Delta e_{t,c}$ is exactly the mismatch between the predicted and observed local increments. Scoring it encourages the forecast to reproduce target values together with their short-range temporal evolution. Under an independent equal-scale reference in which adjacent residuals each have variance σ^2, their difference has reference variance $2\sigma^2$. We therefore apply the same score family to residual innovations using the adjusted reference scale $\sqrt{2}\sigma$:

$$\mathcal{L}_{\mathrm{innov}} = \frac{1}{(H-1)C}\sum_{t=2}^{H}\sum_{c=1}^{C}\phi_{\sqrt{2}\sigma,\nu,w}(\Delta e_{t,c}). \tag{33}$$

Let $\rho \geq 0$ control the contribution of the innovation field. We aggregate the level and innovation fields using the trace-curvature-preserving normalization

$$\mathcal{L}_{\text{res}} = \frac{\mathcal{L}_{\text{level}} + \rho \mathcal{L}_{\text{innov}}}{1 + 2\rho}. \tag{34}$$

This normalization preserves the aggregate Hessian-trace curvature of pointwise MSE at the zero-residual path.

Proposition 3 in the main text states this result.

The proof is given in Appendix B.3. Proposition 3 establishes that the normalization in equation 34 exactly preserves per-variable trace curvature at the zero-residual path for every $\rho \geq 0$. The innovation term still introduces cross-time coupling through the first-difference operator.

A.3 Overlap-constrained five-point ordinal flow

The residual field measures distances in value space, but paths with similar amplitude error can still have different peak–valley order. The ordinal branch therefore scores local shape evolution in three stages. First, each overlapping five-point window is mapped to a differentiable distribution over rankings. Second, the four observations shared by consecutive windows reduce a temporal transition to a source ranking and the insertion rank of the entering observation. Third, predicted and target transition descriptors are compared with a target ordinal-concentration-weighted Hellinger discrepancy.

A.3.1 A differentiable law over five-point orderings

We construct the ordinal relaxation in a scale-normalized coordinate system so that its sharpness reflects local ordering structure. Equation 14 in the main text defines the detached target scale s_c, shared by target and prediction. For either window $V = (v_1, \ldots, v_5) \in \mathbb{R}^5$, let $\bar{v} = 5^{-1} \sum_{i=1}^{5} v_i$ and define the utility of its ith local position as

$$z_i(V; s_c) = \frac{v_i - \bar{v}}{\tau s_c}, \qquad i \in \{1, \ldots, 5\}, \tag{35}$$

where the temperature $\tau > 0$ controls ranking softness. Centering provides an explicit translation-normalized representation, while the shared scale places target and prediction in a common ordinal coordinate system.

Let $\mathfrak{S}_5$ be the set of all $5! = 120$ permutations of the local indices $\{1, \ldots, 5\}$. For $\pi = (\pi_1, \ldots, \pi_5) \in \mathfrak{S}_5$, we use a descending-value convention: π_1 is the local index assigned the highest rank. The Plackett–Luce probability of π is

$$p_\tau(\pi \mid V; s_c) = \prod_{\ell=1}^{4} \frac{\exp z_{\pi_\ell}(V; s_c)}{\sum_{i \in \mathcal{A}_\ell(\pi)} \exp z_i(V; s_c)}, \qquad \mathcal{A}_\ell(\pi) = \{1, \ldots, 5\} \setminus \{\pi_1, \ldots, \pi_{\ell-1}\}, \tag{36}$$

where $\ell \in \{1, \ldots, 4\}$ is the sequential selection stage and $\mathcal{A}_\ell(\pi)$ is the set of local indices not selected before stage ℓ. Enumerate $\mathfrak{S}_5$ as $\{\pi^{(k)}\}_{k=1}^{120}$. For a fixed variable c, each target window W_t and predicted window $\widehat{W}_t$ from equation 28 then induce the categorical state vectors

$$p_t(k) = p_\tau(\pi^{(k)} \mid W_t; s_c), \qquad \hat{p}_t(k) = p_\tau(\pi^{(k)} \mid \widehat{W}_t; s_c), \qquad k \in \{1, \ldots, 120\}. \tag{37}$$

Here $p_t, \hat{p}_t \in \Delta^{119}$, where $\Delta^{K-1} = \{\mathbf{r} \in \mathbb{R}_+^K : \sum_{i=1}^{K} r_i = 1\}$ denotes the probability simplex over a K-state alphabet. The vector p_t is target-derived, $\hat{p}_t$ is prediction-derived, and $t \in \{1, \ldots, H-4\}$. For distinct window values, the law converges to the hard descending permutation as $\tau \to 0$; at finite temperature it remains differentiable and expresses ranking uncertainty.

A.3.2 The shared-four-point transition theorem

Consider six consecutive scalar values $(v_t, \ldots, v_{t+5})$ from one variable of a forecast path. Their adjacent windows are

$$V_t = [v_t, v_{t+1}, v_{t+2}, v_{t+3}, v_{t+4}], \qquad V_{t+1} = [v_{t+1}, v_{t+2}, v_{t+3}, v_{t+4}, v_{t+5}]. \tag{38}$$

The value v_t is the departing observation, $(v_{t+1}, \ldots, v_{t+4})$ are the four retained observations, and v_{t+5} is the entering observation. Overlap imposes an exact combinatorial constraint on ordinal evolution: the successor ordering is completely characterized by the source ordering together with the insertion rank of the entering observation.

Theorem 4 in the main text states this result.

The proof is given in Appendix B.4. Hence the transition alphabet contains exactly $5! \times 5 = 600$ valid ordinal transitions. The construction deletes the departing local index from π_t, preserves the induced order of the four retained indices, and inserts the entering index at rank j_{t+1}. The inverse recovers the source state and insertion rank, proving uniqueness.

The rank-marginalization matrix $\mathbf{M}$ is defined in Equation 17 of the main text. Marginalizing the successor-state law over all permutations that assign the same rank to local index 5 gives

$$\iota_{t+1} = p_{t+1}^{\top}\mathbf{M}, \qquad \hat{\iota}_{t+1} = \hat{p}_{t+1}^{\top}\mathbf{M}, \qquad \iota_{t+1}, \hat{\iota}_{t+1} \in \Delta^4. \tag{39}$$

The vectors ι_{t+1} and $\hat{\iota}_{t+1}$ are five-way insertion-rank laws. On the exact hard-transition alphabet with source state k and insertion rank j, define the factorized differentiable soft descriptors

$$T_t(k, j) = p_t(k)\iota_{t+1}(j), \qquad \widehat{T}_t(k, j) = \hat{p}_t(k)\hat{\iota}_{t+1}(j). \tag{40}$$

After vectorization, $\mathrm{vec}(T_t), \mathrm{vec}(\widehat{T}_t) \in \Delta^{599}$ for $1 \le t \le H - 5$; these factorized soft distributions are supported on the 120×5 legal hard-transition alphabet. For uniquely ranked adjacent windows, $\mathrm{vec}(T_t) \to \delta_{(\pi_t, j_{t+1})}$ as $\tau \to 0$, and analogously for $\mathrm{vec}(\widehat{T}_t)$; here $\delta_{(\pi,j)}$ is the one-hot mass at state (π, j). Thus the factorized soft descriptor converges to the corresponding hard-transition coordinate of Theorem 4.

A.3.3 Factorized Hellinger transition scoring

We compare target and prediction transition descriptors with squared Hellinger distance, which is bounded, symmetric, and stable when some states receive very small mass. Equation 20 in the main text defines the Bhattacharyya coefficient and squared Hellinger distance.

Proposition 5 in the main text states the factorized Hellinger decomposition.

Equivalently, equation 21 reveals the multiplicative decomposition $1 - D_t = A_{\mathrm{src},t} A_{\mathrm{ins},t}$. A transition achieves high affinity only when the predicted path simultaneously recovers the source configuration and the insertion behavior of the entering observation. The factorization evaluates this transition affinity directly from the two marginal agreements. The proof is given in Appendix B.5.

A.3.4 Target ordinal-concentration gating

Target ordinal-concentration gating emphasizes transitions whose target-side ordinal configurations are sharply resolved under the Plackett–Luce relaxation. We measure the concentration of the target source-state and insertion-rank distributions using their normalized entropies. For any probability vector $\mathbf{r} = (r_1, \ldots, r_K)$, define normalized entropy

$$\mathrm{Ent}_K(\mathbf{r}) = -\frac{\sum_{i=1}^{K} r_i \log r_i}{\log K}, \qquad 0 \log 0 := 0. \tag{41}$$

The target ordinal-concentration gate for transition t is

$$I_t = \left[1 - \frac{1}{2}\left(\mathrm{Ent}_{120}(p_t) + \mathrm{Ent}_5(\iota_{t+1})\right)\right]_{[0,1]}, \tag{42}$$

where $[x]_{[0,1]} = \min\{1, \max\{0, x\}\}$ denotes clipping. The gate approaches one when the target source-state and insertion-rank distributions are concentrated, indicating a clearly resolved local ordinal configuration under the chosen relaxation, and decreases when these distributions are diffuse. All target-derived quantities $(s_c, p_t, \iota_{t+1}, I_t)$ are detached. The gate therefore acts as a fixed target-side concentration weight for each transition, while gradients propagate exclusively through the predicted ordinal states $(\hat{p}_t, \hat{\iota}_{t+1})$ to $\widehat{\boldsymbol{Y}}$.

A horizon of length H contains $H-4$ five-point windows and hence $H-5$ adjacent-window transitions. Restoring the suppressed variable index, the ordinal-flow loss is

$$\mathcal{L}_{\mathrm{ord}} = \frac{1}{(H-5)C} \sum_{t=1}^{H-5} \sum_{c=1}^{C} I_{t,c} D_{t,c}. \tag{43}$$

A.4 Unified objective and optimization

The two branches encode complementary path information: the cardinal field is sensitive to metric distortion, whereas the ordinal flow captures order-preserving and order-changing local structure independently of absolute trajectory magnitude. Using the branch coefficient $\lambda \geq 0$ introduced in Section A.1, the final HyOT objective is

$$\mathcal{L}_{\mathrm{HyOT}} = \mathcal{L}_{\mathrm{res}} + \lambda \mathcal{L}_{\mathrm{ord}}. \tag{44}$$

The controls follow the two-branch structure. In the cardinal branch, (σ, ν, w) determine the residual scale, Student-tail shape, and robust-score contribution, with Theorem 2 providing the admissible range of w. The coefficient ρ controls residual innovations while Proposition 3 preserves aggregate local trace curvature. In the ordinal branch, τ determines ranking softness. Finally, λ combines the two branch objectives. Their numerical values are specified in the experimental protocol.

Both branches operate on the same predicted trajectory $\widehat{\boldsymbol{Y}}$ and are optimized jointly through the original forecasting backbone. Training minimizes the empirical average of equation 44 over forecasting instances. The resulting objective directly couples cardinal residual geometry with ordinal transition dynamics in a single training step.

HyOT is used exclusively as a training objective. At inference, the learned backbone retains its original forecasting architecture and prediction procedure. For the fixed five-point ordinal alphabet, the additional training complexity per forecast path is $\mathcal{O}(CH)$.

B Proofs

B.1 Proof of Proposition 1

Write $\beta = (\nu+3)/2$. Around $e=0$, equation 29 has the expansion

$$H_{\mathrm{St}}(e;\sigma,\nu) = \frac{\nu+1}{a} \frac{-1+\beta(e^2/a)}{(1+e^2/a)^2} \tag{45}$$

$$= -\frac{\nu+1}{a} + \frac{(\nu+1)(\nu+7)}{2a^2} e^2 + \mathcal{O}(e^4). \tag{46}$$

Hence $H'_{\mathrm{St}}(0)=0$ and $H''_{\mathrm{St}}(0) = (\nu+1)(\nu+7)/a^2$. Multiplication by $\kappa_{\nu,\sigma} = 2a^2/[(\nu+1)(\nu+7)]$ gives $(\kappa_{\nu,\sigma} H_{\mathrm{St}})''(0) = 2$. The two members of equation 30 therefore have zero first derivative and second derivative 2 at the origin. Their normalized positive mixture has the same derivatives. Because $\phi_{\sigma,\nu,w}$ is smooth and even near zero, Taylor expansion gives $\phi_{\sigma,\nu,w}(e) = \phi_{\sigma,\nu,w}(0) + e^2 + \mathcal{O}(e^4)$.

B.2 Proof of Theorem 2

Let $u = e^2/a$, $\alpha_\nu = (\nu+3)/(\nu+7)$, and $r_\nu(u) = (1-\alpha_\nu u)/(1+u)^3$. Differentiating equation 29 and applying the curvature scale gives

$$\kappa_{\nu,\sigma} H'_{\mathrm{St}}(e;\sigma,\nu) = 2e \frac{1-\alpha_\nu u}{(1+u)^3} = 2e r_\nu(u), \tag{47}$$

and hence

$$\phi'_{\sigma,\nu,w}(e) = \frac{2e}{1+w} \left[1 + w r_\nu(u)\right]. \tag{48}$$

For $u \geq 0$,

$$r'_\nu(u) = \frac{2\alpha_\nu u - (\alpha_\nu + 3)}{(1+u)^4}. \tag{49}$$

The unique interior minimum is attained at $u_\star = (\alpha_\nu + 3)/(2\alpha_\nu)$, with

$$r_\nu(u_\star) = -\frac{4\alpha_\nu^3}{27(\alpha_\nu + 1)^2} = -\frac{(\nu+3)^3}{27(\nu+7)(\nu+5)^2} = -\frac{1}{w_{\max}(\nu)}. \tag{50}$$

Thus $1 + wr_\nu(u) > 0$ for all $u \geq 0$ whenever $0 \leq w < w_{\max}(\nu)$. Also $r_\nu(u) \leq 1$ for $u \geq 0$, so the displayed derivative yields $0 < \phi'_{\sigma,\nu,w}(e)/(2e) \leq 1$. Since $r_\nu(u) \to 0$, the ratio converges to $1/(1+w)$ as $|e| \to \infty$. Finally, $\phi_{\sigma,\nu,w}(e) = e^2/(1+w) + \mathcal{O}(e^{-2})$ in the tails, so the score is coercive; strict direction and coercivity make zero its unique global minimizer.

B.3 Proof of Proposition 3

Suppress the variable index. Let $\mathbf{e} \in \mathbb{R}^H$ be the residual path, let $\mathbf{G} \in \mathbb{R}^{(H-1)\times H}$ be the first-difference matrix from Proposition 3, and let $\mathbf{I}_H \in \mathbb{R}^{H\times H}$ be the identity matrix. At $\mathbf{e} = \mathbf{0}$, Proposition 1 gives

$$\nabla^2_{\mathbf{e}}\ell_{\text{level}} = \frac{2}{H}\mathbf{I}_H, \qquad \nabla^2_{\mathbf{e}}\ell_{\text{innov}} = \frac{2}{H-1}\mathbf{G}^\top\mathbf{G}, \tag{51}$$

Therefore

$$\operatorname{tr}\nabla^2_{\mathbf{e}}\ell_{\text{level}} = 2. \tag{52}$$

Every row of $\mathbf{G}$ contains one -1 and one $+1$, so $\operatorname{tr}(\mathbf{G}^\top\mathbf{G}) = \|\mathbf{G}\|_F^2 = 2(H-1)$, where $\|\cdot\|_F$ is the Frobenius norm. It follows that

$$\operatorname{tr}\nabla^2_{\mathbf{e}}\ell_{\text{innov}} = \frac{2}{H-1}\,2(H-1) = 4. \tag{53}$$

Applying equation 34,

$$\operatorname{tr}\nabla^2_{\mathbf{e}}\ell_{\text{res}} = \frac{2+4\rho}{1+2\rho} = 2, \tag{54}$$

which equals the trace of the Hessian of $H^{-1}\|\mathbf{e}\|_2^2$, where $\|\cdot\|_2$ is the Euclidean norm.

B.4 Proof of Theorem 4

Let π_t be the total order of the labeled source-window elements $(v_t, \ldots, v_{t+4})$ in equation 38. Deleting the departing element v_t from π_t produces one and only one total order of the retained elements $(v_{t+1}, \ldots, v_{t+4})$. For each $j \in \{1, \ldots, 5\}$, inserting the entering element v_{t+5} at position j in that retained order produces one total order of the successor window and therefore one legal successor. Different insertion positions produce different successors, so every source has five legal outgoing transitions.

Conversely, a legal adjacent-window pair exposes its source order π_t directly, and the successor order exposes the unique rank j of v_{t+5}. The maps are inverses. The transition representation is therefore bijective and has cardinality $|\mathfrak{S}_5| \times 5 = 120 \times 5 = 600$.

B.5 Proof of Proposition 5

Using equation 40, the Bhattacharyya affinity between target and prediction transition descriptors is

$$\mathrm{BC}(T_t, \widehat{T}_t) = \sum_{k=1}^{120}\sum_{j=1}^{5}\sqrt{p_t(k)\iota_{t+1}(j)\hat{p}_t(k)\hat{\iota}_{t+1}(j)} \tag{55}$$

$$= \left(\sum_{k=1}^{120}\sqrt{p_t(k)\hat{p}_t(k)}\right)\left(\sum_{j=1}^{5}\sqrt{\iota_{t+1}(j)\hat{\iota}_{t+1}(j)}\right) \tag{56}$$

$$= A_{\text{src},t}A_{\text{ins},t}. \tag{57}$$

Substitution into $d_{\mathrm{H}}^2 = 1 - \mathrm{BC}$ yields equation 21.