DECOMPOSING REPRESENTATION DRIFT VIA INTERVENTIONS

Thomas Y. Chen & Daniel Xu
Columbia University
New York, NY 10027
{chen.thomas,xu.daniel}@columbia.edu

ABSTRACT

Machine learning systems operate in nonstationary settings where both the data-generating environment and the model itself evolve, a phenomenon traditionally studied as dataset shift and concept drift. Prior work often treats internal representations as black-box features or latent variables to be inferred, preventing an estimable separation of drift due to environmental change versus model updates. We introduce a probabilistic causal framework that embeds the feature extractor as a node in a larger structural causal model (SCM). This yields an interventional decomposition of representation drift into environment- and model-driven terms, estimable via variational inference without requiring paired pre/post-update representations, with consistency guarantees on the estimator. We show the decomposition is well-defined under relative identifiability, and connect it to downstream performance through Integral Probability Metric (IPM) bounds. Empirical results validate relative identifiability and robustness of the drift decomposition.

1 INTRODUCTION

Deployed ML systems are nonstationary: data sources, users, and objectives evolve, and models are continually updated. Classic work studies this through *dataset shift* and *concept drift*, which characterize changes in the joint distribution of inputs and labels and their effect on risk (Moreno-Torres et al., 2012; Gama et al., 2014; Webb et al., 2018; Hinder et al., 2024). For modern high-dimensional models, a large portion of change can occur *inside* the model: even when accuracy is stable, embeddings of the same input may move substantially in feature space. This *representation drift* appears in both biological and artificial systems (Driscoll et al., 2022; Caccia et al., 2022; Theofilou et al., 2025), yet its causal sources are not well understood.

We adopt a probabilistic causal view designed to separate *why* representations move. Inputs $X \in \mathcal{X}$ are generated from latent factors whose distribution is modulated by an environment variable E, while representations are produced by a deterministic encoder $Z = f_\Theta(X)$ whose parameters Θ evolve over time. Treating both E and Θ as nodes in a single SCM yields interventional representation laws $P^Z_{e,\theta}$ and allows drift between two time points to be expressed as contrasts of interventions. This perspective matters in practice: drift monitoring increasingly drives adaptation, rollback, or human intervention, and those actions depend on whether drift is due to environment change, model updates, or their interaction; moreover, causal desiderata for representations (e.g., invariance and non-spuriousness) should be preserved under shift (Wang & Jordan, 2024; Schölkopf et al., 2021).

Our contributions are twofold. First, we derive a causal drift decomposition that attributes representation drift to (i) environment-driven change at fixed parameters and (ii) model-driven change under a fixed environment, without requiring paired pre/post representations. We embed this in a latent-variable generative model and derive a variational objective amenable to amortized inference. Second, we link drift to *downstream performance*: extending the SCM with labels and a fixed predictor head, we give an exact causal risk decomposition and IPM-based bounds on risk change over (Z, Y); under additional stability and smoothness assumptions, we obtain representation-only bounds from drift in Z. Appendices address identifiability of the drift decomposition and statistical consistency of the variational drift estimator, using tools from causal representation learning

and identifiability analyses (Brehmer et al., 2022; Wang & Jordan, 2024; Schölkopf et al., 2021; Mamaghan et al., 2023).

2 Related Work

Concept drift, dataset shift, and feature drift. Nonstationary learning traditionally focuses on shifts in $p(x, y)$ and adaptation/detection methods (Moreno-Torres et al., 2012; Gama et al., 2014; Webb et al., 2018; Hinder et al., 2024), with feature drift addressing changing relevance of input dimensions (Barddal et al., 2017). These frameworks rarely model intermediate representations or separate causal pathways for environment change versus model updates.

Representation drift in monitoring and continual learning. Practical monitoring increasingly measures drift in embedding space; DriftLens detects concept drift via distributional changes in representations (Greco et al., 2024a;b), but does not attribute drift to environment versus model updates. In continual learning, representational drift is linked to forgetting and can be shaped by replay/update rules (Caccia et al., 2022), yet this line is primarily optimization-centric rather than causal/generative.

CRL and identifiability. Causal representation learning (CRL) formalizes desiderata such as invariance, non-spuriousness, and sufficiency, and studies when latent factors and their mechanisms are identifiable from high-dimensional observations (Schölkopf et al., 2021; Wang & Jordan, 2024). Recent work establishes identifiability under weak or indirect supervision, including paired-but-unlabeled intervention settings (Brehmer et al., 2022), and develops generative frameworks with identifiability analyses for rich model classes (Mamaghan et al., 2023). We build on these tools by treating representation drift itself as an interventional object in an SCM with evolving environment E and parameters Θ: this yields a variationally estimable decomposition into environment- and model-driven components, extends naturally to causal risk decompositions and performance bounds, and admits path-specific and order-free (Shapley) drift attributions over time.

3 Theoretical Framework

We formalize *representation drift* as change in the induced representation distribution and introduce a causal model in which both the environment and the model parameters are explicit drivers of representations. We then show how this decomposition can be written in a form that is amenable to variational estimation with amortized inference. Finally, we discuss identifiability conditions under which the decomposition is theoretically well-defined.

3.1 Setup and notation

Let $\mathcal{X}$ be the input space and $\mathcal{Z}$ the representation space, with discrete time $t \in \{1, \dots, T\}$. At each time t we observe an environment index $e_t \in \mathcal{E}$ representing the state of the world (data source, population, task mix, etc.), inputs $X_t \sim p_t(x) := p(x \mid E = e_t)$ drawn from the environment-specific distribution, and model parameters $\theta_t \in \Theta$ inducing a feature map $f_{\theta_t} : \mathcal{X} \to \mathcal{Z}$. The representation at time t is the random variable $Z_t := f_{\theta_t}(X_t)$, with marginal

$$p_t^Z(z) \;:=\; p(Z_t = z) \;=\; \int_{\mathcal{X}} \delta\big(z - f_{\theta_t}(x)\big)\, p_t(x)\, \mathrm{d}x, \tag{1}$$

where $\delta(\cdot)$ denotes the Dirac delta, interpreting the integral as a sum when $\mathcal{X}$ is discrete. We define *representation drift* between $t_1 < t_2$ as any change in the distribution p_t^Z. For notational convenience we write

$$e_1 := e_{t_1}, \qquad e_2 := e_{t_2}, \qquad \theta_1 := \theta_{t_1}, \qquad \theta_2 := \theta_{t_2}, \tag{2}$$

with

$$p_1(x) := p(x \mid E = e_1), \qquad p_2(x) := p(x \mid E = e_2), \tag{3}$$

and corresponding representation marginals p_1^Z and p_2^Z.

3.2 A CAUSAL GENERATIVE MODEL FOR INPUTS AND REPRESENTATIONS

We embed the above quantities into an SCM to make precise what it means to intervene on the environment versus the model. Let E be an environment variable, C latent content factors that encode task-relevant semantics (e.g. object identity), S latent style factors that (e.g. illumination, background), X the observed input, Θ model parameters, and Z the representation. We posit structural equations

$$\begin{aligned} E &= g_E(U_E, t), \\ C &= g_C(E, U_C), \\ S &= g_S(E, U_S), \\ X &= g_X(C, S, E, U_X), \\ \Theta &= g_\Theta(t, U_\Theta), \\ Z &= f_\Theta(X). \end{aligned} \tag{4}$$

with mutually independent exogenous noise variables $U_E, U_C, U_S, U_X, U_\Theta$ and t is the time index. The functions $g_E, g_C, g_S, g_X, g_\Theta$ and f_θ are deterministic. The induced causal graph satisfies

$$(U_E, t) \to E \to (C, S) \to X \to Z, \qquad (U_\Theta, t) \to \Theta \to Z, \tag{5}$$

so that E and Θ are distinct ancestors of Z; this is the basic structural separation that will allow us to disentangle their contributions to representation drift.

For fixed (e, θ) we consider the interventional distribution $p(z \mid \mathrm{do}(E = e, \Theta = \theta))$. The following identity links observational representation marginals to interventions; its derivation is standard and given in Appendix A.

Lemma 3.1 (Interventional representation marginal)**.** *Under equation 4,*

$$p(z \mid \mathrm{do}(E = e, \Theta = \theta)) \;=\; \int \delta\big(z - f_\theta(x)\big)\, p(x \mid E = e)\, \mathrm{d}x. \tag{6}$$

Consequently, the time-t representation marginal satisfies $p_t^Z(z) = p(z \mid \mathrm{do}(E = e_t, \Theta = \theta_t))$.

3.3 CAUSAL DECOMPOSITION OF REPRESENTATION DRIFT

We now define environment-driven and model-driven drift between $t_1 < t_2$.

Total drift. For any measurable $A \subseteq \mathcal{Z}$, where $P_t^Z(A) = P_t^X(f_{\theta_t}^{-1}(A))$ is the pushforward measure, define

$$\begin{aligned} \mu_{\mathrm{tot}}(A) &:= P_2^Z(A) - P_1^Z(A) \\ &= p(Z \in A \mid \mathrm{do}(E = e_2, \Theta = \theta_2)) - p(Z \in A \mid \mathrm{do}(E = e_1, \Theta = \theta_1))\,. \end{aligned} \tag{7}$$

Equivalently (when densities exist),

$$\mu_{\mathrm{tot}}(z) := p_2^Z(z) - p_1^Z(z). \tag{8}$$

Intermediate interventional distributions. Define the counterfactual representation distributions

$$\begin{aligned} p_{\mathrm{env}}^Z(z) &:= p(z \mid \mathrm{do}(E = e_2, \Theta = \theta_1))\,, \\ p_{\mathrm{model}}^Z(z) &:= p(z \mid \mathrm{do}(E = e_1, \Theta = \theta_2))\,. \end{aligned} \tag{9}$$

By Lemma 3.1, these admit the explicit forms

$$p_{\mathrm{env}}^Z(z) = \int \delta\big(z - f_{\theta_1}(x)\big)\, p(x \mid E = e_2)\, \mathrm{d}x = \int \delta\big(z - f_{\theta_1}(x)\big)\, p_2(x)\, \mathrm{d}x, \tag{10}$$

$$p_{\mathrm{model}}^Z(z) = \int \delta\big(z - f_{\theta_2}(x)\big)\, p(x \mid E = e_1)\, \mathrm{d}x = \int \delta\big(z - f_{\theta_2}(x)\big)\, p_1(x)\, \mathrm{d}x. \tag{11}$$

Data-driven and model-driven components. For any measurable $A \subseteq \mathcal{Z}$ define

$$\begin{aligned} \mu_{\mathrm{data}}(A) &:= P_{\mathrm{env}}^Z(A) - P_1^Z(A), \\ \mu_{\mathrm{model}}(A) &:= P_2^Z(A) - P_{\mathrm{env}}^Z(A). \end{aligned} \tag{12}$$

Then, by construction, $\mu_{\mathrm{tot}} = \mu_{\mathrm{data}} + \mu_{\mathrm{model}}$. This is a path-specific, order-dependent, decomposition: we first change the environment ($e_1 \to e_2$) at fixed θ_1, then change the model ($\theta_1 \to \theta_2$) at fixed e_2.

3.4 TEST-FUNCTION VIEW AND ESTIMABLE FUNCTIONALS

The signed measures μ_{data} and μ_{model} are infinite-dimensional objects. In practice we often summarize drift via its effect on expectations of test functions $h : \mathcal{Z} \to \mathbb{R}$:

$$\begin{aligned}\Delta_{\mathrm{tot}}(h) &:= \mathbb{E}_{Z\sim p_2^Z}[h(Z)] - \mathbb{E}_{Z\sim p_1^Z}[h(Z)],\\ \Delta_{\mathrm{data}}(h) &:= \mathbb{E}_{Z\sim p_{\mathrm{env}}^Z}[h(Z)] - \mathbb{E}_{Z\sim p_1^Z}[h(Z)],\\ \Delta_{\mathrm{model}}(h) &:= \mathbb{E}_{Z\sim p_2^Z}[h(Z)] - \mathbb{E}_{Z\sim p_{\mathrm{env}}^Z}[h(Z)].\end{aligned} \tag{13}$$

Linearity yields $\Delta_{\mathrm{tot}}(h) = \Delta_{\mathrm{data}}(h) + \Delta_{\mathrm{model}}(h)$. Moreover, these quantities can be rewritten as expectations over X (derivation in Appendix A):

$$\Delta_{\mathrm{data}}(h) = \mathbb{E}_{X\sim p_2}\big[h(f_{\theta_1}(X))\big] - \mathbb{E}_{X\sim p_1}\big[h(f_{\theta_1}(X))\big]. \tag{14}$$

$$\Delta_{\mathrm{model}}(h) = \mathbb{E}_{X\sim p_2}\big[h(f_{\theta_2}(X))\big] - \mathbb{E}_{X\sim p_2}\big[h(f_{\theta_1}(X))\big]. \tag{15}$$

$$\Delta_{\mathrm{tot}}(h) = \mathbb{E}_{X\sim p_2}\big[h(f_{\theta_2}(X))\big] - \mathbb{E}_{X\sim p_1}\big[h(f_{\theta_1}(X))\big]. \tag{16}$$

These expressions suggest that if we can estimate expectations of the form $\mathbb{E}_{X\sim p_t}[h(f_\theta(X))]$ for different choices of t and θ, then we can estimate the drift components. The main difficulty is that for some combinations (for example $\mathbb{E}_{X\sim p_2}[h(f_{\theta_1}(X))]$) we do not have direct observational data. Section 4 introduces a latent generative model to approximate these counterfactual expectations.

Identifiability and consistency. Our decomposition is defined purely in terms of observable environment-conditional input laws $p(x \mid E = e)$ and known encoders f_θ, and is therefore identifiable once these objects are fixed. We formalize this in Appendix D by showing the drift components are uniquely determined by $(p(x \mid E = e_1), p(x \mid E = e_2), \theta_1, \theta_2)$, independent of latent parameterizations of the underlying SCM. We further analyze the statistical behavior of the variational estimator in Appendix E, giving conditions under which the estimated test-function drifts and induced IPM metrics converge to their population counterparts.

4 LATENT GENERATIVE MODEL AND VARIATIONAL ESTIMATION

We specify a latent-variable generative model for (E, C, S, X) consistent with the SCM in Section 3.2, derive a variational lower bound that enables amortized training, and explain how the learned model yields Monte Carlo approximations of the counterfactual expectations in equation 32–equation 33.

4.1 PARAMETRIC GENERATIVE FAMILY

We introduce a parametric family p_ψ that mirrors equation 4 via the following definitions:

$$p_\psi(e) = p_\psi(E = e), \tag{17}$$

$$p_\psi(c \mid e) = p_\psi(C = c \mid E = e), \tag{18}$$

$$p_\psi(s \mid e) = p_\psi(S = s \mid E = e), \tag{19}$$

$$p_\psi(x \mid c, s, e) = p_\psi(X = x \mid C = c, S = s, E = e). \tag{20}$$

It follows that the joint factorizes as

$$p_\psi(e, c, s, x) = p_\psi(e)\, p_\psi(c \mid e)\, p_\psi(s \mid e)\, p_\psi(x \mid c, s, e). \tag{21}$$

At training time we observe $(x_i, e_i)_{i=1}^N$ and treat (c_i, s_i) as latent. The implied environment-conditional input distribution is

$$p_\psi(x \mid e) = \iint p_\psi(c \mid e)\, p_\psi(s \mid e)\, p_\psi(x \mid c, s, e)\, \mathrm{d}c\, \mathrm{d}s. \tag{22}$$

We choose ψ such that $p_\psi(x \mid e) \approx p(x \mid E = e)$ for each e, enabling Monte Carlo approximation of expectations under $p(x \mid E = e)$.

4.2 VARIATIONAL INFERENCE AND ELBO

Exact maximization of $\log p_\psi(x \mid e)$ is intractable due to the latent integral in equation 22. We therefore introduce a variational posterior family $q_\phi(c, s \mid x, e)$ and use the AEVB framework of Kingma & Welling (2013); Kingma (2017). For a single observation (x, e), we have the variational lower bound $\log p_\psi(x \mid e) \geq \mathcal{L}(x, e; \psi, \phi)$, where the ELBO is (derivation in Appendix B)

$$\mathcal{L}(x, e; \psi, \phi) = \mathbb{E}_{q_\phi(c,s|x,e)}\big[\log p_\psi(x \mid c, s, e) + \log p_\psi(c \mid e) + \log p_\psi(s \mid e) - \log q_\phi(c, s \mid x, e)\big]. \tag{23}$$

For a dataset $\{(x_i, e_i)\}_{i=1}^N$ we maximize $\sum_{i=1}^N \mathcal{L}(x_i, e_i; \psi, \phi)$ with stochastic gradients and amortized inference (encoder parameterization of q_ϕ), using the reparameterization trick to obtain low-variance gradient estimators.

4.3 ESTIMATING DRIFT COMPONENTS VIA THE GENERATIVE MODEL

After training, $p_{\hat{\psi}}(x \mid e)$ approximates $p(x \mid E = e)$ via equation 22, so for any environment e and parameters θ,

$$\mathbb{E}_{X\sim p(\cdot|E=e)}\big[h(f_\theta(X))\big] \;\approx\; \mathbb{E}_{X\sim p_{\hat{\psi}}(\cdot|e)}\big[h(f_\theta(X))\big]. \tag{24}$$

We approximate the right-hand side by Monte Carlo. Using equation 32–equation 33, define

$$\begin{aligned}
\widehat{\Delta}_{\mathrm{data}}(h) &:= \widehat{\mathbb{E}}_{e_2,\theta_1}[h] - \widehat{\mathbb{E}}_{e_1,\theta_1}[h], \\
\widehat{\Delta}_{\mathrm{model}}(h) &:= \widehat{\mathbb{E}}_{e_2,\theta_2}[h] - \widehat{\mathbb{E}}_{e_2,\theta_1}[h], \\
\widehat{\Delta}_{\mathrm{tot}}(h) &:= \widehat{\mathbb{E}}_{e_2,\theta_2}[h] - \widehat{\mathbb{E}}_{e_1,\theta_1}[h].
\end{aligned} \tag{25}$$

By construction, $\widehat{\Delta}_{\mathrm{tot}}(h) = \widehat{\Delta}_{\mathrm{data}}(h) + \widehat{\Delta}_{\mathrm{model}}(h)$ exactly (independent of Monte Carlo error).

5 DRIFT METRICS VIA INTEGRAL PROBABILITY METRICS

Section 3.3 defines drift at the level of interventional representation distributions. Here we summarize drift magnitudes via Integral Probability Metrics (IPMs), yielding complementary *data-driven* and *model-driven* drift distances and a simple relationship to total drift. Let $\mathcal{H}$ be a class of measurable functions $h : \mathcal{Z} \to \mathbb{R}$. The induced IPM between distributions P, Q on $\mathcal{Z}$ is

$$d_{\mathcal{H}}(P, Q) := \sup_{h\in\mathcal{H}} \left|\mathbb{E}_{Z\sim P}[h(Z)] - \mathbb{E}_{Z\sim Q}[h(Z)]\right|. \tag{26}$$

Consider the interventional representation distributions $p_1^Z(z)$, $p_2^Z(z)$, and $p_{\mathrm{env}}^Z(z)$. Define three IPM-based drift metrics:

$$D_{\mathrm{tot}}(\mathcal{H}) := d_{\mathcal{H}}(p_1^Z, p_2^Z), \quad D_{\mathrm{data}}(\mathcal{H}) := d_{\mathcal{H}}(p_1^Z, p_{\mathrm{env}}^Z), \quad D_{\mathrm{model}}(\mathcal{H}) := d_{\mathcal{H}}(p_{\mathrm{env}}^Z, p_2^Z). \tag{27}$$

$D_{\mathrm{data}}(\mathcal{H})$ measures the representation shift induced by changing $e_1 \to e_2$ while holding $\theta = \theta_1$, whereas $D_{\mathrm{model}}(\mathcal{H})$ measures the shift induced by changing $\theta_1 \to \theta_2$ while holding $e = e_2$. The IPM $d_{\mathcal{H}}$ satisfies a triangle inequality (proof in Appendix C.2).

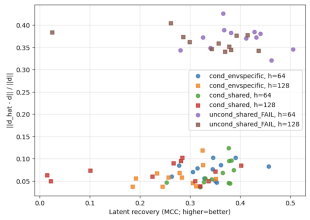
Proposition 5.1 (Drift metric decomposition)**.** *For any function class $\mathcal{H}$, the quantity $D_{\mathrm{tot}}(\mathcal{H}) \leq D_{\mathrm{data}}(\mathcal{H}) + D_{\mathrm{model}}(\mathcal{H})$.*

Proof. Apply Lemma C.1 with $P = p_1^Z$, $Q = p_{\mathrm{env}}^Z$, and $R = p_2^Z$. □

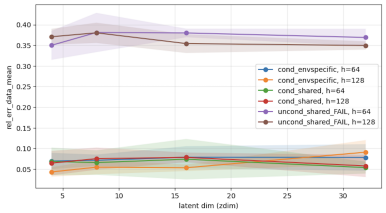
Given the Monte Carlo expectation estimates $\widehat{\mathbb{E}}_{e,\theta}[h]$, define empirical analogues

$$\begin{aligned}
\widehat{D}_{\mathrm{tot}}(\mathcal{H}) &:= \sup_{h\in\mathcal{H}} \left|\widehat{\mathbb{E}}_{e_2,\theta_2}[h] - \widehat{\mathbb{E}}_{e_1,\theta_1}[h]\right|, \\
\widehat{D}_{\mathrm{data}}(\mathcal{H}) &:= \sup_{h\in\mathcal{H}} \left|\widehat{\mathbb{E}}_{e_2,\theta_1}[h] - \widehat{\mathbb{E}}_{e_1,\theta_1}[h]\right|, \\
\widehat{D}_{\mathrm{model}}(\mathcal{H}) &:= \sup_{h\in\mathcal{H}} \left|\widehat{\mathbb{E}}_{e_2,\theta_2}[h] - \widehat{\mathbb{E}}_{e_2,\theta_1}[h]\right|.
\end{aligned} \tag{28}$$

These mirror equation 27 with true expectations replaced by Monte Carlo approximations under $p_{\hat{\psi}}$. In practice, the supremum is computed analytically for some choices of $\mathcal{H}$ or approximated by optimization over a parameterized subclass.



(a) Identifiability: Drift Error vs. MCC



(b) Robustness: Drift Error vs. Latent Dim

Figure 1: **Ablation (data-driven component).** (a) Relative error for Δ_{data} stays low with conditional models even when MCC is low (relative identifiability); the unconditional baseline is high-error throughout. (b) Results are stable across z_{dim}. *Model-driven drift* Δ_{model} *is similar (Appendix).*

6 Ablation: Empirical Validation of Relative Identifiability

Our theoretical framework relies on *relative identifiability* (Proposition D.1): accurate drift decomposition requires matching the environment-conditional marginals $p(x|e)$, but does not require recovering the true latent causal factors (C, S). To validate this empirically, we perform an ablation on a synthetic nonlinear SCM where ground truth drift is known. We systematically vary the generative model's structure and capacity to decouple the quality of latent recovery (disentanglement) from the accuracy of the drift estimate. We generate data (N=12000) from a nonlinear SCM ($X \in \mathbb{R}^{32}$) featuring content/style mixing, non-Gaussian noise, and spurious correlations. We evaluate three variational model families: `Cond-Shared` (our proposed conditional VAE), `Cond-Specific` (environment-specific decoders), and `Uncond-Shared` (a misspecified unconditional baseline). We sweep latent dimensions $z_{\text{dim}} \in \{4, 8, 16, 32\}$ and hidden widths, measuring disentanglement (MCC) and the relative error of the estimated data-driven drift.

Results. Figure 1 demonstrates three key findings. First, drift estimation is decoupled from latent recovery: conditional models frequently achieve low relative error ($< 5\%$) even when disentanglement is poor (MCC < 0.3), supporting our relative identifiability claim. Second, environment conditioning is strictly necessary; the `Uncond-Shared` baseline consistently yields high errors ($> 30\%$) regardless of capacity, confirming that the generative model must explicitly capture the conditional structure $p(x|e)$. Finally, performance is stable across latent capacities ($z_{\text{dim}} \in [4, 32]$), indicating the method is robust to hyperparameter specification and practical for real-world monitoring where the true intrinsic dimension is unknown.

7 Discussion

We elevate *representation drift* from a post-hoc statistic to an identifiable interventional object. By modeling separate causal pathways for environment E and parameters Θ, we distinguish necessary adaptation (D_{data}) from functional geometric changes (D_{model}). This decomposition bridges detection and response. In *compatibility-constrained* settings, risk is dominated by D_{model}, suggesting remedies like stability regularization. Conversely, D_{data} quantifies sensitivity to environmental nuisance. Accurate attribution relies only on *relative identifiability* (matching $p(x|e)$) rather than recovering true latents, ensuring practicality for high-dimensional data. The goal is not zero drift, but to disentangle necessary adaptation from gratuitous instability. The objective is a constrained trade-off: minimize representational change subject to task utility. By rendering drift an estimable causal object, we provide the mechanism to diagnose *why* embeddings moved and guide intervention.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. In *International Conference on Machine Learning*, pp. 214–223. PMLR, 2017.

Jean Paul Barddal, Heitor Murilo Gomes, Fabrício Enembreck, and Bernhard Pfahringer. A survey on feature drift adaptation: Definition, benchmark, challenges and future directions. *Journal of Systems and Software*, 127:278–294, 2017. doi: 10.1016/j.jss.2016.07.005.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38204–38218, 2022.

Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2203.03798.

Laura N. Driscoll, Lea Duncker, and Christopher D. Harvey. Representational drift: Emerging theories for continual learning and experimental future directions. *Current Opinion in Neurobiology*, 76:102609, 2022.

Joao Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Salvatore Greco, Bartolomeo Vacchetti, Daniele Apiletti, and Tania Cerquitelli. Unsupervised concept drift detection from deep learning representations in real-time. *arXiv preprint*, 2024a.

Salvatore Greco, Bartolomeo Vacchetti, Daniele Apiletti, and Tania Cerquitelli. Driftlens: A concept drift detection tool. In *Proceedings of the 27th International Conference on Extending Database Technology (EDBT)*, 2024b.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Fabian Hinder, Valerie Vaquet, and Barbara Hammer. One or two things we know about concept drift—a survey on monitoring in evolving environments. part b: locating and explaining concept drift. *Frontiers in Artificial Intelligence*, 7:1330258, 2024.

Diederik P Kingma. *Variational Inference & Deep Learning: A New Synthesis*. PhD thesis, University of Amsterdam, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Amir Mohammad Karimi Mamaghan, Andrea Dittadi, Stefan Bauer, Karl Henrik Johansson, and Francesco Quinzan. Diffusion based causal representation learning. *arXiv preprint arXiv:2311.05421*, 2023.

José G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012. doi: 10.1016/j.patcog.2011.06.019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Panagiotis Theofilou, Alkiviadis Papaoikonomou, Theodore Antonakopoulos, et al. Stable-drift: A patient-aware latent drift replay method for stabilizing representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

Yixin Wang and Michael I. Jordan. Desiderata for representation learning: A causal perspective. *Journal of Machine Learning Research*, 25:1–65, 2024.

Geoffrey I. Webb, Loong Kuan Lee, Bart Goethals, and François Petitjean. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32(5):1179–1199, 2018.

A Additional Details for Section 3

This appendix contains derivations deferred from Section 3.

A.1 Derivation of Lemma 3.1

For any fixed (e, θ), the interventional distribution is

$$p(z \mid \mathrm{do}(E = e, \Theta = \theta)) = \iint p(z, x \mid \mathrm{do}(E = e, \Theta = \theta))\, \mathrm{d}x.$$

Expanding conditionals,

$$p(z \mid \mathrm{do}(E = e, \Theta = \theta)) = \iint p(z \mid x, \mathrm{do}(E = e, \Theta = \theta))\, p(x \mid \mathrm{do}(E = e, \Theta = \theta))\, \mathrm{d}x. \quad (29)$$

Since $Z = f_\Theta(X)$ deterministically under equation 4,

$$p(z \mid x, \mathrm{do}(E = e, \Theta = \theta)) = \delta\big(z - f_\theta(x)\big).$$

Because Θ has no edge into X in the SCM, intervening on Θ does not affect X, so

$$p(x \mid \mathrm{do}(E = e, \Theta = \theta)) = p(x \mid \mathrm{do}(E = e)).$$

Finally, intervening on E sets $E = e$ in the structural equations, yielding $p(x \mid \mathrm{do}(E = e)) = p(x \mid E = e)$. Substituting gives

$$p(z \mid \mathrm{do}(E = e, \Theta = \theta)) = \int \delta\big(z - f_\theta(x)\big)\, p(x \mid E = e)\, \mathrm{d}x.$$

Setting $(e, \theta) = (e_t, \theta_t)$ yields $p_t^Z(z) = p(z \mid \mathrm{do}(E = e_t, \Theta = \theta_t))$.

A.2 Test-function rewrite of drift components

We show how $\mathbb{E}_{Z \sim p_t^Z}[h(Z)]$ reduces to an expectation over X. Starting from equation 1,

$$\mathbb{E}_{Z \sim p_1^Z}[h(Z)] = \int h(z)\, p_1^Z(z)\, \mathrm{d}z = \int h(z) \left(\int \delta\big(z - f_{\theta_1}(x)\big)\, p_1(x)\, \mathrm{d}x \right) \mathrm{d}z \quad (30)$$

$$= \int \left(\int h(z)\, \delta\big(z - f_{\theta_1}(x)\big)\, \mathrm{d}z \right) p_1(x)\, \mathrm{d}x. \quad (31)$$

The inner integral evaluates to $h(f_{\theta_1}(x))$, hence

$$\mathbb{E}_{Z \sim p_1^Z}[h(Z)] = \int h(f_{\theta_1}(x))\, p_1(x)\, \mathrm{d}x = \mathbb{E}_{X \sim p_1}[h(f_{\theta_1}(X))].$$

Applying the same calculation to p_2^Z and p_{env}^Z (using equation 10) yields

$$\mathbb{E}_{Z \sim p_2^Z}[h(Z)] = \mathbb{E}_{X \sim p_2}[h(f_{\theta_2}(X))], \qquad \mathbb{E}_{Z \sim p_{\mathrm{env}}^Z}[h(Z)] = \mathbb{E}_{X \sim p_2}[h(f_{\theta_1}(X))].$$

Substituting these into the definitions of $\Delta_{\mathrm{data}}(h)$, $\Delta_{\mathrm{model}}(h)$, and $\Delta_{\mathrm{tot}}(h)$ gives the following equations.

$$\Delta_{\mathrm{data}}(h) = \mathbb{E}_{X \sim p_2}\big[h(f_{\theta_1}(X))\big] - \mathbb{E}_{X \sim p_1}\big[h(f_{\theta_1}(X))\big], \quad (32)$$

$$\Delta_{\mathrm{model}}(h) = \mathbb{E}_{X \sim p_2}\big[h(f_{\theta_2}(X))\big] - \mathbb{E}_{X \sim p_2}\big[h(f_{\theta_1}(X))\big], \quad (33)$$

$$\Delta_{\mathrm{tot}}(h) = \mathbb{E}_{X \sim p_2}\big[h(f_{\theta_2}(X))\big] - \mathbb{E}_{X \sim p_1}\big[h(f_{\theta_1}(X))\big]. \quad (34)$$

B Additional Details for Section 4

This appendix provides the full ELBO derivation and makes explicit the Jensen step deferred in the main text.

B.1 ELBO DERIVATION FROM JENSEN'S INEQUALITY

For a single observation (x, e),

$$\log p_\psi(x \mid e) = \log \iint p_\psi(x, c, s \mid e)\, \mathrm{d}c\, \mathrm{d}s. \tag{35}$$

Introduce any density $q_\phi(c, s \mid x, e)$ such that $q_\phi(\cdot, \cdot \mid x, e) > 0$ whenever $p_\psi(x, c, s \mid e) > 0$, and multiply/divide inside the integral:

$$p_\psi(x \mid e) = \iint q_\phi(c, s \mid x, e)\, \frac{p_\psi(x, c, s \mid e)}{q_\phi(c, s \mid x, e)}\, \mathrm{d}c\, \mathrm{d}s. \tag{36}$$

Applying Jensen's inequality to $\log(\cdot)$ (the integrand is non-negative),

$$\log p_\psi(x \mid e) = \log \iint q_\phi(c, s \mid x, e)\, \frac{p_\psi(x, c, s \mid e)}{q_\phi(c, s \mid x, e)}\, \mathrm{d}c\, \mathrm{d}s \tag{37}$$

$$\geq \iint q_\phi(c, s \mid x, e)\, \log \frac{p_\psi(x, c, s \mid e)}{q_\phi(c, s \mid x, e)}\, \mathrm{d}c\, \mathrm{d}s \tag{38}$$

$$= \mathbb{E}_{(c,s)\sim q_\phi(\cdot \mid x, e)} \left[\log p_\psi(x, c, s \mid e) - \log q_\phi(c, s \mid x, e)\right]. \tag{39}$$

Define the right-hand side as the ELBO:

$$\mathcal{L}(x, e; \psi, \phi) := \mathbb{E}_{q_\phi(c,s \mid x,e)} \big[\log p_\psi(x, c, s \mid e) - \log q_\phi(c, s \mid x, e)\big]. \tag{40}$$

Using the conditional factorization

$$p_\psi(x, c, s \mid e) = p_\psi(c \mid e)\, p_\psi(s \mid e)\, p_\psi(x \mid c, s, e), \tag{41}$$

we obtain the expanded form reported in the main text:

$$\mathcal{L}(x, e; \psi, \phi) = \mathbb{E}_{q_\phi(c,s \mid x,e)} \big[\log p_\psi(x \mid c, s, e) + \log p_\psi(c \mid e) + \log p_\psi(s \mid e) - \log q_\phi(c, s \mid x, e)\big]. \tag{42}$$

B.2 DATASET OBJECTIVE AND AMORTIZED INFERENCE

For a dataset $\{(x_i, e_i)\}_{i=1}^N$, the standard variational objective is

$$\mathcal{L}(\psi, \phi) := \sum_{i=1}^{N} \mathcal{L}(x_i, e_i; \psi, \phi),$$

which is optimized with stochastic gradients. In amortized inference, $q_\phi(c, s \mid x, e)$ is parameterized (e.g., by an encoder network) to map (x, e) to a distribution over (c, s). When q_ϕ is reparameterizable, the expectation terms in equation 42 admit low-variance gradient estimators via the reparameterization trick, enabling end-to-end training consistent with Kingma & Welling (2013); Kingma (2017).

B.3 MONTE CARLO APPROXIMATION OF COUNTERFACTUAL EXPECTATIONS

Fix (e, θ) and consider $\mathbb{E}_{X\sim p_{\hat{\psi}}(\cdot \mid e)}[h(f_\theta(X))]$, where

$$p_{\hat{\psi}}(x \mid e) = \iint p_{\hat{\psi}}(c \mid e)\, p_{\hat{\psi}}(s \mid e)\, p_{\hat{\psi}}(x \mid c, s, e)\, \mathrm{d}c\, \mathrm{d}s.$$

A Monte Carlo estimator is obtained by ancestral sampling:

$$(c^{(k)}, s^{(k)}) \sim p_{\hat{\psi}}(c \mid e) p_{\hat{\psi}}(s \mid e), \qquad x^{(k)} \sim p_{\hat{\psi}}(\cdot \mid c^{(k)}, s^{(k)}, e),$$

and then averaging $h(f_\theta(x^{(k)}))$:

$$\widehat{\mathbb{E}}_{e,\theta}[h] := \frac{1}{K} \sum_{k=1}^{K} h\big(f_\theta(x^{(k)})\big).$$

Under standard integrability assumptions, $\widehat{\mathbb{E}}_{e,\theta}[h] \to \mathbb{E}_{X\sim p_{\hat{\psi}}(\cdot \mid e)}[h(f_\theta(X))]$ almost surely as $K \to \infty$ by the law of large numbers.

C Additional Details for Section 5

C.1 Background and examples of IPMs

IPMs recover standard distances by suitable choices of $\mathcal{H}$. For example, if $\mathcal{H}$ is the unit ball of an RKHS, then $d_{\mathcal{H}}$ is the Maximum Mean Discrepancy (Gretton et al., 2012); if $\mathcal{H}$ is the set of 1-Lipschitz functions, then $d_{\mathcal{H}}$ coincides with the Wasserstein-1 distance (Arjovsky et al., 2017).

C.2 Triangle inequality for IPMs

Lemma C.1 (Triangle inequality for IPMs)**.** *Let P, Q, and R be probability distributions on $\mathcal{Z}$. Then*

$$d_{\mathcal{H}}(P, R) \leq d_{\mathcal{H}}(P, Q) + d_{\mathcal{H}}(Q, R). \tag{43}$$

Proof. By definition,

$$d_{\mathcal{H}}(P, R) = \sup_{h \in \mathcal{H}} \left| \mathbb{E}_P[h(Z)] - \mathbb{E}_R[h(Z)] \right|. \tag{44}$$

For any fixed $h \in \mathcal{H}$, add and subtract $\mathbb{E}_Q[h(Z)]$:

$$\begin{aligned} |\mathbb{E}_P[h(Z)] - \mathbb{E}_R[h(Z)]| &= |\mathbb{E}_P[h(Z)] - \mathbb{E}_Q[h(Z)] + \mathbb{E}_Q[h(Z)] - \mathbb{E}_R[h(Z)]| && (45) \\ &\leq |\mathbb{E}_P[h(Z)] - \mathbb{E}_Q[h(Z)]| + |\mathbb{E}_Q[h(Z)] - \mathbb{E}_R[h(Z)]|, && (46) \end{aligned}$$

by the triangle inequality on $\mathbb{R}$. Taking the supremum over $h \in \mathcal{H}$ yields

$$\begin{aligned} d_{\mathcal{H}}(P, R) &= \sup_{h \in \mathcal{H}} |\mathbb{E}_P[h(Z)] - \mathbb{E}_R[h(Z)]| && (47) \\ &\leq \sup_{h \in \mathcal{H}} |\mathbb{E}_P[h(Z)] - \mathbb{E}_Q[h(Z)]| + \sup_{h \in \mathcal{H}} |\mathbb{E}_Q[h(Z)] - \mathbb{E}_R[h(Z)]| && (48) \\ &= d_{\mathcal{H}}(P, Q) + d_{\mathcal{H}}(Q, R). && (49) \end{aligned}$$

□

C.3 When the drift bound is tight

Applying Lemma C.1 with $P = p_1^Z$, $Q = p_{\text{env}}^Z$, and $R = p_2^Z$ yields Proposition 5.1. The inequality in 5.1 becomes an equality whenever there exists $h^\star \in \mathcal{H}$ such that: (i) $h^\star$ attains the supremum simultaneously for each pair $(p_1^Z, p_{\text{env}}^Z)$, $(p_{\text{env}}^Z, p_2^Z)$, and (p_1^Z, p_2^Z), and (ii) the corresponding signed differences in expectations have aligned signs so that the real-valued triangle inequality is tight for $h^\star$. In general such a common maximizer need not exist, so additivity is not guaranteed, but $D_{\text{data}}(\mathcal{H})$ and $D_{\text{model}}(\mathcal{H})$ remain interpretable controls on total drift.

D Identifiability of the Drift Decomposition

We now discuss conditions under which the decomposition of representation drift into data-driven and model-driven components is *identifiable*, in the sense that it is uniquely determined by the joint distribution of observed variables and the sequence of model parameters.

D.1 Problem formulation

Consider two SCMs, $\mathcal{M}$ and $\mathcal{M}'$, each of the form defined in Section 3.2, possibly with different hidden latent variables and structural functions, but sharing the same observable variables (E, X, Θ, Z) and the same representation functions f_θ. Suppose that:

1. For every environment value $e \in \mathcal{E}$, the environment-conditional input distributions are identical:
$$p_{\mathcal{M}}(x \mid E = e) = p_{\mathcal{M}'}(x \mid E = e) \quad \text{for all } x \in \mathcal{X}. \tag{50}$$
2. The sequences of model parameters at times t_1 and t_2 are the same in both models:
$$\theta_1^{\mathcal{M}} = \theta_1^{\mathcal{M}'} = \theta_1, \qquad \theta_2^{\mathcal{M}} = \theta_2^{\mathcal{M}'} = \theta_2. \tag{51}$$

We ask whether the induced drift components μ_{data} and μ_{model} are necessarily the same in both models.

D.2 RELATIVE IDENTIFIABILITY OF DRIFT COMPONENTS

We first state a basic result showing that the decomposition is determined entirely by the observable quantities $p(x \mid E = e_1)$, $p(x \mid E = e_2)$, θ_1, and θ_2.

Proposition D.1 (Relative identifiability of drift components)**.** *Under assumptions equation 50–equation 51, the data-driven and model-driven drift measures are identical in $\mathcal{M}$ and $\mathcal{M}'$. That is, for any measurable set $A \subseteq \mathcal{Z}$,*

$$\mu_{\text{data}}^{\mathcal{M}}(A) = \mu_{\text{data}}^{\mathcal{M}'}(A), \tag{52}$$

$$\mu_{\text{model}}^{\mathcal{M}}(A) = \mu_{\text{model}}^{\mathcal{M}'}(A), \tag{53}$$

and hence the total drift decomposition is the same in both models.

Proof. We prove the statement for μ_{data}; the proof for μ_{model} is analogous.

By definition (see equation 12), $\mu_{\text{data}}(A)$ is the difference

$$\mu_{\text{data}}(A) = p_{\text{env}}^{Z}(A) - p_1^{Z}(A), \tag{54}$$

where p_{env}^{Z} and p_1^{Z} are the distributions in equation 10 and equation 1. In model $\mathcal{M}$, we have

$$p_{\text{env}}^{Z,\mathcal{M}}(A) = \int_{\mathcal{X}} \mathbf{1}\{f_{\theta_1}(x) \in A\}\, p_{\mathcal{M}}(x \mid E = e_2)\, \mathrm{d}x, \tag{55}$$

$$p_1^{Z,\mathcal{M}}(A) = \int_{\mathcal{X}} \mathbf{1}\{f_{\theta_1}(x) \in A\}\, p_{\mathcal{M}}(x \mid E = e_1)\, \mathrm{d}x, \tag{56}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. The same expressions hold in model $\mathcal{M}'$, with $p_{\mathcal{M}}$ replaced by $p_{\mathcal{M}'}$. By assumption equation 50,

$$p_{\mathcal{M}}(x \mid E = e_j) = p_{\mathcal{M}'}(x \mid E = e_j) \quad \text{for all } x \in \mathcal{X},\ j \in \{1, 2\}. \tag{57}$$

Since the integrands are equal pointwise in x, the integrals are equal:

$$p_{\text{env}}^{Z,\mathcal{M}}(A) = p_{\text{env}}^{Z,\mathcal{M}'}(A), \tag{58}$$

$$p_1^{Z,\mathcal{M}}(A) = p_1^{Z,\mathcal{M}'}(A). \tag{59}$$

Therefore,

$$\mu_{\text{data}}^{\mathcal{M}}(A) = p_{\text{env}}^{Z,\mathcal{M}}(A) - p_1^{Z,\mathcal{M}}(A) = p_{\text{env}}^{Z,\mathcal{M}'}(A) - p_1^{Z,\mathcal{M}'}(A) = \mu_{\text{data}}^{\mathcal{M}'}(A). \tag{60}$$

This holds for all measurable sets $A \subseteq \mathcal{Z}$. The argument for μ_{model} uses the analogous expressions with p_2^{Z} and p_{env}^{Z} and the same equality of environment-conditional input distributions. Finally, the total drift measure satisfies $\mu_{\text{tot}} = \mu_{\text{data}} + \mu_{\text{model}}$ in both models, so it is also identical. □

Proposition D.1 shows that the decomposition we propose is uniquely determined by observable environment-conditional input distributions and the model parameters. In particular, any two SCMs that agree on these observable objects will agree on the data-driven and model-driven components of drift, even if they differ in their latent variable parameterizations.

D.3 DISCUSSION OF ABSOLUTE IDENTIFIABILITY

The result above is a *relative* identifiability statement: it shows that once we fix the observable joint distribution of (E, X) and the model parameters (θ_1, θ_2), the drift decomposition is uniquely determined. It does not, by itself, guarantee that the environment-conditional distributions $p(x \mid E = e)$ or the model parameters are identifiable from finite data.

Absolute identifiability of the latent variables (C, S) and their causal relations generally requires additional assumptions, such as multi-environment diversity, constrained functional forms, or access to interventions or auxiliary variables. Recent work on causal representation learning provides conditions under which latent causal variables can be recovered (up to certain equivalence classes) from high-dimensional observations. In our setting, such results can be used to argue that the factors

through which the environment influences the inputs, and hence the way environment changes induce data-driven drift, can be interpreted in a stable, approximately invariant manner across time. A full treatment of latent identifiability in our model would require additional structural assumptions and is deferred to an extended version of this work.

From the perspective of drift decomposition, however, the key point is that the quantities entering the definitions of μ_{data} and μ_{model} are *observable in principle*: they depend only on $p(x \mid E = e)$ and on the known representation maps f_{θ_t}. Our variational framework provides a practical way to approximate $p(x \mid E = e)$ when some of the cross-environment, cross-model expectations needed in equation 32–equation 33 cannot be directly computed from logged data.

E Statistical Consistency of the Variational Drift Estimator

We now study when the estimated drift components and drift metrics converge to their population counterparts. We focus on the estimators induced by the variational generative model in Section 4. The main idea is to separate the error into (i) a *modeling error* due to approximating $p(x \mid E = e)$ by $p_{\hat{\psi}}(x \mid e)$, and (ii) a *Monte Carlo error* due to approximating expectations by finite samples.

E.1 Assumptions

We state a set of assumptions under which consistency holds. These are standard in the analysis of variational and generative models.

(A1) Well-specified environment model. For each environment value $e \in \mathcal{E}$, there exists a parameter vector $\psi^\star$ such that

$$p_{\psi^\star}(x \mid e) = p(x \mid E = e) \quad \text{for all } x \in \mathcal{X}. \tag{61}$$

That is, the true environment-conditional input distributions lie in the model family.

(A2) Consistent estimation of $\psi^\star$. Let $\hat{\psi}_N$ be the maximizer (or approximate maximizer) of the empirical ELBO $\mathcal{L}(\psi, \phi)$ based on N i.i.d. samples. As $N \to \infty$,

$$\hat{\psi}_N \to \psi^\star \tag{62}$$

in probability.

(A3) Bounded test functions. The function class $\mathcal{H}$ consists of measurable functions $h : \mathcal{Z} \to \mathbb{R}$ that are uniformly bounded: there exists $B < \infty$ such that $|h(z)| \le B$ for all $h \in \mathcal{H}$ and all $z \in \mathcal{Z}$.

(A4) Monte Carlo sampling. For each environment e and parameter vector θ, Monte Carlo expectations are estimated using K i.i.d. samples $X^{(1)}, \dots, X^{(K)} \sim p_{\hat{\psi}_N}(\cdot \mid e)$. As $N \to \infty$, we also let $K = K(N) \to \infty$.

Assumption (A1) is a standard well-specification condition; it guarantees that there is a "best" parameter vector $\psi^\star$ that matches the true environment-conditional distributions exactly. Assumption (A2) states that the variational learning procedure recovers this parameter in the large data limit. Assumption (A3) ensures that all expectations are finite and that uniform convergence over $\mathcal{H}$ is possible. Assumption (A4) allows us to control the Monte Carlo error via the law of large numbers.

E.2 Pointwise consistency for fixed test functions

We first prove consistency for the drift components evaluated at a fixed test function $h \in \mathcal{H}$.

Theorem E.1 (Pointwise consistency of drift components)**.** *Fix a test function $h \in \mathcal{H}$ satisfying (A3). Under assumptions (A1)–(A4), the following convergence in probability holds as $N \to \infty$:*

$$\widehat{\Delta}_{\text{data}}(h) \xrightarrow{p} \Delta_{\text{data}}(h), \tag{63}$$

$$\widehat{\Delta}_{\text{model}}(h) \xrightarrow{p} \Delta_{\text{model}}(h), \tag{64}$$

$$\widehat{\Delta}_{\text{tot}}(h) \xrightarrow{p} \Delta_{\text{tot}}(h). \tag{65}$$

Proof. We prove the statement for $\widehat{\Delta}_{\mathrm{data}}(h)$; the arguments for $\widehat{\Delta}_{\mathrm{model}}(h)$ and $\widehat{\Delta}_{\mathrm{tot}}(h)$ are analogous.

Recall that

$$\Delta_{\mathrm{data}}(h) = \mathbb{E}_{X\sim p_2}[h(f_{\theta_1}(X))] - \mathbb{E}_{X\sim p_1}[h(f_{\theta_1}(X))], \tag{66}$$

and the estimator is

$$\widehat{\Delta}_{\mathrm{data}}(h) = \widehat{\mathbb{E}}_{e_2,\theta_1}[h] - \widehat{\mathbb{E}}_{e_1,\theta_1}[h], \tag{67}$$

where

$$\widehat{\mathbb{E}}_{e,\theta}[h] = \frac{1}{K}\sum_{k=1}^{K} h(f_\theta(X^{(k)})), \qquad X^{(k)} \sim p_{\hat{\psi}_N}(\cdot \mid e) \text{ i.i.d.} \tag{68}$$

We decompose the error as

$$\widehat{\Delta}_{\mathrm{data}}(h) - \Delta_{\mathrm{data}}(h) = \left(\widehat{\mathbb{E}}_{e_2,\theta_1}[h] - \mathbb{E}_{X\sim p_2}[h(f_{\theta_1}(X))]\right) - \left(\widehat{\mathbb{E}}_{e_1,\theta_1}[h] - \mathbb{E}_{X\sim p_1}[h(f_{\theta_1}(X))]\right). \tag{69}$$

Using the triangle inequality,

$$\left|\widehat{\Delta}_{\mathrm{data}}(h) - \Delta_{\mathrm{data}}(h)\right| \le \left|\widehat{\mathbb{E}}_{e_2,\theta_1}[h] - \mathbb{E}_{X\sim p_2}[h(f_{\theta_1}(X))]\right| + \left|\widehat{\mathbb{E}}_{e_1,\theta_1}[h] - \mathbb{E}_{X\sim p_1}[h(f_{\theta_1}(X))]\right|. \tag{70}$$

We now analyze a generic term of the form

$$\left|\widehat{\mathbb{E}}_{e,\theta_1}[h] - \mathbb{E}_{X\sim p(\cdot|E=e)}[h(f_{\theta_1}(X))]\right|. \tag{71}$$

Add and subtract the expectation under $p_{\hat{\psi}_N}(\cdot \mid e)$:

$$\left|\widehat{\mathbb{E}}_{e,\theta_1}[h] - \mathbb{E}_{X\sim p(\cdot|E=e)}[h(f_{\theta_1}(X))]\right| \tag{72}$$

$$\le \left|\widehat{\mathbb{E}}_{e,\theta_1}[h] - \mathbb{E}_{X\sim p_{\hat{\psi}_N}(\cdot|e)}[h(f_{\theta_1}(X))]\right| + \left|\mathbb{E}_{X\sim p_{\hat{\psi}_N}(\cdot|e)}[h(f_{\theta_1}(X))] - \mathbb{E}_{X\sim p(\cdot|E=e)}[h(f_{\theta_1}(X))]\right|. \tag{73}$$

The first term in equation 73 is a Monte Carlo error. Conditional on $\hat{\psi}_N$, the random variables $h(f_{\theta_1}(X^{(1)})), \ldots, h(f_{\theta_1}(X^{(K)}))$ are i.i.d. with finite mean and bounded by B in absolute value by (A3). By the weak law of large numbers,

$$\widehat{\mathbb{E}}_{e,\theta_1}[h] \xrightarrow{p} \mathbb{E}_{X\sim p_{\hat{\psi}_N}(\cdot|e)}[h(f_{\theta_1}(X))] \quad \text{as } K \to \infty, \tag{74}$$

so the first term converges to zero in probability as $K \to \infty$.

The second term in equation 73 is a modeling error. By (A1), there exists $\psi^\star$ such that $p_{\psi^\star}(x \mid e) = p(x \mid E = e)$ for all x. By (A2), $\hat{\psi}_N \to \psi^\star$ in probability as $N \to \infty$. Under mild continuity conditions on the mapping $\psi \mapsto p_\psi(x \mid e)$ and dominated convergence (ensured by boundedness of h and f_{θ_1}), this implies

$$\mathbb{E}_{X\sim p_{\hat{\psi}_N}(\cdot|e)}[h(f_{\theta_1}(X))] \xrightarrow{p} \mathbb{E}_{X\sim p_{\psi^\star}(\cdot|e)}[h(f_{\theta_1}(X))] = \mathbb{E}_{X\sim p(\cdot|E=e)}[h(f_{\theta_1}(X))], \tag{75}$$

so the second term also converges to zero in probability as $N \to \infty$.

Combining both terms, we see that for each $e \in \{e_1, e_2\}$,

$$\left|\widehat{\mathbb{E}}_{e,\theta_1}[h] - \mathbb{E}_{X\sim p(\cdot|E=e)}[h(f_{\theta_1}(X))]\right| \xrightarrow{p} 0 \quad \text{as } N \to \infty. \tag{76}$$

Substituting into equation 70 and using the fact that the sum of two random variables that converge to zero in probability also converges to zero in probability, we obtain

$$\widehat{\Delta}_{\mathrm{data}}(h) - \Delta_{\mathrm{data}}(h) \xrightarrow{p} 0 \quad \text{as } N \to \infty, \tag{77}$$

which proves the first claim. The proofs for $\widehat{\Delta}_{\mathrm{model}}(h)$ and $\widehat{\Delta}_{\mathrm{tot}}(h)$ follow exactly the same structure, with different choices of (e, θ) in the expectations. □

E.3 UNIFORM CONSISTENCY OVER FUNCTION CLASSES AND IPM CONVERGENCE

We now extend the pointwise result to uniform convergence over the function class $\mathcal{H}$. This allows us to deduce consistency of the IPM-based drift metrics defined in Section 5.

Theorem E.2 (Uniform consistency over $\mathcal{H}$)**.** *In addition to (A1)–(A4), suppose that $\mathcal{H}$ is a Glivenko–Cantelli class with respect to the family of distributions $\{p(\cdot \mid E = e) : e \in \{e_1, e_2\}\}$ and $\{p_\psi(\cdot \mid e) : \psi$ in a neighborhood of $\psi^\star\}$. Then*

$$\sup_{h\in\mathcal{H}} \left|\widehat{\Delta}_{\mathrm{data}}(h) - \Delta_{\mathrm{data}}(h)\right| \xrightarrow{p} 0, \tag{78}$$

$$\sup_{h\in\mathcal{H}} \left|\widehat{\Delta}_{\mathrm{model}}(h) - \Delta_{\mathrm{model}}(h)\right| \xrightarrow{p} 0, \tag{79}$$

$$\sup_{h\in\mathcal{H}} \left|\widehat{\Delta}_{\mathrm{tot}}(h) - \Delta_{\mathrm{tot}}(h)\right| \xrightarrow{p} 0 \tag{80}$$

as $N \to \infty$.

Proof. We prove the first claim; the other two follow by the same argument with the appropriate substitutions.

Recall

$$\Delta_{\mathrm{data}}(h) = \mathbb{E}_{X\sim p(\cdot|E=e_2)}\big[h(f_{\theta_1}(X))\big] - \mathbb{E}_{X\sim p(\cdot|E=e_1)}\big[h(f_{\theta_1}(X))\big],$$

and

$$\widehat{\Delta}_{\mathrm{data}}(h) = \widehat{\mathbb{E}}_{e_2,\theta_1}[h] - \widehat{\mathbb{E}}_{e_1,\theta_1}[h], \qquad \widehat{\mathbb{E}}_{e,\theta_1}[h] = \frac{1}{K}\sum_{k=1}^{K} h\big(f_{\theta_1}(X_e^{(k)})\big),$$

where conditionally on $\hat{\psi}_N$, the samples $X_e^{(1)}, \ldots, X_e^{(K)}$ are i.i.d. from $p_{\hat{\psi}_N}(\cdot \mid e)$.

Define the shorthand class of composed functions

$$\mathcal{G}_{\theta_1} := \{\, g_h(\cdot) := h(f_{\theta_1}(\cdot)) \;:\; h \in \mathcal{H} \,\}.$$

Since $|h| \le B$ on $\mathcal{Z}$ by (A3), we have $|g_h| \le B$ on $\mathcal{X}$ for all $h \in \mathcal{H}$.

We start from the uniform error and apply the triangle inequality:

$$\begin{aligned}\sup_{h\in\mathcal{H}}\big|\widehat{\Delta}_{\mathrm{data}}(h) - \Delta_{\mathrm{data}}(h)\big| &= \sup_{h\in\mathcal{H}}\Big|\big(\widehat{\mathbb{E}}_{e_2,\theta_1}[h] - \mathbb{E}_{p(\cdot|e_2)}[g_h]\big) - \big(\widehat{\mathbb{E}}_{e_1,\theta_1}[h] - \mathbb{E}_{p(\cdot|e_1)}[g_h]\big)\Big| \\ &\le \sup_{h\in\mathcal{H}}\big|\widehat{\mathbb{E}}_{e_2,\theta_1}[h] - \mathbb{E}_{p(\cdot|e_2)}[g_h]\big| + \sup_{h\in\mathcal{H}}\big|\widehat{\mathbb{E}}_{e_1,\theta_1}[h] - \mathbb{E}_{p(\cdot|e_1)}[g_h]\big|.\end{aligned} \tag{81}$$

It therefore suffices to show that for each fixed $e \in \{e_1, e_2\}$,

$$\sup_{h\in\mathcal{H}}\big|\widehat{\mathbb{E}}_{e,\theta_1}[h] - \mathbb{E}_{p(\cdot|e)}[g_h]\big| \xrightarrow{p} 0. \tag{82}$$

Fix such an e. Add and subtract the expectation under the fitted model distribution $p_{\hat{\psi}_N}(\cdot \mid e)$ and use the triangle inequality:

$$\sup_{h\in\mathcal{H}}\big|\widehat{\mathbb{E}}_{e,\theta_1}[h] - \mathbb{E}_{p(\cdot|e)}[g_h]\big| \le \sup_{h\in\mathcal{H}}\big|\widehat{\mathbb{E}}_{e,\theta_1}[h] - \mathbb{E}_{p_{\hat{\psi}_N}(\cdot|e)}[g_h]\big| + \sup_{h\in\mathcal{H}}\big|\mathbb{E}_{p_{\hat{\psi}_N}(\cdot|e)}[g_h] - \mathbb{E}_{p(\cdot|e)}[g_h]\big|. \tag{83}$$

We show that both terms on the right converge to 0 in probability.

Monte Carlo (uniform) term. Condition on $\hat{\psi}_N$. Then $\widehat{\mathbb{E}}_{e,\theta_1}[h]$ is the empirical mean of $g_h(X)$ over i.i.d. samples $X \sim p_{\hat{\psi}_N}(\cdot \mid e)$, and the first term in equation 83 is precisely the uniform deviation of empirical means from expectations over the class $\mathcal{G}_{\theta_1}$ under the distribution $p_{\hat{\psi}_N}(\cdot \mid e)$. By the assumption of Theorem E.2, $\mathcal{H}$ is Glivenko–Cantelli with respect to the family $\{p_\psi(\cdot \mid e) : \psi$ in a neighborhood of $\psi^\star\}$; since f_{θ_1} is fixed and measurable, the composed class $\mathcal{G}_{\theta_1}$ inherits the

same uniform Glivenko–Cantelli property for these distributions (equivalently, $\mathcal{G}_{\theta_1}$ is Glivenko–Cantelli for each $p_\psi(\cdot \mid e)$ in that neighborhood). Hence, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left| \widehat{\mathbb{E}}_{e,\theta_1}[h] - \mathbb{E}_{p_{\hat{\psi}_N}(\cdot|e)}[g_h] \right| > \varepsilon \;\middle|\; \hat{\psi}_N \right) \longrightarrow 0 \qquad \text{as } K \to \infty, \tag{84}$$

whenever $\hat{\psi}_N$ lies in the specified neighborhood. By (A2), $\hat{\psi}_N \to \psi^\star$ in probability, so $\mathbb{P}(\hat{\psi}_N \text{ lies in the neighborhood}) \to 1$. Combining this with equation 84 and $K = K(N) \to \infty$ from (A4) yields the unconditional convergence

$$\sup_{h\in\mathcal{H}} \left| \widehat{\mathbb{E}}_{e,\theta_1}[h] - \mathbb{E}_{p_{\hat{\psi}_N}(\cdot|e)}[g_h] \right| \xrightarrow{p} 0 \qquad \text{as } N \to \infty. \tag{85}$$

Modeling term. Consider the second term in equation 83. By (A1), $p_{\psi^\star}(\cdot \mid e) = p(\cdot \mid E = e)$, so it suffices to show

$$\sup_{h\in\mathcal{H}} \left| \mathbb{E}_{p_{\hat{\psi}_N}(\cdot|e)}[g_h] - \mathbb{E}_{p_{\psi^\star}(\cdot|e)}[g_h] \right| \xrightarrow{p} 0. \tag{86}$$

By the Glivenko–Cantelli assumption in the theorem statement, $\mathcal{H}$ is Glivenko–Cantelli uniformly over the family $\{p_\psi(\cdot \mid e) : \psi \text{ near } \psi^\star\}$. In particular, the map

$$\psi \;\mapsto\; \sup_{h\in\mathcal{H}} \left| \mathbb{E}_{p_{\psi}(\cdot|e)}[g_h] - \mathbb{E}_{p_{\psi^\star}(\cdot|e)}[g_h] \right|$$

is continuous at $\psi^\star$ under the topology induced by this uniform law of large numbers (e.g., continuity in total variation suffices; more generally it is enough that $\psi \mapsto p_\psi(\cdot \mid e)$ is continuous in the metric $d_{\mathcal{G}_{\theta_1}}(P, Q) := \sup_{h\in\mathcal{H}} |\mathbb{E}_P[g_h] - \mathbb{E}_Q[g_h]|$). Since $\hat{\psi}_N \to \psi^\star$ in probability by (A2), the continuous mapping theorem yields equation 86.

Combining equation 85 and equation 86 in equation 83 establishes equation 82. Substituting equation 82 for $e = e_1$ and $e = e_2$ into equation 81 gives

$$\sup_{h\in\mathcal{H}} \left| \widehat{\Delta}_{\text{data}}(h) - \Delta_{\text{data}}(h) \right| \xrightarrow{p} 0.$$

The proofs for $\widehat{\Delta}_{\text{model}}$ and $\widehat{\Delta}_{\text{tot}}$ are identical, replacing $(e_1, \theta_1), (e_2, \theta_1)$ by $(e_2, \theta_1), (e_2, \theta_2)$ and by $(e_1, \theta_1), (e_2, \theta_2)$, respectively, throughout. This yields the stated uniform convergence for all three drift components. □

An immediate corollary is consistency of the IPM-based drift metrics.

Corollary E.3 (Consistency of IPM drift metrics)**.** *Under the assumptions of Theorem E.2,*

$$\widehat{D}_{\text{data}}(\mathcal{H}) \xrightarrow{p} D_{\text{data}}(\mathcal{H}), \tag{87}$$

$$\widehat{D}_{\text{model}}(\mathcal{H}) \xrightarrow{p} D_{\text{model}}(\mathcal{H}), \tag{88}$$

$$\widehat{D}_{\text{tot}}(\mathcal{H}) \xrightarrow{p} D_{\text{tot}}(\mathcal{H}). \tag{89}$$

Proof. By definition of the IPM,

$$D_{\text{data}}(\mathcal{H}) = \sup_{h\in\mathcal{H}} |\Delta_{\text{data}}(h)|, \qquad \widehat{D}_{\text{data}}(\mathcal{H}) = \sup_{h\in\mathcal{H}} |\widehat{\Delta}_{\text{data}}(h)|. \tag{90}$$

For any $h \in \mathcal{H}$,

$$\left| |\widehat{\Delta}_{\text{data}}(h)| - |\Delta_{\text{data}}(h)| \right| \le \left| \widehat{\Delta}_{\text{data}}(h) - \Delta_{\text{data}}(h) \right|. \tag{91}$$

Taking the supremum over $h \in \mathcal{H}$ on both sides yields

$$\left| \widehat{D}_{\text{data}}(\mathcal{H}) - D_{\text{data}}(\mathcal{H}) \right| \le \sup_{h\in\mathcal{H}} \left| \widehat{\Delta}_{\text{data}}(h) - \Delta_{\text{data}}(h) \right|. \tag{92}$$

By Theorem E.2, the right-hand side converges to zero in probability, so the left-hand side does as well. The proofs for $\widehat{D}_{\text{model}}(\mathcal{H})$ and $\widehat{D}_{\text{tot}}(\mathcal{H})$ are identical with the appropriate substitutions. □

Taken together, Theorems E.1 and E.2 and Corollary E.3 show that our variational generative framework yields statistically consistent estimators of both the *causal drift components* for individual test functions and the *IPM-based drift metrics* that summarize these components across a function class. As a result, the data-driven and model-driven contributions to representation drift are not only conceptually well-defined but also estimable in a principled and asymptotically sound manner.

F ADDITIONAL DETAILS FOR SECTION 6

Table 1 details the numerical results for the model-driven drift component Δ_{model}, which mirror the data-driven trends from this experiment. We observe a clear decoupling of drift estimation from latent recovery: conditional models (e.g., `cond_envspecific`) consistently achieve low errors (typically < 0.08) even when disentanglement is poor (e.g., MCC ≈ 0.19 in Run 12). Furthermore, performance remains stable across varying model capacities ($z_{\text{dim}} \in \{4, \ldots, 32\}$), whereas the misspecified `uncond_shared_FAIL` baseline exhibits errors an order of magnitude higher (≈ 0.30), confirming that environment-conditional structure is strictly necessary for identifiability.

Table 1: Full numerical results for the ablation study. **MCC**: Mean Correlation Coefficient (latent recovery). **Rel Err (Model)**: Relative error in estimating Δ_{model}. **Recon**: VAE reconstruction loss.

ID	Model Spec	Hidden	Z-Dim	MCC	Rel Err (Model)	Recon
0	cond_envspecific	64	4	0.345	0.060	0.356
1	cond_envspecific	64	4	0.350	0.040	0.355
2	cond_envspecific	64	4	0.457	0.066	0.355
3	cond_envspecific	64	8	0.264	0.048	0.356
4	cond_envspecific	64	8	0.278	0.067	0.356
5	cond_envspecific	64	8	0.305	0.055	0.356
6	cond_envspecific	64	16	0.315	0.062	0.356
7	cond_envspecific	64	16	0.328	0.044	0.356
8	cond_envspecific	64	16	0.346	0.081	0.355
9	cond_envspecific	64	32	0.354	0.037	0.356
10	cond_envspecific	64	32	0.362	0.068	0.355
11	cond_envspecific	64	32	0.391	0.081	0.355
12	cond_envspecific	128	4	0.185	0.029	0.355
13	cond_envspecific	128	4	0.245	0.030	0.356
14	cond_envspecific	128	4	0.351	0.043	0.356
15	cond_envspecific	128	8	0.234	0.054	0.356
16	cond_envspecific	128	8	0.257	0.046	0.356
17	cond_envspecific	128	8	0.311	0.030	0.356
18	cond_envspecific	128	16	0.191	0.045	0.356
19	cond_envspecific	128	16	0.284	0.046	0.356
20	cond_envspecific	128	16	0.306	0.036	0.356
21	cond_envspecific	128	32	0.279	0.054	0.356
22	cond_envspecific	128	32	0.325	0.094	0.356
23	cond_envspecific	128	32	0.326	0.068	0.356
24	cond_shared	64	4	0.320	0.029	0.356
25	cond_shared	64	4	0.378	0.075	0.356
26	cond_shared	64	4	0.387	0.059	0.356
27	cond_shared	64	8	0.253	0.037	0.356
28	cond_shared	64	8	0.331	0.044	0.356
29	cond_shared	64	8	0.383	0.076	0.355
30	cond_shared	64	16	0.343	0.043	0.356
31	cond_shared	64	16	0.377	0.098	0.356
32	cond_shared	64	16	0.380	0.035	0.356
33	cond_shared	64	32	0.328	0.038	0.356
34	cond_shared	64	32	0.365	0.054	0.356
35	cond_shared	64	32	0.377	0.036	0.356
36	cond_shared	128	4	0.021	0.039	0.356
37	cond_shared	128	4	0.282	0.075	0.356
38	cond_shared	128	4	0.310	0.039	0.356
39	cond_shared	128	8	0.014	0.050	0.356
40	cond_shared	128	8	0.224	0.048	0.356
41	cond_shared	128	8	0.284	0.081	0.356
42	cond_shared	128	16	0.101	0.058	0.356
43	cond_shared	128	16	0.267	0.072	0.356
44	cond_shared	128	16	0.351	0.057	0.356
45	cond_shared	128	32	0.321	0.031	0.356
46	cond_shared	128	32	0.337	0.040	0.356
47	cond_shared	128	32	0.402	0.067	0.356
48	uncond_shared_FAIL	64	4	0.382	0.303	0.407
49	uncond_shared_FAIL	64	4	0.462	0.254	0.407