Conditionally Identifiable Latent Representation for Multivariate Time Series with Structural Dynamics

Minkey Chang, Jae Young Kim[1]

[1]Seoul National University

Abstract

We propose the Identifiable Variational Dynamic Factor Model (iVDFM) for multivariate time series. The model combines variational inference with *partial* identifiability guarantees up to a known ambiguity class. Innovations follow a conditional exponential-family prior over observed auxiliary context and regime embeddings. Linear diagonal dynamics map innovations to factors and preserve that class, so factors are partially identifiable up to permutation and component-wise affine maps. We train by maximizing the ELBO and estimate uncertainty in both latent trajectories and forecasts. Under explicit assumptions, we prove partial identifiability up to the stated class and provide the full proof in the appendix. We evaluate factor recovery on synthetic DGPs, intervention behavior on synthetic SCMs, and forecasting on real benchmarks with CRPS and MSE.

1 Introduction

Time-series modeling often requires assigning uncertainty and making downstream intervention queries in a latent space. When latent coordinates are not identifiable up to a controlled ambiguity class, the same observations admit multiple latent factorizations with different coordinate semantics, so uncertainty and intervention statements become ill-defined. Our goal is therefore to learn latent factors whose meaning is stable up to a known, small ambiguity class of transformations.

Classical dynamic factor models recover temporal structure but are identifiable only up to arbitrary invertible linear mixings of factors (Stock and Watson, 2002). Identifiable VAEs (iVAEs) reduce ambiguity in static settings to permutation and component-wise affine maps by conditioning the latent prior on auxiliary context (Khemakhem et al., 2020), but they do not directly target sequential models. We propose the *Identifiable Variational Dynamic Factor Model (iVDFM)*, which shifts the identification target to the *innovation process* that drives the dynamics: we impose a conditional non-Gaussian exponential-family prior on innovations and propagate them to factors through linear diagonal dynamics. Under explicit assumptions, this yields *partial* identifiability of both innovations and factor trajectories up to permutation and component-wise affine maps.

iVDFM is the dynamic-factor counterpart of the iVAE construction of Khemakhem et al. (2020): innovations $\boldsymbol{\eta}_{1:T}$ take the role of their sources $\mathbf{z}$; the pair $(\mathbf{u}_t, \mathbf{e}_t)$, with $\mathbf{e}_t = \psi(\mathbf{u}_t)$ a deterministic regime embedding, takes the role of the auxiliary $\mathbf{u}$; and the temporal observation map $\boldsymbol{y}_t = g(\mathcal{F}_t(\boldsymbol{\eta}_{1:t})) + \boldsymbol{\varepsilon}_t$, composing the learned dynamics $\mathcal{F}_t$ with a decoder g, replaces the static mixing $\mathbf{x} = \mathbf{f}(\mathbf{z})$. Innovation-level identifiability then follows from Khemakhem et al. (2020, Theorem 1) in this correspondence; the step we add is carrying the ambiguity class $\mathcal{T}$ through the dynamics to factor trajectories, and assembling a dynamic factor model around this core.

The technical content of that carry-through is a propagation lemma (Lemma 2): under diagonal or block-diagonal companion-form dynamics, $\mathcal{T}$ transfers from innovations to factor trajectories, and the same implication fails under dense dynamics or Gaussian innovations. On the inference side, the ELBO factorizes along the Markov structure and training proceeds on sliding windows, which keeps identification (a property of the generative prior) separate from the variational approximation (a property of the encoder). The regime embedding is a deterministic mixture of the auxiliary, which extends the conditional innovation family while preserving exogeneity. Factor recovery on synthetic dynamic DGPs, intervention analysis on synthetic SCMs, and probabilistic forecasting on real benchmarks then show that innovation-level partial identifiability is compatible with competitive predictive performance.

2 Related Work

Latent-variable models are typically identifiable only up to an *ambiguity class* of transformations that leave the observation law unchanged; the practical question is how small that class can be made. Classical methods such as PCA (Pearson, 1901) and ICA (Hyvarinen and Oja, 2000) recover structured directions under strong covariance or independence assumptions that do not extend cleanly to nonlinear or noisy settings. VAEs (Kingma and Welling, 2014) relax those assumptions but in exchange leave the latent coordinates broadly non-unique, and identifiable VAEs (Khemakhem et al., 2020) recover a small class in the static case by conditioning the latent prior on auxiliary variables.

For multivariate time series, classical dynamic factor models (DFMs) (Stock and Watson, 2002) summarize many observed variables with linear latent dynamics, but factors are identified only up to arbitrary invertible linear mixings (a rotational indeterminacy in the Gaussian case; see Appendix B). Structural VAR approaches (Sims, 1980; Stock and Watson, 2018) identify interpretable shocks via restrictions or instruments, though they rely on linear assumptions.

Recent deep sequence models such as iTransformer (Liu et al., 2023), TimeMixer (Wang et al., 2024a), S4 (Gu et al., 2022), and Mamba (Gu and Dao, 2023) emphasize predictive accuracy more than latent identifiability. Deep Dynamic Factor Models (Andreini et al., 2020) and Deep Kalman Filters (Krishnan et al., 2015) combine probabilistic temporal structure with nonlinear mappings, but they typically do not provide a narrow latent ambiguity class. Work on auxiliary-conditioned sequential identifiability (Hyvarinen et al., 2019; Balsells-Rodas et al., 2024) shows how temporal structure and auxiliary variation can shrink ambiguity under specific assumptions.

A useful lens on this landscape is the size of the residual ambiguity group. Gaussian DFMs and linear state-space models are identified only up to the full general linear group of invertible $r \times r$ real matrices, $\mathrm{GL}(r)$, acting on factors; Gaussian linear dynamical systems (LDSs) stay inside this group because the Gaussian family is closed under invertible linear maps (Appendix B). Static iVAEs and their non-Gaussian extensions (Khemakhem et al., 2020; Hyvarinen et al., 2019) shrink the group to the much smaller class $\mathcal{T}$ of permutations and component-wise affine maps, contingent on assumptions about the auxiliary variable. iVDFM inherits $\mathcal{T}$ at the innovation level and, via Lemma 2, preserves $\mathcal{T}$ at the factor level through diagonal dynamics, trading the Gaussian closure and dense dynamics of classical DFMs for non-Gaussian innovations and component-wise (or block-diagonal companion) propagation. Switching-state identifiable models (Balsells-Rodas et al., 2024) also reach a sub-$\mathrm{GL}(r)$ ambiguity, but through discrete regime switches; iVDFM instead uses a deterministic regime mixture that preserves exogeneity of the auxiliary variable without committing to hard transitions.

From a causal perspective, this matters because intervention semantics depend on stable latent coordinates. If latent factors are defined only up to a broad class, causal conclusions can vary across equivalent parameterizations (Scholkopf et al., 2021; Spirtes et al., 2000). Identifiability up to a known and controlled class is therefore a prerequisite for well-defined intervention queries (Pearl, 2009; Peters et al., 2017).

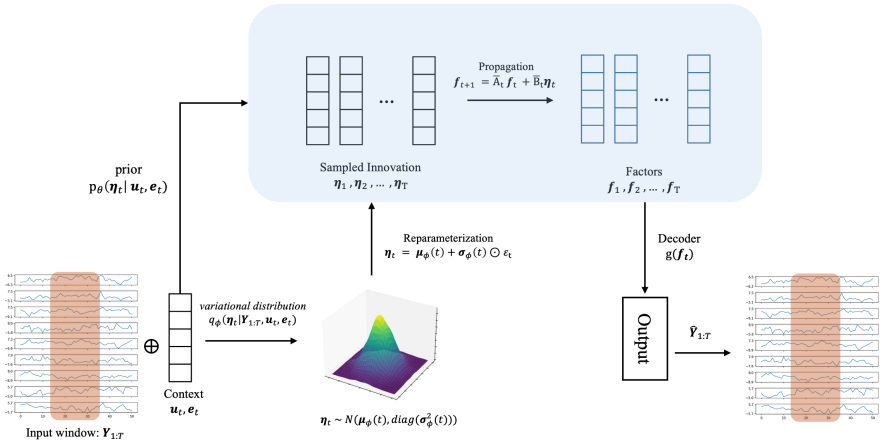


Figure 1: iVDFM in three steps. (a) Extraction. Observations $\boldsymbol{y}_{1:T}$ and context $(\mathbf{u}_t, \mathbf{e}_t)$ are mapped by the encoder to a variational distribution $q_\phi(\boldsymbol{\eta}_t \mid \cdot)$ over innovations, while the prior $p_\theta(\boldsymbol{\eta}_t \mid \mathbf{u}_t, \mathbf{e}_t)$ conditions on the same context. (b) Propagation. Innovations $\boldsymbol{\eta}_t$ drive factors $\boldsymbol{f}_t$ through diagonal dynamics $\bar{A}_t$, $\bar{B}_t$. (c) Forecasting and generation. Factors are decoded by g to reconstructed or forecast observations $\widehat{\boldsymbol{y}}_t$.

3 Identifiable Variational Dynamic Factor Model

The *Identifiable Variational Dynamic Factor Model (iVDFM)* concentrates all latent stochasticity in innovations $\boldsymbol{\eta}_t$ drawn from a conditional prior and propagates them to factors through diagonal time-varying linear dynamics, with parameters estimated by variational inference (Figure 1). Diagonality makes factor trajectories deterministic functions of the innovation history, so we can analyze identifiability at the innovation level and carry it over to factors.

The generative model then reduces to three equations,

$$\begin{aligned}
\boldsymbol{\eta}_t &\sim p_\theta(\cdot \mid \mathbf{u}_t, \mathbf{e}_t) && \text{(conditional exponential-family innovation prior)},\\
\boldsymbol{f}_{t+1} &= \bar{A}_t \boldsymbol{f}_t + \bar{B}_t \boldsymbol{\eta}_t && \text{(diagonal or block-diagonal companion-form dynamics)},\\
\boldsymbol{y}_t &= g(\boldsymbol{f}_t) + \boldsymbol{\varepsilon}_t && \text{(injective MLP decoder with fixed-scale noise)},
\end{aligned}$$

whose components we describe before stating the identifiability result (Theorem 3).

3.1 Innovation

The innovation $\boldsymbol{\eta}_t \in \mathbb{R}^r$ carries a conditional non-Gaussian exponential-family prior that factorizes across components,

$$p(\boldsymbol{\eta}_t \mid \mathbf{u}_t, \mathbf{e}_t) = \prod_{i=1}^{r} h_i(\eta_{i,t}) \exp\Big(\mathbf{T}_i(\eta_{i,t})^\top \boldsymbol{\lambda}_i(\mathbf{u}_t, \mathbf{e}_t) - \Psi_i(\boldsymbol{\lambda}_i(\mathbf{u}_t, \mathbf{e}_t))\Big), \tag{1}$$

with sufficient statistics $\mathbf{T}_i \in \mathbb{R}^k$, natural parameters $\boldsymbol{\lambda}_i(\mathbf{u}_t, \mathbf{e}_t) \in \mathbb{R}^k$, base measure h_i, and log-partition Ψ_i. Attaching identifiability to this prior, rather than to factors directly, keeps a clean separation between what is identified (innovation coordinates) and how it is used (propagation and decoding): sufficient variation of the natural-parameter map $\boldsymbol{\lambda}(\mathbf{u}_t, \mathbf{e}_t)$ across the auxiliary $(\mathbf{u}_t, \mathbf{e}_t)$

identifies innovations up to permutation and component-wise affine maps (Hyvarinen and Oja, 2000; Khemakhem et al., 2020).

The regime context $\mathbf{e}_t$ is a deterministic soft mixture,

$$\mathbf{e}_t = \sum_{j=1}^{K} \pi_{t,j} \mathbf{e}_j, \qquad \boldsymbol{\pi}_t = \mathrm{softmax}(\mathrm{RegimeNet}(\mathbf{u}_t)),$$

over learnable embeddings $\mathbf{e}_j \in \mathbb{R}^d$ produced by a small MLP; the pair $(\mathbf{u}_t, \mathbf{e}_t)$ plays the role of the iVAE auxiliary, and taking the expected embedding rather than sampling a discrete regime keeps $\mathbf{e}_t$ a deterministic function of $\mathbf{u}_t$. The induced prior $\sum_j \pi_{t,j}\, p_j(\boldsymbol{\eta} \mid \mathbf{u}_t)$ with regime-specific natural parameters $\boldsymbol{\lambda}_j(\mathbf{u}_t) = \boldsymbol{\lambda}(\mathbf{u}_t, \mathbf{e}_j)$ enlarges the family of conditional innovation distributions reachable by a single $\boldsymbol{\lambda}(\mathbf{u}_t)$. Only the weighted combination of the regime parameters $(\mathbf{e}_j, A_j, B_j)$ enters the observation law, so individual regime parameters are not separately resolved; the identification guarantee applies to innovations and factors, not to the components of the mixture. The determinism requirement extends to any auxiliary one might plug in: calendar features, known event indicators, and similar domain metadata fold into $\mathbf{u}_t$ without issue, while stochastic context or variables computed from $\boldsymbol{y}_t$ forfeit exogeneity.

Identifiability at the innovation level rests on how much $\boldsymbol{\lambda}(\mathbf{u}, \mathbf{e})$ varies with the auxiliary. The natural-parameter map is required to have full column rank at $rk + 1$ distinct auxiliary values (Khemakhem et al., 2020); for scalar sufficient statistics (k=1, as in the Laplace innovations we use) this reduces to $r + 1$. Richer auxiliary variation sharpens the conclusion, while weaker variation leaves room for distinct parameterizations to induce the same observation law.

3.2 Dynamics

Factors propagate from innovations through diagonal dynamics,

$$\boldsymbol{f}_{t+1} = \bar{A}_t \boldsymbol{f}_t + \bar{B}_t \boldsymbol{\eta}_t, \tag{2}$$

with diagonal $\bar{A}_t, \bar{B}_t \in \mathbb{R}^{r \times r}$ formed by the same regime weights, $\bar{A}_t = \sum_j \pi_{t,j} A_j$ and $\bar{B}_t = \sum_j \pi_{t,j} B_j$, each A_j, B_j diagonal. Each factor coordinate then depends only on its own history and its own innovation, so the innovation-level ambiguity class $\mathcal{T}$ transfers unchanged to factor trajectories. AR(p) dependence fits into the same picture via the standard companion-form stacking (Lütkepohl, 2005; Hamilton, 1994), which preserves block-diagonality across factor coordinates and hence the same component-wise propagation (Appendix A.2).

Definition 1 (Identifiability up to a class)**.** *Innovations (or factors) are* identifiable up to a class of transformations $\mathcal{T}$ *if, for any two parameterizations that induce the same $p(\boldsymbol{y}_{1:T}, \boldsymbol{u}_{1:T})$, the corresponding latent variables are related by a transformation in $\mathcal{T}$ almost surely.*

Concretely, any $\tau \in \mathcal{T}$ acts as $\tilde{\eta}_{i,t} = a_{\pi(i)}\, \eta_{\pi(i),t} + b_{\pi(i)}$ for a permutation π and component-wise scalars $a_i \neq 0$, b_i; there is no cross-coordinate mixing. So the recovered coordinates name the same directions up to relabeling, sign, shift, and per-coordinate scale, which is what lets intervention and uncertainty statements remain stable under equivalent parameterizations. This is the qualitative gap from rotationally ambiguous Gaussian factor models, where actions and explanations sit in a coordinate system with a much larger ambiguity group.

Lemma 2 (Propagation of $\mathcal{T}$ through diagonal dynamics)**.** *Let $\boldsymbol{f}_t = \mathcal{F}_t(\boldsymbol{\eta}_{1:t})$ denote the deterministic innovation-to-factor map induced by the diagonal dynamics in Eq. 2 or the block-diagonal companion form. If innovations are identified up to $\mathcal{T}$, then so are factor trajectories: any reparameterization $\tilde{\boldsymbol{\eta}}_t = \tau(\boldsymbol{\eta}_t)$ with $\tau \in \mathcal{T}$ induces a factor reparameterization $\tilde{\boldsymbol{f}}_t = \bar{\tau}_t(\boldsymbol{f}_t)$ with $\bar{\tau}_t \in \mathcal{T}$ for all t.*

The proof is in Appendix A.3 (Step 2); the key property is that $\mathcal{T}$ is preserved, not enlarged, when innovations pass through the dynamics. The same argument breaks down under dense (non-diagonal) dynamics or Gaussian innovations, where additional symmetries in the induced observation law enlarge the ambiguity group (Appendix B).

Theorem 3 (Identifiability of dynamic factors)**.** *Consider the generative model of this section (innovation prior Eq. 1, diagonal dynamics Eq. 2 or block-diagonal companion form, observation model with injective g and fixed noise). Assume:*

1. *The innovation prior is conditional exponential-family with linearly independent sufficient statistics $\mathbf{T}_i \in \mathbb{R}^k$, and the natural-parameter map $\boldsymbol{\lambda}(\mathbf{u}, \mathbf{e})$ has full column rank for at least $rk + 1$ distinct values of $(\mathbf{u}, \mathbf{e})$; with scalar sufficient statistics ($k = 1$) this reduces to $r + 1$.*
2. *$\boldsymbol{\pi}_t$ depends only on $\mathbf{u}_t$, with diagonal A_j, B_j per regime.*
3. *The decoder g is injective on the support of $\boldsymbol{f}_t$.*
4. *Observation noise has full support and σ^2 is fixed.*
5. *Base measures have full support, $r \leq N$, and all mappings are measurable and integrable.*

Then the innovations $\{\boldsymbol{\eta}_t\}_{t=1}^T$ are partially *identifiable from $(\boldsymbol{y}_{1:T}, \boldsymbol{u}_{1:T})$ up to $\mathcal{T}$ (Definition 1, with $\mathcal{T}$ the class of permutation and component-wise affine maps). Combined with Lemma 2, the latent factor trajectories $\{\boldsymbol{f}_t\}_{t=1}^T$ and the associated dynamic parameters are partially identifiable up to the same class.*

The innovation-level part follows from Khemakhem et al. (2020) applied with auxiliary $(\mathbf{u}_t, \mathbf{e}_t)$; Lemma 2 carries the conclusion to factor trajectories. The full proof is in Appendix A.

3.3 Estimation

Observations $\boldsymbol{y}_t \in \mathbb{R}^N$ arise from factors through an injective MLP decoder $g : \mathbb{R}^r \to \mathbb{R}^N$ with fixed-scale additive noise, $\boldsymbol{y}_t = g(\boldsymbol{f}_t) + \boldsymbol{\varepsilon}_t$. The posterior over innovations is approximated by a component-factorized Gaussian encoder,

$$q_\phi(\boldsymbol{\eta}_t \mid \boldsymbol{y}_t, \mathbf{u}_t) = \mathcal{N}\big(\boldsymbol{\mu}_\phi(t), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(t))\big)\,,$$

which does not need $\mathbf{e}_t$ as an explicit input because it is already a deterministic function of $\mathbf{u}_t$. The encoder sees a single time step: identifiability is a property of the generative prior, so a longer-context encoder would not shrink $\mathcal{T}$, and restricting the context reduces the chance that the encoder routes information through $\mathbf{u}_t$ at the expense of $\boldsymbol{y}_t$. Sampling uses the standard reparameterization (Kingma and Welling, 2014), $\boldsymbol{\eta}_t = \boldsymbol{\mu}_\phi(t) + \boldsymbol{\sigma}_\phi(t) \odot \boldsymbol{\epsilon}_t$ with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, I)$, and sampled innovations are propagated through the dynamics to form factors.

We train by maximizing the evidence lower bound (ELBO) jointly over generative parameters θ (decoder, prior including $\{\mathbf{e}_j\}$ and RegimeNet, dynamics) and encoder parameters ϕ. Under the mean-field variational distribution $q_\phi(\boldsymbol{\eta}_{1:T} \mid \boldsymbol{y}_{1:T}, \mathbf{u}_{1:T}) = \prod_{t=1}^T q_\phi(\boldsymbol{\eta}_t \mid \boldsymbol{y}_t, \mathbf{u}_t)$ and with factors deterministic given innovations, the bound factorizes across time:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{q_\phi(\boldsymbol{\eta}_{1:T} \mid \cdot)}\left[\sum_{t=1}^T \log p_\theta(\boldsymbol{y}_t \mid \mathcal{F}_t(\boldsymbol{\eta}_{1:t}))\right] \\ - \sum_{t=1}^T \mathrm{KL}\big(q_\phi(\boldsymbol{\eta}_t \mid \boldsymbol{y}_t, \mathbf{u}_t) \,\big\|\, p_\theta(\boldsymbol{\eta}_t \mid \mathbf{u}_t, \mathbf{e}_t)\big)\,,\end{aligned} \tag{3}$$

with $\boldsymbol{f}_t = \mathcal{F}_t(\boldsymbol{\eta}_{1:t})$ unrolled from $\boldsymbol{f}_{t+1} = \bar{A}_t \boldsymbol{f}_t + \bar{B}_t \boldsymbol{\eta}_t$ and a learned initial state $\boldsymbol{f}_0$. We optimize $\mathcal{L}$ on sliding windows of length T with stride $T/2$ (Appendix E.2); partial windows at the sequence boundary are dropped rather than zero-padded, so the dynamics are never unrolled on artificial context. Windowed estimation is a finite-data surrogate; Theorem 3 characterizes the invariance class of the underlying generative family, so neither the windowing scheme nor local adaptation through $(\mathbf{u}_t, \mathbf{e}_t)$ enters the identifiability argument.

4 Experiments

Our experiments probe three questions raised by the identifiability claim: whether iVDFM actually recovers latent factors on synthetic data with known ground truth, whether the learned innovation coordinates support *do*-interventions on a known structural causal model, and whether enforcing partial identifiability compromises forecasting performance on real benchmarks. Factor recovery compares against DFM, DDFM, VAE, and iVAE; forecasting adds TimeMixer and related deep baselines. Code, configurations, and protocols are in Appendix C.

4.1 Factor Recovery

We split the evaluation into *dynamic* data-generating processes (DGPs; AR dynamics driven by innovations) and *static* DGPs, aligning recovered factors to ground truth via permutation-invariant maximum-correlation matching and reporting the mean correlation coefficient (MCC), subspace distance, smoothness, and Trace R^2 (Tables 1–2). All methods see the same settings (T=200, N=20, r=5) with matched simulation budgets, and we report mean ± standard deviation across 10 seeds so the static-versus-dynamic contrast is not a single-seed artefact.

Table 1: Factor recovery on the dynamic DGP (T=200, N=20, r=5). Metrics are MCC, subspace distance, smoothness, and Trace R^2 after MCC-based matching. Higher values are better for MCC and Trace R^2, while lower values are better for subspace distance and smoothness. Reported values are mean ± std over simulations.

Model	MCC ↑	Subspace ↓	Smoothness ↓	Trace R^2 ↑
iVDFM	**0.6548 ± 0.0549**	**0.4345 ± 0.1022**	1.7392 ± 0.1118	**0.8202 ± 0.0603**
DDFM	0.5202 ± 0.0360	0.8441 ± 0.0564	2.0060 ± 0.1068	0.4998 ± 0.0435
iVAE	0.6146 ± 0.0560	0.5149 ± 0.1264	1.5864 ± 0.0710	0.7664 ± 0.0871
VAE	0.6065 ± 0.0587	0.5418 ± 0.1286	1.5844 ± 0.0741	0.7526 ± 0.0838
DFM	0.4681 ± 0.0310	0.9565 ± 0.0414	**0.3149 ± 0.0566**	0.3783 ± 0.0324

Table 2: Factor recovery, static DGP (T=200, N=20, r=5). Metrics are computed after MCC-based matching; we report MCC, subspace distance, smoothness, and Trace R^2 (mean ± std over simulations).

Model	MCC ↑	Subspace ↓	Smoothness ↓	Trace R^2 ↑
iVDFM	0.5683 ± 0.0661	0.6949 ± 0.0544	1.7930 ± 0.1281	0.6178 ± 0.0460
DDFM	0.5989 ± 0.0673	0.6147 ± 0.0540	2.3296 ± 0.0911	0.6745 ± 0.0502
iVAE	0.5678 ± 0.0751	0.5055 ± 0.0953	2.2504 ± 0.1186	0.7766 ± 0.0589
VAE	**0.6405 ± 0.0945**	**0.4520 ± 0.0675**	2.3255 ± 0.1438	**0.8078 ± 0.0472**
DFM	0.3374 ± 0.0397	1.1192 ± 0.0448	**0.0994 ± 0.0173**	0.2476 ± 0.0446

On the dynamic DGP, iVDFM attains the highest MCC and Trace R^2 (MCC 0.6548 ± 0.0549; Trace R^2 0.8202 ± 0.0603), with the gain visible at both coordinate alignment (MCC) and trajectory-level

Table 4: CRPS and MSE (standardized), averaged over horizons. Rows: dataset; columns: per-model CRPS and MSE (std.).

Dataset	iVDFM		iTrans.		TimeMix		TimeXer		DDFM	
	CRPS	MSE	CRPS	MSE	CRPS	MSE	CRPS	MSE	CRPS	MSE
ETTh1	0.282	1.258	0.693	1.296	0.276	0.801	0.542	1.189	0.334	1.342
ETTh2	0.268	1.016	0.501	1.018	0.230	0.401	0.452	0.990	0.278	1.013
ETTm1	0.315	1.221	0.706	1.288	0.229	0.547	0.598	1.250	0.356	1.277
ETTm2	0.340	1.098	0.644	1.330	0.232	0.434	0.531	1.395	0.460	2.308
Weather	0.271	1.184	0.316	0.703	0.163	0.341	0.281	0.707	0.164	0.439

explanatory structure (Trace R^2). On the static DGP, VAE achieves the best MCC and Trace R^2 (0.6405 ± 0.0945; 0.8078 ± 0.0472); iVDFM matches iVAE on MCC (0.5683 vs. 0.5678) and retains stronger smoothness than the other nonlinear baselines. The gap between the two settings is where the identification signal lives: temporal innovation structure is what iVDFM exploits, so the dynamic DGP is where the mechanism is active, while on the static DGP there is nothing for the dynamic side of the model to add.

4.2 Causal Intervention

Table 3: Causal intervention on synthetic SCMs (iVDFM). IRF-MSE, IRF-MAE: mean squared / absolute error between ground-truth and model-implied impulse response; Sign Acc: fraction of (time step, variable) pairs with correct sign; IRF Corr: Pearson correlation with ground truth. Mean $\pm$ std over five simulations.

SCM	IRF-MSE ↓	IRF-MAE ↓	Sign Acc ↑	IRF Corr ↑
Base (linear)	0.89 ± 1.01	0.45 ± 0.23	0.58 ± 0.25	0.21 ± 0.75
Regime	1.28 ± 0.92	0.77 ± 0.27	0.60 ± 0.27	0.13 ± 0.62
Chain	2.18 ± 2.40	0.97 ± 0.40	0.58 ± 0.33	0.20 ± 0.62

To validate that the learned representation supports *do*-interventions at the innovation level, we generate synthetic series from a structural causal model (SCM), fit iVDFM on observational samples, and then apply *do* by fixing one innovation component at a target time and propagating its effect through the learned dynamics and decoder. We compare the model-implied impulse response functions (IRFs) with ground-truth IRFs using IRF-MSE, sign accuracy, and IRF correlation (Table 3).

The chain setting is the hardest case because its dynamics violate diagonality: iVDFM keeps IRF-MSE bounded but loses sign and correlation fidelity. The degradation tracks the argument behind Lemma 2 — propagation relies on cross-factor independence — and points to richer transition structures as the natural relaxation for intervention-faithful dynamics.

4.3 Forecasting

We evaluate distributional forecasts on ETTh1/2, ETTm1/2, and Weather at horizons 96–720 using the continuous ranked probability score (CRPS; Matheson and Winkler, 1976; Gneiting and Raftery, 2007) and standardized MSE with rolling origins (Table 4).

Factor order and regime count are tuned by validation forecast loss. Across datasets iVDFM's CRPS and standardized MSE sit alongside several strong baselines; the reading we draw from this is that the innovation-level constraints used to obtain partial identifiability are compatible with competitive

distributional forecasting, while the latent space continues to carry the controlled ambiguity $\mathcal{T}$ used in the recovery and intervention analyses above.

5 Application: Exchange Rates and Epidemiology

As qualitative case studies we fit iVDFM to two real panels of multivariate time series. In both cases observations $\boldsymbol{y}_t$ are the stacked series (currencies or surveillance measurements), the auxiliary $\mathbf{u}_t$ carries calendar encodings and standardized trend features, and we interpret the learned loadings on observed variables. Coordinates are only identified up to $\mathcal{T}$, so the useful output is the *structure* of the loadings rather than their raw sign or scale.

5.1 Exchange Rates

On daily exchange-rate series for AUD, JPY, CAD, CNY, and NZD against USD, a two-factor representation (r=2) is enough to separate broad co-movement from the part of the variation that is more series-specific. The second factor loads strongly and with the same sign on AUD, JPY, CAD, and NZD (roughly 0.85 to 0.94) while CNY remains weakly connected, which is consistent with a market-wide USD component concentrated on the freely moving rates in the panel. The first factor carries moderate negative correlations on the same currencies (-0.48 to -0.60) and is again weak on CNY, behaving as a secondary re-pricing direction rather than a second copy of the same common mode. Because the coordinates are tied to an innovation process with controlled ambiguity rather than to an unconstrained embedding, shifts along these trajectories can be discussed as candidate shocks up to $\mathcal{T}$. In practice, $\mathcal{T}$ preserves the separation between the market-wide direction and the secondary re-pricing direction across retraining runs, whereas a rotationally ambiguous embedding would allow those two directions to mix freely and leave shock-level interpretation ill-posed.

5.2 Epidemiology

For weekly influenza-like illness surveillance we again use two factors (r=2), but the structure of the loadings tells a different story. One latent trajectory moves closely with standardized outpatient visits and aligns strongly with both ILI rates (about -0.69, -0.71, and -0.64; Figure 2): when this coordinate shifts, all three surveillance series move together, as expected of a common epidemic-burden axis. The second coordinate is weaker on the two ILI percentages (~ -0.30) yet non-negligible on outpatient visits (~ -0.40), consistent with residual utilization-sensitive variation that the main burden axis does not fully absorb. For monitoring purposes the useful output is precisely this two-factor decomposition: one coordinate summarizes shared epidemic intensity while the other captures utilization-linked structure that a one-dimensional summary would hide. Partial identifiability is what makes this decomposition portable across retraining: the burden axis remains attachable to the ILI series up to $\mathcal{T}$, so downstream dashboards and alert rules do not need to re-learn which coordinate corresponds to which quantity after each refit.

6 Conclusion

iVDFM extends variational latent modeling to a dynamic factor setting with conditional partial identifiability: under a conditional exponential-family innovation prior and diagonal dynamics, innovations and therefore factors are identified up to a controlled ambiguity class of permutations and component-wise affine maps, with inference carried out via the ELBO and efficient AR(p) rollouts. Empirically, the model recovers factors on synthetic dynamic DGPs, matches intervention behavior on synthetic SCMs, and remains competitive on probabilistic forecasting benchmarks. Tying latent

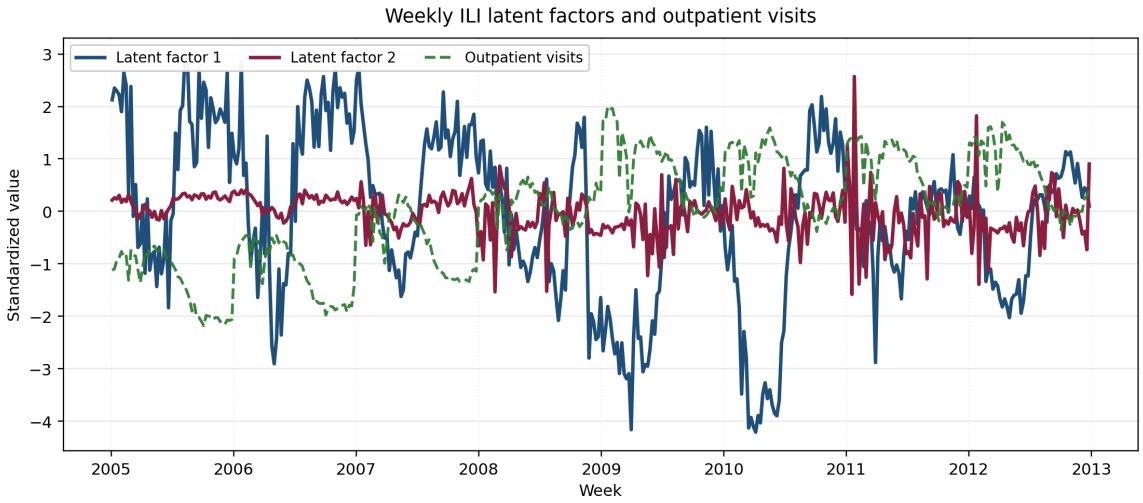


Figure 2: Recovered iVDFM factors for the weekly illness application (r=2), with standardized outpatient visits overlay.

trajectories to innovation-level dynamics with an explicit ambiguity class makes downstream analysis more stable across runs and better aligned with intervention semantics (Section 5 illustrates this on exchange-rate and epidemiology series).

A complementary design lever is the auxiliary $(\mathbf{u}_t, \mathbf{e}_t)$ itself. The theory only requires it to be a deterministic function of observed variables, which leaves room for substantially richer context than the calendar and trend features used here – for instance language-model embeddings of event text or domain reports, which are deterministic encodings of observed text and can inject much more variation into $\boldsymbol{\lambda}(\mathbf{u}_t, \mathbf{e}_t)$ than low-dimensional covariates. The caveat is that if the underlying text is itself downstream of $\boldsymbol{y}_t$, treating it as exogenous silently breaks the auxiliary assumption; in that regime promoting the variable into $\boldsymbol{y}_t$ is the cleaner route. Viewed this way, Theorem 3 gives a concrete criterion for deciding when a candidate context source should enter as auxiliary versus as observation.

The guarantees rest on fixed observation noise and diagonal dynamics, which are what keep the ambiguity group small. Relaxing these assumptions – structured cross-factor interactions in the dynamics, or heteroskedastic or heavy-tailed observation noise – is the natural next direction, at the cost of re-characterizing the residual symmetry group and sharpening the identifiability conditions. Within the current setting, the construction is already enough to support the innovation-level intervention queries of Section 4 and the panel-level interpretation of Section 5; the extensions above are what would be required when cross-factor feedback or heteroskedasticity is itself the object of study rather than a nuisance to bound.

Acknowledgments and Disclosure of Funding

We thank the reviewers for their careful reading and constructive feedback.

References

Paolo Andreini, Cosimo Izzo, and Giovanni Ricco. Deep dynamic factor models. *arXiv preprint arXiv:2007.11887*, 2020.

Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 2639–2672. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/balsells-rodas24a.html. arXiv:2305.15925.

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

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2022.

James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

Aapo Hyvarinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 2019. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

Ilyes Khemakhem, Diederik Kingma, Riccardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 2207–2217, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Rahul G. Krishnan, Uri Shalit, and David A. Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, 2017.

Bernhard Scholkopf, Dominik Janzing, Jonas Peters, Sebastian Bauer, Kun Xu, Patrick Bloebaum, Kun Zhang, and Joris Mooij. Toward causal representation learning. *Proceedings of the National Academy of Sciences*, 118(23):e2021297118, 2021.

Christopher A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

James H. Stock and Mark W. Watson. Identification and estimation of dynamic causal effects in macroeconomics using external instruments. *The Economic Journal*, 128(610):917–948, 2018.

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. *arXiv preprint arXiv:2405.14616*, 2024a.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 37, 2024b. arXiv:2402.19072.

Michael Zhang, Khaled Saab, Michael Poli, Tri Dao, Karan Goel, and Christopher Ré. Effectively modeling time series with simple discrete state spaces. In *International Conference on Learning Representations*, 2023.

This supplementary material contains proofs (identifiability and degeneracy), experiment details, additional results, and ablation studies.

Appendix A. Identifiability Proof

This appendix gives the full proof of Theorem 3 (Section 3). We use the same generative model, Definition 1, and notation as in the main text.

A.1 Setup and Notation

The generative model includes regime from the outset. Innovations follow the conditional prior in Eq. 1, where auxiliary $\mathbf{u}_t$ is observed and regime embedding $\mathbf{e}_t$ is deterministic given $\mathbf{u}_t$. Dynamics follow $\boldsymbol{f}_{t+1} = \bar{A}_t \boldsymbol{f}_t + \bar{B}_t \boldsymbol{\eta}_t$ with diagonal $\bar{A}_t$, $\bar{B}_t$, or the equivalent block-diagonal companion form. Observations satisfy $\boldsymbol{y}_t = g(\boldsymbol{f}_t) + \boldsymbol{\varepsilon}_t$, with $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)$, injective g, and fixed σ^2.

We say two parameterizations $(\theta, \boldsymbol{\eta}, \boldsymbol{f})$ and $(\tilde{\theta}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{f}})$ *induce the same observation distribution* if $p_\theta(\boldsymbol{y}_{1:T}, \boldsymbol{u}_{1:T}) = p_{\tilde{\theta}}(\boldsymbol{y}_{1:T}, \boldsymbol{u}_{1:T})$ as densities (or equivalently, the induced laws agree). By the chain rule and the fact that $\boldsymbol{f}_{1:T}$ is a deterministic function of $\boldsymbol{\eta}_{1:T}$ under the dynamics, the joint factorizes as

$$p(\boldsymbol{y}_{1:T}, \boldsymbol{u}_{1:T}, \boldsymbol{\eta}_{1:T}, \boldsymbol{f}_{1:T}) = p(\boldsymbol{u}_{1:T}) \prod_{t=1}^{T} p(\boldsymbol{\eta}_t \mid \mathbf{u}_t, \mathbf{e}_t)\, \delta(\boldsymbol{f}_t - \mathcal{F}_t(\boldsymbol{\eta}_{1:t}))\; p(\boldsymbol{y}_t \mid \boldsymbol{f}_t), \tag{4}$$

where $\mathcal{F}_t$ is the deterministic innovation-to-factor map induced by the dynamics and δ is a Dirac mass enforcing the dynamic constraint (we overload notation across discrete and continuous cases). We will use equality of the observation law together with the iVAE identifiability conditions (injective decoder, sufficient auxiliary variation, and non-Gaussian exponential-family components) to characterize the ambiguity class of $\boldsymbol{\eta}_t$.

Deterministic conditioning variable. The regime path enters only through $\mathbf{e}_t = \psi(\mathbf{u}_t)$ with measurable deterministic map ψ (soft assignments or a single regime are both covered). Hence the effective auxiliary variable is $(\mathbf{u}_t, \psi(\mathbf{u}_t))$, which is observed up to deterministic post-processing. This satisfies the exogeneity requirement in the iVAE argument used below.

A.2 Companion form (AR(p))

For AR(p), we stack the current and $p-1$ lagged factor vectors into $\boldsymbol{s}_t = (\boldsymbol{f}_t^\top, \boldsymbol{f}_{t-1}^\top, \ldots, \boldsymbol{f}_{t-p+1}^\top)^\top \in \mathbb{R}^{rp}$. The stacked state follows $\boldsymbol{s}_{t+1} = \mathcal{A}\boldsymbol{s}_t + \mathcal{B}\boldsymbol{\eta}_t$ with block-diagonal $\mathcal{A}$ and $\mathcal{B}$. Unrolling gives

$$\boldsymbol{s}_t = \mathcal{A}^t \boldsymbol{s}_0 + \sum_{\ell=0}^{t-1} \mathcal{A}^\ell \mathcal{B}\, \boldsymbol{\eta}_{t-1-\ell}.$$

With $\mathcal{C}$ such that $\boldsymbol{f}_t = \mathcal{C}\boldsymbol{s}_t$, we obtain

$$\boldsymbol{f}_t = \mathcal{C}\mathcal{A}^t \boldsymbol{s}_0 + \sum_{\ell=0}^{t-1} \mathcal{H}_\ell \boldsymbol{\eta}_{t-1-\ell}, \quad \mathcal{H}_\ell := \mathcal{C}\mathcal{A}^\ell \mathcal{B}.$$

Because $\mathcal{A}$ and $\mathcal{B}$ are block diagonal, each $\mathcal{H}_\ell$ is diagonal in factor index. So $\mathcal{F}_t$ is *component-wise*: factor i at time t depends only on $(\eta_{i,1}, \ldots, \eta_{i,t})$. This representation also supports efficient long-horizon evaluation, such as FFT or Krylov methods (Zhang et al., 2023).

Component-wise operator form. Write $f_{i,t} = \mathcal{F}_{i,t}(\eta_{i,1:t})$ for measurable operators $\mathcal{F}_{i,t}$ induced by diagonal (or block-diagonal companion) dynamics. Then

$$\boldsymbol{f}_t = \mathcal{F}_t(\boldsymbol{\eta}_{1:t}) = \big(\mathcal{F}_{1,t}(\eta_{1,1:t}), \dots, \mathcal{F}_{r,t}(\eta_{r,1:t})\big)^\top.$$

This representation is used in Step 2 of the proof (the propagation lemma, Lemma 2 in the main text) to show preservation of the same ambiguity class from innovations to factors.

A.3 Proof of Theorem 3

Proof. Let $(\theta, \boldsymbol{\eta}, \boldsymbol{f})$ and $(\tilde{\theta}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{f}})$ be two parameterizations that induce the same $p(\boldsymbol{y}_{1:T}, \boldsymbol{u}_{1:T})$. We show that the corresponding latent variables are related by a transformation in $\mathcal{T}$ almost surely.

Step 1 (Innovations, invocation of Khemakhem et al. 2020). The conditioning variable $(\mathbf{u}_t, \mathbf{e}_t)$ is deterministic given $\mathbf{u}_t$ and observed, so it is a valid auxiliary variable in the iVAE setting. The innovation prior has the form

$$p(\boldsymbol{\eta}_t \mid \mathbf{u}_t, \mathbf{e}_t) = \prod_{i=1}^{r} h_i(\eta_{i,t}) \exp\big(\mathbf{T}_i(\eta_{i,t})^\top \boldsymbol{\lambda}_i(\mathbf{u}_t, \mathbf{e}_t) - \Psi_i(\boldsymbol{\lambda}_i(\mathbf{u}_t, \mathbf{e}_t))\big)\,, \tag{5}$$

with h_i the base measure, $\mathbf{T}_i \in \mathbb{R}^k$ the sufficient statistics, and Ψ_i the log-partition function. Under Assumption 1 of Theorem 3, the natural-parameter map $\boldsymbol{\lambda}(\mathbf{u}, \mathbf{e})$ has full column rank at at least $rk + 1$ distinct conditioning values (reducing to $r + 1$ when $k = 1$). Therefore, by Theorem 1 of Khemakhem et al. (2020), any alternative parameterization with the same observation law must satisfy

$$\tilde{\eta}_{i,t} = a_i(\eta_{\pi(i),t}) \quad \text{a.s.},$$

for some permutation π on $\{1, \dots, r\}$ and component-wise affine (or monotone invertible under regularity) maps $\{a_i\}$. We use a_i to denote the ambiguity transformation to avoid re-using h_i, which already names the base measure. The rank condition excludes broader linear mixing classes. Hence innovations are partially identifiable up to $\mathcal{T}$.

Step 2 (Propagation to factors, Lemma 2). By the operator form above, each coordinate satisfies

$$f_{i,t} = \mathcal{F}_{i,t}(\eta_{i,1:t}), \qquad \tilde{f}_{i,t} = \tilde{\mathcal{F}}_{i,t}(\tilde{\eta}_{i,1:t}).$$

From Step 1, $\tilde{\eta}_{i,t} = a_i(\eta_{\pi(i),t})$ almost surely. Since dynamics are component-wise and no cross-index coupling is allowed, each transformed coordinate still depends on a single permuted innovation coordinate. Therefore there exist induced maps $\bar{a}_{i,t}$ such that

$$\tilde{f}_{i,t} = \bar{a}_{i,t}\big(f_{\pi(i),t}\big) \quad \text{a.s.}$$

In the linear diagonal case with affine a_i, $\bar{a}_{i,t}$ is affine as well, so the same class $\mathcal{T}$ is preserved from innovations to factors. This fails with full matrices because then $\mathcal{F}_t$ mixes coordinates and the class expands toward general linear transforms. This step is the *only* new identifiability argument in the paper beyond a direct citation of Khemakhem et al. (2020).

Step 3 (Observation model). Equality of $p(\boldsymbol{y}_{1:T}, \boldsymbol{u}_{1:T})$ implies equality of each conditional law $p(\boldsymbol{y}_t \mid \mathbf{u}_{1:T})$. By Step 2, latent factors differ only by $\mathcal{T}$. Injectivity of g on factor support excludes additional non-identifiable collapse in observation space. Fixing σ^2 removes the scale-noise trade-off that would otherwise create an extra ambiguity. Hence the observational model does not enlarge $\mathcal{T}$.

Step 4 (Conclusion). Combining Steps 1–3, both innovations and factors are identified up to the same class $\mathcal{T}$ whenever two parameterizations induce the same observational law. This is exactly Definition 1. □

A.4 Remarks

Identifiability is established at the innovation level through variation of $\boldsymbol{\lambda}$ over $(\mathbf{u}, \mathbf{e})$, then inherited by $\{\boldsymbol{f}_t\}$ through deterministic component-wise dynamics. Companion-form and Krylov implementations are computational devices and do not alter the argument. In particular, $\mathcal{H}_t$ need not be invertible. Under variational inference, identifiability remains a property of the generative model, so the variational distribution may be chosen with any conditioning set (current frame, longer window, or full sequence) without changing the theorem.

What breaks identifiability. Two modeling choices are especially problematic. First, conditioning on a continuous latent variable, for example $\boldsymbol{\eta}_t \sim p(\boldsymbol{\eta} \mid \mathbf{u}_t, \mathbf{z}_t)$, breaks the iVAE argument because conditioning is no longer on fixed observed context. Second, directly feeding observations or attention features into innovation definitions, such as $\boldsymbol{\eta}_t = \boldsymbol{\mu}(\mathbf{u}_t, \boldsymbol{y}_{t-L:t}) + \ldots$, makes the auxiliary variable non-exogenous. In both cases, the identification argument no longer applies.

Encoder. In our implementation the encoder conditions on $(\boldsymbol{y}_t, \mathbf{u}_t)$ and not on $\mathbf{e}_t$ (since $\mathbf{e}_t$ is a deterministic function of $\mathbf{u}_t$). Any richer conditioning is admissible: identifiability is a property of the generative model, so the variational distribution may depend on longer context or on $\mathbf{e}_t$ without changing Theorem 3.

Appendix B. Degeneracy of Gaussian Innovations

We present the argument first for the *single-regime* case ($K = 1$). In that case, $\mathbf{e}_t$ is fixed, so we write the prior as $p(\boldsymbol{\eta}_t \mid \boldsymbol{u}_t)$. The same degeneracy extends to $K > 1$ when conditioning on both $\boldsymbol{u}_t$ and $\mathbf{e}_t$.

B.1 Setup: Gaussian Innovations in Exponential-Family Form

Suppose innovations are conditionally Gaussian,

$$\boldsymbol{\eta}_t \mid \boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{u}_t), \Sigma(\boldsymbol{u}_t)), \tag{6}$$

with $\Sigma(\boldsymbol{u}_t)$ positive definite. The log-density is

$$\log p(\boldsymbol{\eta}_t \mid \boldsymbol{u}_t) = -\tfrac{1}{2}\boldsymbol{\eta}_t^\top \Sigma(\boldsymbol{u}_t)^{-1}\boldsymbol{\eta}_t + \boldsymbol{\eta}_t^\top \Sigma(\boldsymbol{u}_t)^{-1}\boldsymbol{\mu}(\boldsymbol{u}_t) + c(\boldsymbol{u}_t), \tag{7}$$

where $c(\boldsymbol{u}_t)$ does not depend on $\boldsymbol{\eta}_t$. In exponential-family form, sufficient statistics are $\mathbf{T}(\boldsymbol{\eta}_t) = (\boldsymbol{\eta}_t, \mathrm{vec}(\boldsymbol{\eta}_t\boldsymbol{\eta}_t^\top))$ and natural parameters $\boldsymbol{\lambda}(\boldsymbol{u}_t)$ are functions of $\Sigma(\boldsymbol{u}_t)^{-1}$ and $\Sigma(\boldsymbol{u}_t)^{-1}\boldsymbol{\mu}(\boldsymbol{u}_t)$.

Lemma 4 (Gaussian family is closed under linear maps)**.** *For any invertible $R \in \mathbb{R}^{r\times r}$, the transformed variable $\tilde{\boldsymbol{\eta}}_t = R\boldsymbol{\eta}_t$ is also Gaussian given $\boldsymbol{u}_t$, with mean $R\boldsymbol{\mu}(\boldsymbol{u}_t)$ and covariance $R\Sigma(\boldsymbol{u}_t)R^\top$. The log-density of $\tilde{\boldsymbol{\eta}}_t \mid \boldsymbol{u}_t$ lies in the same (linear plus quadratic in $\boldsymbol{\eta}$) family as that of $\boldsymbol{\eta}_t \mid \boldsymbol{u}_t$. Thus the observation likelihood $p(\boldsymbol{y}_{1:T} \mid \boldsymbol{u}_{1:T})$ is unchanged under the reparameterization $\boldsymbol{\eta}_t \mapsto R\boldsymbol{\eta}_t$ (with the decoder and dynamics adjusted accordingly).*

The lemma implies that the ambiguity class of any Gaussian-innovation model includes all invertible linear maps. As a result, full-rank variation in $\boldsymbol{\lambda}(\mathbf{u}, \mathbf{e})$ cannot reduce the class to permutation and component-wise affine maps.

Density transformation detail. Let $\tilde{\boldsymbol{\eta}}_t = R\boldsymbol{\eta}_t$ with invertible R. By change of variables,

$$p_{\tilde{\eta}}(\tilde{\boldsymbol{\eta}}_t \mid \boldsymbol{u}_t) = p_\eta(R^{-1}\tilde{\boldsymbol{\eta}}_t \mid \boldsymbol{u}_t)\,|\det(R^{-1})|.$$

Substituting the Gaussian density gives

$$\tilde{\boldsymbol{\eta}}_t \mid \boldsymbol{u}_t \sim \mathcal{N}\big(R\boldsymbol{\mu}(\boldsymbol{u}_t),\ R\Sigma(\boldsymbol{u}_t)R^\top\big).$$

Hence the transformed family remains Gaussian for every invertible R.

B.2 Main Result: Gaussian Innovations Are Not Identifiable

Proposition 5 (Gaussian innovations are not identifiable up to $\mathcal{T}$)**.** *Let $\boldsymbol{\eta}_t \mid \boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{u}_t), \Sigma(\boldsymbol{u}_t))$ with $\Sigma(\boldsymbol{u}_t)$ positive definite. For any invertible $R \in \mathbb{R}^{r\times r}$, the transformed innovations $\tilde{\boldsymbol{\eta}}_t = R\boldsymbol{\eta}_t$ satisfy*

$$\tilde{\boldsymbol{\eta}}_t \mid \boldsymbol{u}_t \sim \mathcal{N}\big(R\boldsymbol{\mu}(\boldsymbol{u}_t),\ R\Sigma(\boldsymbol{u}_t)R^\top\big)\,, \tag{8}$$

and, by Lemma 4, the observation likelihood is unchanged under the corresponding full-model reparameterization. Therefore the ambiguity class includes all invertible linear maps. Innovations are identifiable at most up to invertible linear transformation, not up to the smaller class $\mathcal{T}$ (permutation and component-wise affine maps) that is available in the non-Gaussian case (Appendix A).

Even when $\boldsymbol{\mu}(\boldsymbol{u}_t)$ or $\Sigma(\boldsymbol{u}_t)$ vary with $\boldsymbol{u}_t$, the conditional family remains Gaussian and closed under R. Auxiliary variation alone therefore cannot break rotational or scaling symmetry.

Sketch of observational invariance. Consider linear dynamics and decoder for clarity,

$$\boldsymbol{f}_{t+1} = A\boldsymbol{f}_t + B\boldsymbol{\eta}_t, \qquad \boldsymbol{y}_t = C\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t.$$

Define transformed parameters

$$\tilde{\boldsymbol{f}}_t = R\boldsymbol{f}_t, \quad \tilde{A} = RAR^{-1}, \quad \tilde{B} = RB, \quad \tilde{C} = CR^{-1}.$$

Then

$$\tilde{\boldsymbol{f}}_{t+1} = \tilde{A}\tilde{\boldsymbol{f}}_t + \tilde{B}\tilde{\boldsymbol{\eta}}_t, \qquad \tilde{\boldsymbol{y}}_t = \tilde{C}\tilde{\boldsymbol{f}}_t + \boldsymbol{\varepsilon}_t = C\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t = \boldsymbol{y}_t.$$

So the induced observation law is unchanged. The same idea extends to nonlinear decoders by composing decoder maps with R^{-1}.

Example ($r = 2$)**.** For $\boldsymbol{\eta}_t \mid \boldsymbol{u}_t \sim \mathcal{N}(\mathbf{0}, \Sigma(\boldsymbol{u}_t))$, any orthogonal Q yields $\tilde{\boldsymbol{\eta}}_t = Q\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, Q\Sigma(\boldsymbol{u}_t)Q^\top)$. The observation distribution is unchanged, so the model cannot distinguish original from rotated innovations.

B.3 Implications for Dynamic Factor Models

In classical Gaussian dynamic factor models (Stock and Watson, 2002), factors evolve as

$$\boldsymbol{f}_{t+1} = A\boldsymbol{f}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{9}$$

with a linear decoder. The joint distribution of observations is then invariant under $\boldsymbol{f}_t \mapsto R\boldsymbol{f}_t$, $A \mapsto RAR^{-1}$, $\Sigma \mapsto R\Sigma R^\top$ for any invertible R—the well-known rotational indeterminacy. Temporal dependence or higher-order (e.g. companion-form) dynamics do not remove this ambiguity, because linear Gaussian dynamics remain closed under linear maps.

Restricting A to be diagonal avoids cross-factor mixing in the dynamics but does not restore identifiability. The ambiguity class is fixed at the *innovation* level: any invertible $\tilde{\boldsymbol{\eta}}_t = R\boldsymbol{\eta}_t$ preserves the Gaussian family and can be paired with a transformed decoder and dynamics to yield the same observations. Diagonal A only constrains how factors evolve over time, not the invariance of the innovation distribution under linear mixing.

B.4 Contrast with Non-Gaussian Innovations and Takeaway

For non-Gaussian exponential families with linearly independent sufficient statistics (for example Laplace or Student-t), the log-density is not invariant to general linear mixing (Hyvarinen and Oja, 2000). In this setting, $\tilde{\boldsymbol{\eta}}_t = R\boldsymbol{\eta}_t$ preserves the product form only when R is a permutation of a scaled identity, which corresponds to permutation and component-wise scaling. Conditioning on $\boldsymbol{u}_t$

then yields natural-parameter variation that can satisfy full-rank conditions in iVAE (Khemakhem et al., 2020). Innovations are then partially identifiable up to permutation and component-wise affine maps (or monotone invertible maps under regularity). iVDFM uses this non-Gaussian route so identifiability is obtained at the innovation level and then propagated through dynamics (Theorem 3).

Connection to finite-sample diagnostics. This proposition is a population-level identifiability statement. Finite-sample recovery scores such as MCC can still appear similar between Gaussian and non-Gaussian runs, because those scores depend on optimization, sample size, and matching heuristics. Similar finite-sample scores do not imply equal identifiability classes.

The degeneracy of Gaussian innovations is thus a property of the Gaussian family, not of the dynamic structure. iVDFM avoids it by using conditional non-Gaussian innovations.

Appendix C. Experiment Details

This appendix gives details for the experiments in Section 4: factor recovery, causal intervention, and forecasting.

C.1 Factor Recovery

Synthetic DGPs. We use two synthetic DGPs adapted from Khemakhem et al. (2020) and Andreini et al. (2020). In the static setting, latents are generated independently across time conditional on an auxiliary variable and observations are formed through nonlinear mixing. In the dynamic setting, latent factors follow AR dynamics driven by i.i.d. non-Gaussian innovations (Laplace with unit-variance scaling), and observations are generated by a nonlinear decoder. Both settings use $T = 200$, $N = 20$, and $r = 5$. This design isolates the role of temporal dynamics while keeping measurement structure comparable.

Baselines. iVDFM is compared to DFM (Stock and Watson, 2002), DDFM (Andreini et al., 2020), VAE (Kingma and Welling, 2014), and iVAE (Khemakhem et al., 2020). DFM is a linear dynamic factor model; without further restrictions it is only identifiable up to an arbitrary invertible linear mixing of factors. DDFM uses nonlinear measurement maps but does not impose a controlled ambiguity class. VAE does not use auxiliary conditioning. iVAE applies conditional identification per time step, with time used as auxiliary context and no explicit temporal dependence in latent dynamics.

Metrics. Let $F \in \mathbb{R}^{T\times r}$ denote ground-truth factors and $\hat{F} \in \mathbb{R}^{T\times r}$ recovered factors. MCC (higher is better) is computed from max-weight assignment over Pearson correlations between standardized columns of F and $\hat{F}$:

$$\mathrm{MCC} = \frac{1}{r} \max_{\pi\in S_r} \sum_{i=1}^{r} |C_{i,\pi(i)}|.$$

Subspace distance (lower is better) is the principal-angle distance between column spans. With orthonormal bases Q_F, $Q_{\hat{F}}$ and singular values σ_i of $Q_F^\top Q_{\hat{F}}$ clamped to $[0, 1]$, we use

$$\mathrm{Subspace} = \frac{1}{r} \sum_{i=1}^{r} \arccos(\sigma_i).$$

Smoothness (lower is better) is the average step-to-step ℓ_2 norm of recovered trajectories. DFM tends to be smoothest because of linear latent dynamics (Stock and Watson, 2002). Trace R^2 (higher is better) is the fraction of true factor variance explained by the recovered span (Andreini et al., 2020), with range $[0, 1]$.

Protocol. We run $N_{\text{sim}} = 10$ simulations per suite (dynamic and static) with fixed seeds so all models are evaluated on the same data. We report mean and standard deviation for every metric. For stability and comparability, each observed series is standardized to zero mean and unit variance within each synthetic dataset before fitting. Common settings are window = 200, max epochs = 200, and learning rate = 0.002. To avoid one-sided tuning, we apply the same Optuna budget per model within each suite. For iVDFM, VAE, and iVAE, we maximize MCC on a fixed synthetic dataset from the same suite and then evaluate with those selected hyperparameters on the suite simulations. Results appear in Tables 1 and 2.

Reproducibility details. Each simulation uses a fixed triplet of seeds for data generation, model initialization, and minibatch ordering. We reuse the same seed triplets across models within each simulation index. This controls for dataset and optimization randomness when comparing recovery metrics. In dynamic runs, we discard a short warm-up prefix from generated latent trajectories before evaluation to reduce dependence on initialization transients.

C.2 Causal Intervention

SCM. We use a minimal time-series SCM (Pearl, 2009) with exogenous innovations as the only source of randomness. State evolution: $\boldsymbol{f}_t = A\boldsymbol{f}_{t-1} + \boldsymbol{\epsilon}_t$, $\boldsymbol{f}_0 = \mathbf{0}$, with $A \in \mathbb{R}^{r\times r}$ stable. Observation map: $\boldsymbol{y}_t = C\boldsymbol{f}_t$ (base/linear) or $\boldsymbol{y}_t = g(\boldsymbol{f}_t)$ (general). Shocks $\boldsymbol{\epsilon}_t$ are i.i.d. non-Gaussian (e.g. Laplace), giving the causal graph $\boldsymbol{\epsilon}_t \to \boldsymbol{f}_t \to \boldsymbol{y}_t$.

Do-intervention and IRF. An intervention at (t_0, i) sets $\epsilon_{t_0}^{(i)} = c$ and leaves all other shocks unchanged. The impulse response function (IRF) at horizon h is the difference in expected observations at $t_0 + h$ with and without intervention. Under linear C, the ground-truth IRF is

$$\text{IRF}_{\text{true}}(h) = c \cdot CA^h\mathbf{e}_i \in \mathbb{R}^N,$$

where $\mathbf{e}_i$ denotes the ith standard basis vector in $\mathbb{R}^r$ (distinct from the regime embedding $\mathbf{e}_j$ of the main text). In the base SCM, $A = \rho I$. In the regime SCM, A depends on the regime. In the chain SCM, A is upper triangular so factor i affects factor $i + 1$.

Model-implied IRF (iVDFM). iVDFM learns innovations $\boldsymbol{\eta}_t$ and dynamics $\boldsymbol{f}_t = \mathcal{F}_t(\boldsymbol{\eta}_{1:t})$ with observations $\boldsymbol{y}_t = g_\theta(\boldsymbol{f}_t)$. Under the main-text assumptions, $\boldsymbol{\eta}_t$ aligns with SCM shocks $\boldsymbol{\epsilon}_t$ up to the same ambiguity class (Appendix A). We define model-implied IRFs in five steps. First, fit on observational data. Second, construct $\boldsymbol{\eta}^{\text{do}}$ by setting $(\boldsymbol{\eta}_{t_0}^{\text{do}})^{(i)} = c$. Third, propagate through learned dynamics to obtain $\boldsymbol{f}^{\text{do}}$. Fourth, decode both baseline $\widehat{\boldsymbol{y}}_t$ and intervention trajectory $\boldsymbol{y}_t^{\text{do}}$. Fifth, compute

$$\widehat{\text{IRF}}(h) = \boldsymbol{y}_{t_0+h}^{\text{do}} - \widehat{\boldsymbol{y}}_{t_0+h}.$$

This evaluation is specific to models with an interpretable innovation process.

Metrics and protocol. We compare $\widehat{\text{IRF}}$ and IRF_{true} over horizon-variable grids. IRF-MSE (lower is better) is mean squared error on that grid. Sign accuracy (higher is better) is the fraction of (h, n) pairs where $\text{sign}(\widehat{\text{IRF}}_{h,n}) = \text{sign}(\text{IRF}_{\text{true},h,n})$. We also report IRF-MAE and Pearson correlation (Appendix Table 8). Sign accuracy is a strict criterion and can be more brittle than MSE/MAE under short sequences because intervention-coordinate alignment is estimated from finite data. We generate data from the three SCM variants and train iVDFM only on observational samples with fixed hyperparameters for reproducibility. We run five simulations per setting. Main text results are in Table 3, and stress tests are in Table 8. Experiments use a single GPU.

Intervention setup details. Unless otherwise stated, interventions are one-shot impulses at a fixed intervention time t_0 with amplitude c chosen from a symmetric range around zero. We evaluate all intervention coordinates and average metrics across coordinates and simulation runs. For each run, baseline and intervention rollouts share the same posterior sample path outside the intervened coordinate so that differences isolate intervention effects.

C.3 Forecasting

Datasets and task. We use five real-world multivariate time-series datasets: ETTh1 and ETTh2 (electricity transformer temperature, hourly), ETTm1 and ETTm2 (15-minute), and Weather. For each dataset we train on a fixed train/val split and evaluate out-of-sample forecasts at horizons $H \in \{96, 192, 336, 720\}$. Predictions are made in standardized (scaled) space: we fit a scaler on the training set and evaluate metrics without inverse-transforming, so that MSE and CRPS are comparable across variables and datasets.

Baselines. iVDFM is compared to iTransformer (Liu et al., 2023), TimeMixer (Wang et al., 2024a), TimeXer (Wang et al., 2024b), DDFM (Andreini et al., 2020), and MLPMultivariate. All baselines are trained and evaluated on identical splits and horizons. iVDFM uses fixed per-dataset hyperparameters for reproducibility, with no Optuna during evaluation. Configuration details are in Appendix E.

Metrics. CRPS (lower is better) is computed from quantile forecasts (10th, 50th, 90th percentiles) and averaged across horizon and variables. Standardized MSE (lower is better) is computed between median point forecasts and targets in standardized space. We report both metrics per dataset, model, and horizon. Table 4 gives horizon-averaged values by dataset, and Table 9 reports all horizons.

Protocol. Common settings are window = 96, stride = 25, and up to 10 rolling forecast origins per dataset-horizon pair. Each model is trained once for each dataset and horizon, and metrics are aggregated over origins. iVDFM encoder, decoder, prior architecture, and training hyperparameters are fixed per dataset in the experiment configuration. Results are reported in Table 4 (main text) and Table 9 (appendix).

Evaluation consistency checks. Before reporting, we verify that each model uses identical scaler statistics from the training split and that rolling origins are aligned across models. We also check that all reported CRPS values are computed from the same quantile set and that point forecasts for MSE correspond to the median prediction used in CRPS summaries.

Appendix D. Additional Results

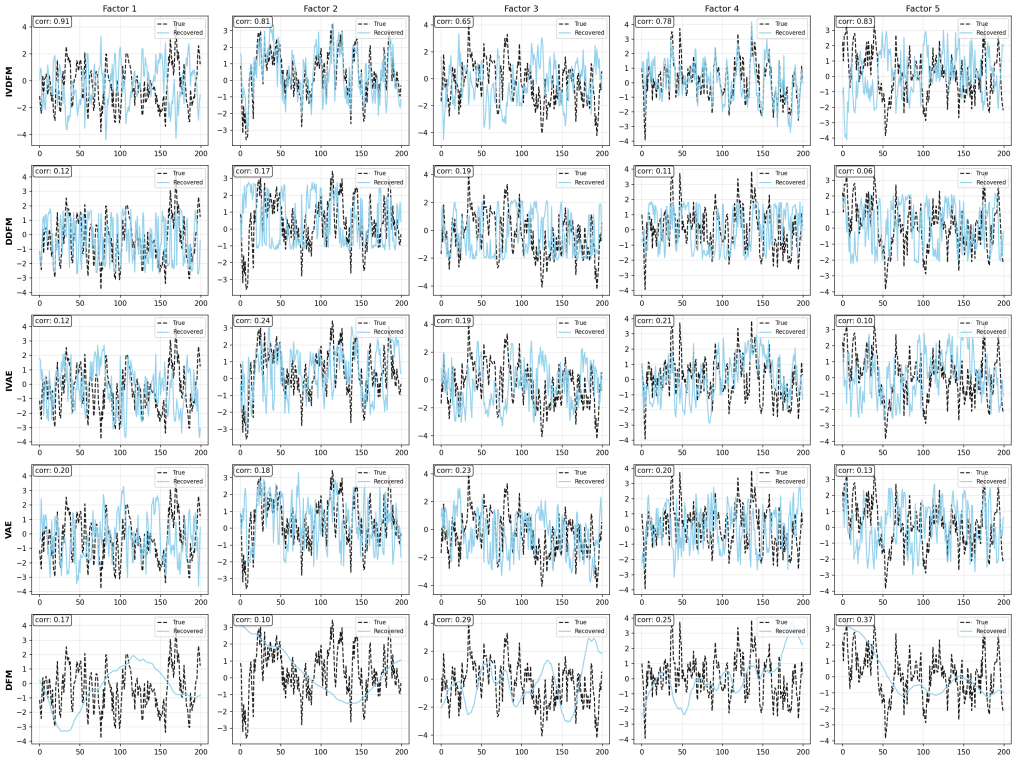
D.1 Synthetic Factor Recovery Visualization

This appendix provides full-page visualizations of synthetic factor recovery for both **dynamic** and **static** DGPs. Each grid shows recovered factors for each model together with ground-truth factors. Columns are matched using the same MCC assignment used in evaluation. For readability, recovered factors are rescaled to the mean and standard deviation of true factors. This visual rescaling does not change correlations or reported metrics.

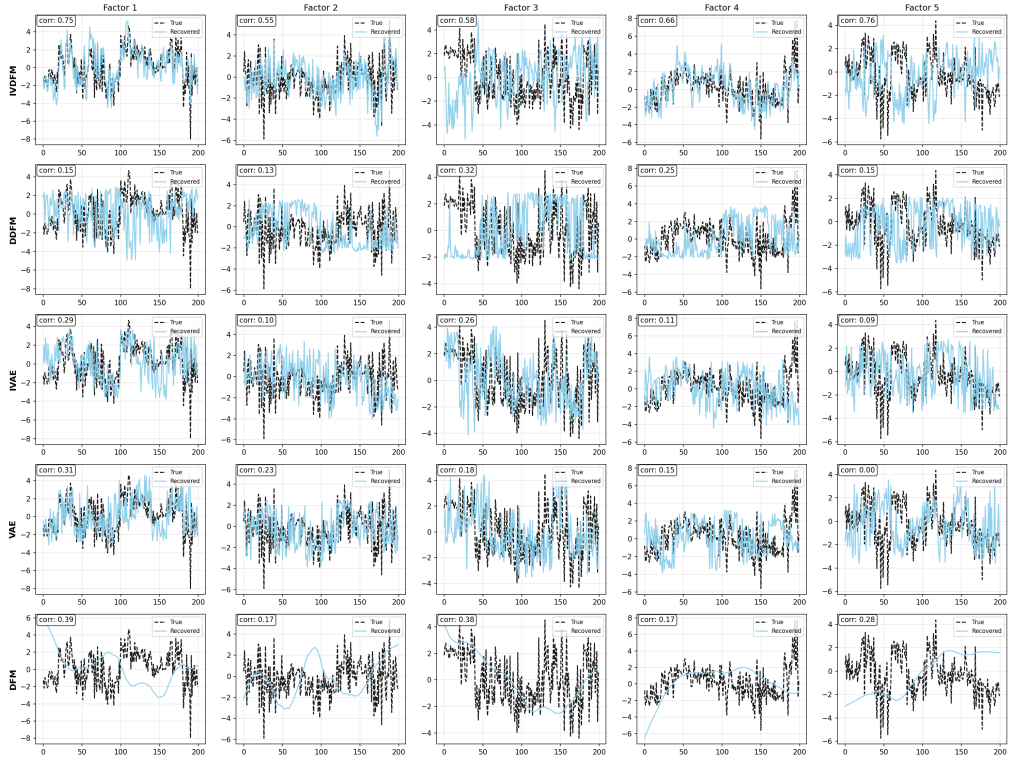
D.2 Factor Recovery Diagnostics

Factor-recovery metrics (MCC, subspace distance, Trace R^2) measure alignment to ground truth after permutation-invariant matching. They are affected by finite samples and optimization. They do not directly measure the size of the latent ambiguity class. The diagnostics below therefore add three complementary checks: iVAE-aligned static runs, finite-sample sweeps, and innovation-distribution comparisons.

iVAE static diagnostic. This diagnostic tests whether iVAE performance is consistent when the synthetic setup better matches a canonical iVAE-aligned regime and training choices. We therefore run a static diagnostic with longer samples and simpler mixing: T=2000, N=20, r=5, n_{seg}=20, and a single linear mixing layer without intermediate nonlinearity. We train VAE and iVAE for 400 epochs and report mean ± std over 5 simulations.



(a) Dynamic synthetic factor recovery (5×5). Columns aligned by MCC-based matching; recovered factors rescaled to true factor scale for visualization.



(b) Static synthetic factor recovery (5×5). Columns aligned by MCC-based matching; recovered factors rescaled to true factor scale for visualization.

Table 5: iVAE static diagnostic (static synthetic DGP, iVAE-aligned). Mean ± std over 5 simulations.

Model	MCC ↑	Subspace ↓	Smoothness ↓	Trace R^2 ↑
VAE	0.7160 ± 0.0609	0.2665 ± 0.1176	2.2715 ± 0.0286	0.8681 ± 0.0638
iVAE	0.7265 ± 0.0828	0.2305 ± 0.0915	2.2710 ± 0.0635	0.8869 ± 0.0374

Table 6: Dynamic factor recovery finite-sample sweep for iVDFM (Laplace innovations), training window equal to T. Mean ± std over 5 simulations.

T	MCC ↑	Subspace ↓	Smoothness ↓	Trace R^2 ↑
100	0.608 ± 0.088	0.597 ± 0.112	2.050 ± 0.037	0.731 ± 0.074
200	0.666 ± 0.036	0.377 ± 0.065	1.723 ± 0.099	0.857 ± 0.043
500	0.660 ± 0.039	0.307 ± 0.021	1.665 ± 0.040	0.888 ± 0.017
1000	0.703 ± 0.092	0.337 ± 0.044	1.661 ± 0.025	0.857 ± 0.041

Finite-sample sweep (varying T). To assess finite-sample effects in the dynamic suite, we vary the sequence length T while keeping N=20, r=5, and Laplace innovations. We set the training window equal to T so each simulation contributes one full trajectory, which avoids confounding the finite-sample comparison with a fixed shorter window. Table 6 reports iVDFM performance (mean ± std over 5 simulations) for $T \in \{100, 200, 500, 1000\}$. Subspace distance and smoothness decrease and Trace R^2 increases with T; MCC trends upward with some run-to-run noise at the larger end.

Gaussian vs non-Gaussian innovations (diagnostic). Our identifiability discussion (Appendix B) shows that Gaussian innovations admit a larger ambiguity class at the population level. The main factor-recovery metrics, however, measure matched alignment and do not directly quantify ambiguity-class size. As a diagnostic, Table 7 compares dynamic factor recovery for iVDFM in three settings: (i) Laplace shocks in the synthetic DGP (baseline), (ii) Gaussian shocks in the synthetic DGP, and (iii) Laplace shocks with a Gaussian innovation prior in iVDFM. In this synthetic setup, aggregate recovery metrics are of similar scale.

D.3 Causal Intervention Stress Tests (Full Table)

Table 8 reports causal intervention stress tests (SCM comparison and varying T, misspecified r for base, regime, and chain SCMs) for iVDFM. Main-text Table 3 reports the SCM comparison only.

Table 7: Innovation-distribution diagnostics for dynamic factor recovery (iVDFM). Mean ± std over 5 simulations.

Setting	MCC ↑	Subspace ↓	Smoothness ↓	Trace R^2 ↑
DGP Laplace; iVDFM Laplace	0.6696 ± 0.0696	0.4455 ± 0.0842	1.6905 ± 0.0810	0.8103 ± 0.0471
DGP Gaussian; iVDFM Laplace	0.6879 ± 0.0769	0.3315 ± 0.0387	1.7547 ± 0.0495	0.8742 ± 0.0279
DGP Laplace; iVDFM Gaussian	0.6805 ± 0.0779	0.3531 ± 0.0489	1.7221 ± 0.0767	0.8720 ± 0.0319

Table 8: iVDFM causal intervention stress tests: varying T and misspecified r (base, regime, chain SCMs). IRF-MSE, IRF-MAE: mean squared / absolute error; Sign Acc: fraction of (time step, variable) pairs with correct sign; IRF Corr: Pearson correlation with ground truth. Mean ± std over five simulations.

Setting	IRF-MSE ↓	IRF-MAE ↓	Sign Acc ↑	IRF Corr ↑
SCM comparison ($T = 200$, $r = 3$)				
Base (linear)	0.89 ± 1.01	0.45 ± 0.23	0.58 ± 0.25	0.21 ± 0.75
Regime	1.28 ± 0.92	0.77 ± 0.27	0.60 ± 0.27	0.13 ± 0.62
Chain	2.18 ± 2.40	0.97 ± 0.40	0.58 ± 0.33	0.20 ± 0.62
Varying T (base SCM, $r = 3$)				
$T = 100$	0.75 ± 0.54	0.44 ± 0.16	0.02 ± 0.01	-0.08 ± 0.52
$T = 200$	0.89 ± 1.01	0.45 ± 0.23	0.58 ± 0.25	0.21 ± 0.75
$T = 500$	0.76 ± 0.72	0.43 ± 0.19	0.60 ± 0.28	0.27 ± 0.71
$T = 1000$	0.77 ± 0.70	0.43 ± 0.19	0.60 ± 0.28	0.24 ± 0.69
Misspecified r (base SCM, $T = 200$, true $r = 3$)				
$r = 2$ (fit)	1.16 ± 1.17	0.52 ± 0.28	0.41 ± 0.30	-0.16 ± 0.77
$r = 3$ (fit)	0.89 ± 1.01	0.45 ± 0.23	0.58 ± 0.25	0.21 ± 0.75
$r = 4$ (fit)	1.01 ± 0.84	0.50 ± 0.22	0.41 ± 0.25	-0.27 ± 0.58
$r = 5$ (fit)	0.79 ± 0.59	0.47 ± 0.18	0.42 ± 0.28	-0.04 ± 0.81
Varying T (Regime SCM, $r = 3$)				
$T = 100$	1.32 ± 1.02	0.77 ± 0.26	0.03 ± 0.01	0.25 ± 0.28
$T = 200$	1.28 ± 0.92	0.77 ± 0.27	0.60 ± 0.27	0.13 ± 0.62
$T = 500$	0.42 ± 0.33	0.22 ± 0.09	0.60 ± 0.28	0.29 ± 0.69
$T = 1000$	0.44 ± 0.31	0.23 ± 0.09	0.58 ± 0.24	0.17 ± 0.64
Misspecified r (Regime SCM, $T = 200$, true $r = 3$)				
$r = 2$ (fit)	1.51 ± 1.20	0.82 ± 0.34	0.40 ± 0.28	-0.14 ± 0.67
$r = 3$ (fit)	1.28 ± 0.92	0.77 ± 0.27	0.60 ± 0.27	0.13 ± 0.62
$r = 4$ (fit)	1.43 ± 1.08	0.81 ± 0.32	0.41 ± 0.26	-0.10 ± 0.66
$r = 5$ (fit)	1.24 ± 0.87	0.77 ± 0.29	0.44 ± 0.28	0.10 ± 0.73
Varying T (Chain SCM, $r = 3$)				
$T = 100$	1.95 ± 1.27	0.98 ± 0.27	0.03 ± 0.01	0.11 ± 0.28
$T = 200$	2.18 ± 2.40	0.97 ± 0.40	0.58 ± 0.33	0.20 ± 0.62
$T = 500$	2.06 ± 2.20	0.95 ± 0.37	0.62 ± 0.34	0.17 ± 0.58
$T = 1000$	2.07 ± 2.23	0.95 ± 0.37	0.62 ± 0.34	0.11 ± 0.58
Misspecified r (Chain SCM, $T = 200$, true $r = 3$)				
$r = 2$ (fit)	2.39 ± 2.33	1.03 ± 0.39	0.49 ± 0.34	-0.04 ± 0.68
$r = 3$ (fit)	2.18 ± 2.40	0.97 ± 0.40	0.58 ± 0.33	0.20 ± 0.62
$r = 4$ (fit)	2.04 ± 1.36	0.99 ± 0.29	0.48 ± 0.29	0.01 ± 0.65
$r = 5$ (fit)	1.43 ± 0.67	0.90 ± 0.24	0.62 ± 0.32	0.34 ± 0.50

D.4 Forecasting Results

Table 9 reports CRPS and MSE (standardized) for all datasets and horizons. Dataset names are shown vertically; horizon is in a separate column.

Appendix E. Ablation Study

We conduct an iVDFM-only ablation on ETTh1 and selected ETT/Weather splits. All runs use the same pipeline as Section 4.3. Unless noted, we report CRPS and standardized MSE at horizon 96 and link results to the main forecasting section.

E.1 Design Choices and Degenerate Cases

Prior and context. Using constant context $u_t \equiv u_0$ with a linear prior network yields a degenerate baseline: the conditional innovation prior does not vary across time. This is distinct from Gaussian degeneracy in Appendix B. We compare Laplace, Gaussian, and Student-t innovations; Gaussian remains degenerate under the identifiability conditions. We use Laplace with time-step context.

Number of regimes K. The choice $K > 1$ was introduced as an expressivity fix, not a theoretical add-on (Section 3). We swept $K \in \{1, 3, 5, 7, 9\}$ on ETTh1 at horizon 96 with three seeds each (Table 10). The improvement of small K over $K = 1$ is modest but reproducible: $K = 3$ lowers CRPS from 0.265 to 0.256 and sMSE from 1.061 to 1.032 (about 3% on each score). Further increasing K does not help on this setting; $K = 7$ and $K = 9$ are worse than $K = 1$ in both scores. We therefore do not claim that a larger number of regimes is universally beneficial; the mixture prior is a useful expressivity lever, but the useful range is narrow and setting-dependent. The cross-dataset default `num_regimes`= 7 used in the main forecasting table was selected on an earlier joint ETT/Weather sweep and reflects a compromise across datasets rather than a per-dataset optimum. The trade-off is that mixture components are not themselves identifiable (only their deterministic expectation $\mathbf{e}_t$ is), so we do not read individual regime embeddings as interpretable structural states.

Decoder and post-hoc correction. We compare linear, residual (MLP plus linear), and full-MLP decoders for the observation map. Linear decoding underperforms; a residual decoder improves, so we use the full MLP decoder (hidden size 128, one layer). We do not apply post-hoc residual correction.

E.2 Hyperparameters and Architecture

Protocol and optimization. Aligning the evaluation protocol with benchmark settings (rolling stride equal to the horizon and max origins 10) improves MSE. We use Laplace innovations, tune the learning rate around $4\text{–}6 \times 10^{-4}$, and use dropout around 0.08. We retain layer normalization, a two-layer encoder/decoder with hidden size 128, and set regime temperature $\tau = 0.2$ with `num_regimes=7`.

Architecture variants. Asymmetric bottlenecks and alternative activations (Swish, LeakyReLU) underperform ReLU; residual connections in encoder/decoder do not help. We keep a factor order of 2 for ETTh1/ETTh2 and 3 for ETTm1/ETTm2.

E.3 Chain-SCM Diagonal-Violation Sweep

Section 4 reports that IRF recovery degrades when the true SCM violates diagonal dynamics. To quantify this, we sweep the chain-SCM off-diagonal strength (parameter `chain_strength` in

Table 9: CRPS and MSE (standardized). Rows: dataset (horizon); columns: per-model CRPS and MSE (standardized).

Dataset	H	iVDFM		iTransformer		TimeMixer		TimeXer		DDFM		MLP	
		CRPS	MSE (std.)	CRPS	MSE (std.)	CRPS	MSE (std.)	CRPS	MSE (std.)	CRPS	MSE (std.)	CRPS	MSE (std.)
ETTh1	96	0.2683	1.0907	0.6805	1.2977	0.2492	0.6444	0.5518	1.1886	0.3172	1.4191	0.2275	0.6376
	192	0.2831	1.2684	0.6868	1.2812	0.2610	0.7614	0.5095	1.0815	0.3329	1.4004	0.2359	0.6832
	336	0.2835	1.3041	0.6936	1.3671	0.2828	0.8134	0.5683	1.2357	0.3390	1.2963	0.2324	0.6647
	720	0.2939	1.3684	0.7099	1.2374	0.3115	0.9848	0.5394	1.2506	0.3468	1.2502	0.2871	0.8883
ETTh2	96	0.2595	1.0307	0.5826	1.0791	0.2239	0.4882	0.4684	1.0791	0.2793	1.1222	0.1869	0.3663
	192	0.2509	0.9018	0.5156	0.9543	0.2291	0.3539	0.4616	0.8881	0.2765	1.0148	0.1699	0.3365
	336	0.2493	0.8795	0.4381	1.0926	0.2138	0.3688	0.4406	1.0679	0.2694	0.8808	0.1406	0.2302
	720	0.3126	1.2534	0.4701	0.9460	0.2549	0.3948	0.4371	0.9258	0.2848	1.0320	0.1782	0.3487
ETTm1	96	0.3237	1.2088	0.7856	1.4427	0.2309	0.4816	0.6996	1.5203	0.3525	1.5346	0.1844	0.4514
	192	0.3003	1.1049	0.7127	1.2799	0.2020	0.4598	0.6075	1.2303	0.3357	1.1912	0.1934	0.5271
	336	0.3062	1.1663	0.6483	1.1407	0.2385	0.5738	0.5320	1.0434	0.3394	1.0315	0.2006	0.5211
	720	0.3300	1.4042	0.6775	1.2881	0.2455	0.6719	0.5521	1.2070	0.3974	1.3520	0.2251	0.6366
ETTm2	96	0.2975	0.8705	0.8116	2.1123	0.2476	0.5310	0.6167	2.2748	0.5591	3.1659	0.1958	0.3789
	192	0.3114	0.9570	0.6778	1.4037	0.2329	0.4155	0.5579	1.5041	0.4921	2.5408	0.1805	0.3908
	336	0.3564	1.1939	0.5755	0.9626	0.2204	0.3914	0.4785	0.9450	0.4235	1.9969	0.1674	0.3612
	720	0.3934	1.3712	0.5107	0.8413	0.2281	0.3988	0.4710	0.8558	0.3633	1.5285	0.1798	0.4608
Weather	96	0.1969	0.6850	0.2949	0.5596	0.1460	0.2936	0.2704	0.5658	0.1362	0.3011	0.1711	0.3376
	192	0.2296	0.9232	0.2815	0.5242	0.1527	0.3527	0.2512	0.5182	0.1391	0.3517	0.1733	0.3535
	336	0.2954	1.3521	0.3608	0.9217	0.1675	0.3532	0.3163	0.9116	0.1791	0.5065	0.1889	0.4050
	720	0.3621	1.7743	0.3256	0.8048	0.1870	0.3630	0.2852	0.8330	0.2022	0.5977	0.1953	0.4044

Table 10: iVDFM K-ablation on ETTh1 at horizon 96. Mean ± std over 3 seeds (2026, 1337, 42). Other hyperparameters fixed per `config/experiment/forecasting.yaml`.

K	sMSE ↓	CRPS ↓	MAE ↓
1	1.061 ± 0.023	0.265 ± 0.005	0.739 ± 0.015
3	$\mathbf{1.032 \pm 0.071}$	$\mathbf{0.256 \pm 0.013}$	0.745 ± 0.032
5	1.061 ± 0.124	0.267 ± 0.028	0.759 ± 0.070
7	1.113 ± 0.083	0.273 ± 0.016	0.789 ± 0.045
9	1.114 ± 0.042	0.274 ± 0.010	0.791 ± 0.027

Table 11: Chain-SCM sweep. Each row is mean ± std over 5 simulations; IRF-MSE ↓, IRF-MAE ↓, Sign Acc ↑, IRF-Corr ↑. Row `chain_strength=0.2` is the main-text chain setting.

`chain_strength`	IRF-MSE	IRF-MAE	Sign Acc	IRF-Corr
0.0	0.89 ± 0.91	0.45 ± 0.20	0.58 ± 0.23	0.21 ± 0.67
0.1	1.31 ± 1.39	0.66 ± 0.28	0.70 ± 0.29	0.27 ± 0.59
0.2	2.18 ± 2.15	0.97 ± 0.36	0.58 ± 0.30	0.20 ± 0.56
0.3	4.02 ± 2.94	1.44 ± 0.40	0.61 ± 0.30	0.15 ± 0.49
0.5	15.26 ± 7.15	2.99 ± 0.73	0.56 ± 0.31	0.15 ± 0.34

`synthetic.causal.CausalChainSynthetic`) and re-run the causal-intervention protocol (5 simulations per setting, $T = 200$, $r = 3$, `do_time`=100, `do_value`=3.0). Table 11 shows that IRF-MSE grows with the amount of cross-factor mixing introduced by the subdiagonal: a zero off-diagonal reduces to the base SCM case and already has non-trivial error (learned coordinates are only aligned up to $\mathcal{T}$), while at `chain_strength`=0.5 the error is an order of magnitude larger. Sign accuracy and IRF correlation decrease more slowly but noisily, consistent with the interpretation that diagonal propagation fails to reproduce cross-factor influence paths while still capturing the direction of the dominant own-factor response.

E.4 Counterfactual Distribution under Partial Identifiability

To probe whether the identified innovation coordinates support counterfactual forecasting, we compare counterfactual trajectories of the learned model against the true SCM under a fair protocol: both ensembles share the same observational history up to `do_time` and vary only in post-intervention shocks. The true ensemble is generated by re-sampling SCM shocks $\varepsilon_{t>t_0}$ with $\varepsilon_{t_0,d} = c$ fixed; the model ensemble resamples $\boldsymbol{\eta}_{t>t_0}$ from the learned prior with the aligned coordinate pinned to c and decodes. We use the base SCM ($A{=}\rho I$, linear C) with $T{=}200$, $r{=}3$, `do_time`=100, $c{=}3.0$, horizon 20, and 100 samples per ensemble. Alignment per SCM dimension is chosen by picking the model innovation coordinate whose IRF best correlates with the true IRF (horizon 5).

The model's counterfactual distribution remains further from the true counterfactual than the null reference across all dimensions, i.e. the learned decoder does not reproduce the full multivariate distribution of the intervention. However, the directional effect $\mathbb{E}[Y^{do}] - \mathbb{E}[Y^{\text{null}}]$ is partially captured whenever the aligned coordinate has non-trivial IRF correlation (cosine 0.2–0.3 for dims $0, 1$; essentially zero for dim 2, where alignment itself fails). This is consistent with the partial-identifiability story: counterfactual accuracy is upper-bounded by the quality of coordinate alignment to the true SCM basis, so an imperfect finite-sample recovery (MCC around 0.6–0.7 in the dynamic suite) propagates directly to imperfect counterfactual magnitudes.

Table 12: Counterfactual trajectory comparison under the shared-history protocol. Columns: IRF-alignment correlation between model and true IRF for the selected coordinate; mean energy score of the model vs. true counterfactual ensemble (lower is better; the last column gives the energy score between the true no-intervention and true intervened ensembles as a reference scale); cosine between model and true expected causal effects $\mathbb{E}[Y^{do}] - \mathbb{E}[Y^{\text{null}}]$ averaged over horizon.

SCM dim	align corr	ES (model vs true)	ES (null vs do, true)	effect cosine
0	0.66	4.80	3.82	0.32
1	0.52	4.35	3.44	0.21
2	0.01	5.01	3.37	−0.04

E.5 Dataset-Specific Settings

We fix per-dataset settings in configuration for reproducibility (no Optuna). For ETTh1/ETTh2, we use `factor_order=2`, `decoder_var`≈`0.02--0.06`, and window length 96–1100. For ETTm1, we use `factor_order=3`, window length 2000, `decoder_var=0.03`, and `max_epochs=60`. ETTm2 and Weather use shared defaults (factor order 3, window 96).