LAMBDA: Diffusion-Step Alignment for Learning Robust Cross-Modal Representations in Time Series

Anonymous Authors[1]

Abstract

Paired structured time series from heterogeneous health sensors often observe the same evolving physiological or biomechanical state through different measurement channels. We study how this underlying structure can be used to improve bidirectional conditional diffusion models, $p_\theta(X \mid Y)$ and $p_\phi(Y \mid X)$, for such paired data. We introduce LAMBDA, a lightweight training objective that aligns local encoder neighborhoods at matched diffusion steps through a windowed sequence-contrastive loss and a covariance-matching loss. On paired locomotion signals and canonical dynamical systems, stepwise alignment improves cross-modal reconstruction fidelity, distributional similarity, and downstream representation quality over the same diffusion backbone trained without alignment. These results suggest that diffusion-step local alignment is a useful inductive bias for structured health time series with shared underlying dynamics.

1. Introduction

Structured physiological and biomechanical measurements are increasingly collected as synchronized multivariate time series: motion capture, force plates, wearable inertial sensors, and clinical devices all measure different views of an evolving state of human locomotion. Conditional diffusion models provide a flexible generative backbone for reconstructing or imputing one view from another (Sohl-Dickstein et al., 2015; Ho et al., 2020; Shen and Kwok, 2023; Yuan and Qiao, 2024). In many health settings, however, the relation is bidirectional: kinematics may predict kinetics (Dey et al., 2021), kinetics may predict kinematics, and each modality can become missing, noisy, or expensive to measure. Training two conditional denoisers independently ignores that both denoising processes should preserve the same local temporal phase and dynamics.

We ask whether paired conditional diffusion models can be improved by coupling their intermediate representations at matched diffusion steps. Our hypothesis is that if X and Y are synchronized observations of a shared latent process, then local windows of the denoising latents should be compatible up to a smooth view-specific transformation (Sauer et al., 1991). By leveraging the complementary information between paired modalities, we introduce LAMBDA (Local Latent Embedding Alignment), a method that trains paired conditional diffusion models for the two modalities, $p_\theta(X \mid Y)$ and $p_\phi(Y \mid X)$, and couples their denoising trajectories through stepwise local latent manifold alignment (Fig. 1). Specifically, the method aligns local subsequences of the modality-specific encoder features at the each diffusion step with (i) a first-order sequence-contrastive term that pulls together temporally matched windows and separates mismatched windows, and (ii) a second-order covariance term that matches local latent geometry. Our main contributions are: (1) a paired bidirectional diffusion framework for cross-modal structured time-series generation; (2) a diffusion-compatible local latent alignment objective (LAMBDA) with no inference-time overhead; and (3) an empirical evaluation on human locomotion and synthetic dynamical systems, including representation probes against standard alignment losses.

2. Related Work

Diffusion for sequential and cross-modal data. Denoising diffusion models have become a strong generative framework since the nonequilibrium formulation of Sohl-Dickstein et al. (2015) and the DDPM objective of Ho et al. (2020), with later work demonstrating competitive synthesis in high-dimensional domains (Dhariwal and Nichol, 2021). For sequences, diffusion has been adapted to time-series prediction and generation (Shen and Kwok, 2023; Yuan and Qiao, 2024) and to audio-like temporal signals (Kong et al., 2020). Most conditional diffusion studies, however, emphasize a single mapping direction. Our setting instead treats paired sensors as two conditionally generative views and asks how to couple the two denoising processes.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Representation and multiview alignment. Contrastive and redundancy-reduction objectives, including SimCLR, Barlow Twins, VICReg, and contrastive predictive coding, align views for representation learning (Chen et al., 2020; Zbontar et al., 2021; Bardes et al., 2021; Oord et al., 2018). Multiview methods such as deep CCA and temporal latent-variable models similarly exploit shared structure while allowing view-specific variation (Benton et al., 2019; Casale et al., 2018; Gondur et al., 2023). These approaches usually learn a static shared space or introduce an explicit multiview generative model. LAMBDA is complementary: it keeps modality-specific conditional diffusion samplers but regularizes their intermediate, time-local denoising neighborhoods at matched diffusion steps.

3. Method

3.1. Paired conditional diffusion

Let $\{(X_i, Y_i)\}_{i=1}^N$ denote paired trajectories, with $X_i \in \mathbb{R}^{L\times d_X}$ and $Y_i \in \mathbb{R}^{L\times d_Y}$. We train two conditional denoisers, $p_\theta(X \mid Y)$ and $p_\phi(Y \mid X)$, that reconstruct each modality at full temporal resolution from a noisy target and a clean conditioning signal. For a variance schedule $\{\beta_t\}_{t=1}^T$, define $\alpha_t = 1 - \beta_t$ and $\bar\alpha_t = \prod_{s=1}^t \alpha_s$. The closed-form forward noising process is

$$X_t = \sqrt{\bar\alpha_t} X_0 + \sqrt{1-\bar\alpha_t}\epsilon_X, \quad \epsilon_X \sim \mathcal{N}(0, I), \quad (1)$$

$$Y_t = \sqrt{\bar\alpha_t} Y_0 + \sqrt{1-\bar\alpha_t}\epsilon_Y, \quad \epsilon_Y \sim \mathcal{N}(0, I). \quad (2)$$

The denoisers predict $\widehat{X}_0 = p_\theta(X_t, g_Y(Y_0), t)$ and $\widehat{Y}_0 = p_\phi(Y_t, g_X(X_0), t)$, where g_X and g_Y are condition encoders. The ordinary bidirectional diffusion objective is

$$\mathcal{L}_{\text{den}} = \|X_0 - \widehat{X}_0\|_2^2 + \|Y_0 - \widehat{Y}_0\|_2^2. \quad (3)$$

3.2. Local neighborhood alignment

Each denoiser encoder produces step-indexed latents $Z_{X,t} = h_X(X_t, t)$ and $Z_{Y,t} = h_Y(Y_t, t)$ in $\mathbb{R}^{L\times d_Z}$. We divide these latents into P temporally matched windows, $Z_{X,t}^{(p)}$ and $Z_{Y,t}^{(p)}$, with window length S. The first-order term is a temporal contrastive loss adapted from contrastive predictive learning (Oord et al., 2018):

$$\mathcal{L}_{\text{con}} = -\frac{1}{P}\sum_{p=1}^{P} \log \frac{\exp(s(Z_{X,t}^{(p)}, Z_{Y,t}^{(p)})/\tau)}{\sum_{q=1}^{P}\exp(s(Z_{X,t}^{(p)}, Z_{Y,t}^{(q)})/\tau) + \mathcal{N}_B}, \quad (4)$$

where $s(\cdot,\cdot)$ is cosine similarity and $\mathcal{N}_B$ denotes additional mismatched windows from other sequences in the minibatch. This term rewards phase-consistent correspondence while discouraging trivial alignment to unrelated windows.

The second-order term matches the local covariance structure of corresponding windows:

$$\mathcal{L}_{\text{cov}} = \frac{1}{P}\sum_{p=1}^{P} \left\| \text{cov}\left(Z_{X,t}^{(p)}\right) - \text{cov}\left(Z_{Y,t}^{(p)}\right) \right\|_F^2. \quad (5)$$

The full training loss is

$$\mathcal{L} = \mathcal{L}_{\text{den}} + \lambda(\mathcal{L}_{\text{con}} + \mathcal{L}_{\text{cov}}). \quad (6)$$

where λ is designed as a learnable parameter. The alignment is performed at training time only. At test time, either model samples conditionally by the standard reverse diffusion procedure. This preserves the modularity of independent conditional generators while biasing their hidden denoising trajectories toward locally compatible phase structure.

Training details. For each minibatch we draw a diffusion step t, noise both modalities, run both denoisers, and compute $\mathcal{L}_{\text{den}}$ and $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{cov}}$ on the encoder outputs from that same step. Windows are sampled at matched temporal indices, so positives are synchronized subsequences from the same paired trajectory and negatives are temporally or subject-wise mismatched windows. The tradeoff λ controls the strength of cross-view coupling; setting $\lambda = 0$ recovers two independent conditional diffusion models. Because no projection head or matching network is retained at inference, LAMBDA changes the training objective but not the sampler, memory footprint, or conditional interface.

The alignment assumption is motivated by multiview dynamical systems. If two observation maps o_X and o_Y view the same state Z_k along a trajectory, then delay-coordinate arguments imply that local windows can be mutually reconstructive under generic conditions (Takens, 2006; Sauer et al., 1991). LAMBDA does not require an explicit shared latent-variable model; it only encourages the learned denoising features to respect this local correspondence.

4. Experiments

Biomechanical structured time series. We evaluate on an open-source locomotion dataset containing synchronized treadmill walking measurements over different speeds and inclines (Embry et al., 2018). The three modalities are joint angles, joint moments, and ground-reaction forces (GRF). Each experiment trains a pair of transformer-based DDPMs for one modality pair and evaluates both conditional directions: $X \mid Y$ and $Y \mid X$. This domain is a natural health testbed because the modalities are physically coupled, yet differ in measurement cost and availability.

Synthetic dynamical systems. To test the same idea under controlled dynamics, we also use the Lorenz attractor and the double pendulum in non-chaotic and mildly chaotic regimes. These systems vary in phase-space geometry and

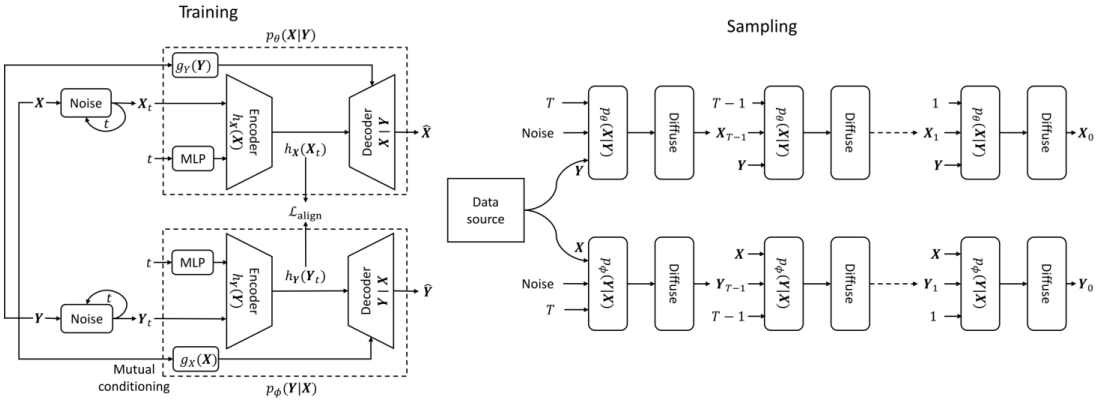

Figure 1. Training and sampling overview. Paired conditional denoisers generate X from Y and Y from X. During training, LAMBDA aligns local encoder neighborhoods, $h_X(X_t, t)$ and $h_Y(Y_t, t)$, at matched diffusion steps. During sampling, the learned denoisers are used independently and no additional alignment module is required.

sensitivity to initial conditions, allowing us to ask whether local alignment helps beyond a single empirical dataset.

Evaluation. We report mean-squared error (MSE) between generated and ground-truth paired trajectories; distributional similarity using a time-series adaptation of Frechet Inception Distance (FID) (Yu et al., 2021); and predictive score, where a sequence model trained on generated data is tested on real data (Yoon et al., 2019). We also evaluate representation quality by freezing diffusion encoder outputs and training linear or nonlinear classifiers to predict locomotion task labels. Lower values are better for MSE, FID, and predictive score; higher values are better for probing accuracy.

Protocol. Biomechanical sequences are segmented into fixed windows that span approximately two gait cycles. Train–test splits leave out participants and locomotion profiles rather than random individual windows, so the reported values reflect generalization to unseen users and speed–incline combinations. All comparisons use the same transformer-based DDPM backbone, optimizer settings, and conditional inputs; only the alignment objective is changed. This design isolates the effect of stepwise latent coupling from architecture or sampling differences.

5. Results

Cross-modal generation improves under local alignment. Table 1 and Figure. 2 summarizes the main biomechanical results. Adding LAMBDA improves distributional similarity in all six conditional directions and improves or matches pointwise MSE across all pairs. The largest gains occur for angles–GRF, the most heterogeneous pairing: FID improves from 66.7 to 40.4 for $X \mid Y$ and from 24.8 to 5.8 for $Y \mid X$, while the predictive score for $X \mid Y$ drops from 0.30 to 0.16. These changes indicate that the learned conditional trajectories are not merely closer pointwise, but also more

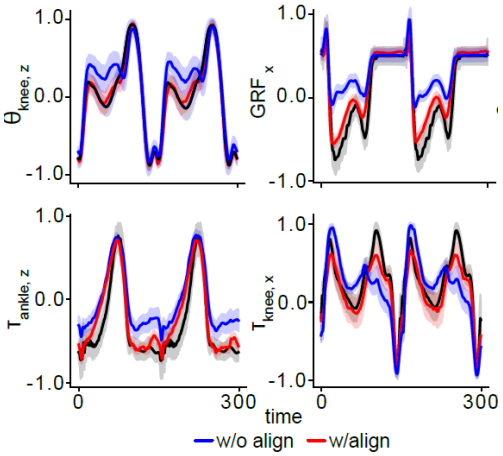

Figure 2. Comparison of real (black) and generated trajectories of joint angles, moments, and GRF using models trained with (red) and without (blue) latent alignment of diffusion models. The shaded region represents the standard deviation.

useful as samples from the real time-series distribution.

The effect is not limited to locomotion. Table 2 shows that the same stepwise alignment improves Lorenz and chaotic double-pendulum reconstruction, while the non-chaotic pendulum is already nearly solved by the baseline. The pattern is consistent with the motivation: local alignment is most useful when the inverse relation between views is nonlinear, phase-sensitive, or distributionally complex.

Alignment also improves representations. Against plug-in representation losses applied to the same diffusion backbone, LAMBDA gives the strongest average downstream probes. Averaged over the six biomechanical conditional directions, LAMBDA reaches 0.807 linear-probe accuracy and 0.827 nonlinear-probe accuracy, compared with 0.775 and 0.767 for the strongest alternatives among SimCLR (Chen et al., 2020), Barlow Twins (Zbontar et al., 2021), VICReg (Bardes

Table 1. Cross-modal generation on biomechanical time series. Each row compares the same conditional diffusion backbone trained without alignment and with LAMBDA. Values are mean ± standard deviation; lower is better.

Modality pair	Direction	MSE		FID		Pred.	
		w/o align	LAMBDA	w/o align	LAMBDA	w/o align	LAMBDA
Angles–Moments	$X \mid Y$	0.18 ± 0.03	**0.14 ± 0.02**	37.8 ± 9.6	**32.4 ± 7.2**	0.18 ± 0.06	**0.16 ± 0.03**
Angles–Moments	$Y \mid X$	0.08 ± 0.02	**0.07 ± 0.01**	20.4 ± 12.1	**14.2 ± 2.8**	0.08 ± 0.01	**0.07 ± 0.01**
Angles–GRF	$X \mid Y$	0.22 ± 0.03	**0.19 ± 0.03**	66.7 ± 51.5	**40.4 ± 8.3**	0.30 ± 0.23	**0.16 ± 0.02**
Angles–GRF	$Y \mid X$	0.07 ± 0.03	**0.06 ± 0.03**	24.8 ± 34.2	**5.8 ± 3.6**	0.12 ± 0.12	**0.08 ± 0.07**
Moments–GRF	$X \mid Y$	0.08 ± 0.02	**0.07 ± 0.02**	16.5 ± 4.0	**13.7 ± 3.2**	0.08 ± 0.01	**0.07 ± 0.01**
Moments–GRF	$Y \mid X$	0.03 ± 0.02	**0.03 ± 0.02**	6.6 ± 2.5	**4.3 ± 2.5**	0.07 ± 0.04	**0.05 ± 0.04**

Table 2. Cross-modal generation performance (MSE) on canonical dynamical systems.

System	Dir.	w/o align	LAMBDA
Lorenz	$X \mid Y$	0.678	**0.425**
Lorenz	$Y \mid X$	0.135	**0.004**
Pendulum, non-chaotic	$X \mid Y$	$2.5e-3$	**2.5e-3**
Pendulum, non-chaotic	$Y \mid X$	$6.6e-3$	**6.4e-3**
Pendulum, chaotic	$X \mid Y$	0.042	**0.028**
Pendulum, chaotic	$Y \mid X$	0.031	**0.021**

Table 3. Average locomotion-task probe accuracy over six conditional directions. Higher is better.

Alignment loss	Linear	Nonlinear
Barlow Twins	0.620	0.665
VICReg	0.617	0.663
Latent MSE	0.742	0.767
SimCLR	0.775	0.750
LAMBDA	**0.807**	**0.827**

et al., 2021), and latent MSE (Table 3). UMAP visualizations (McInnes et al., 2018) in the full manuscript further show that aligned models merge modality-specific latent spaces by locomotion task rather than by sensor view. These results support the interpretation that diffusion-step local alignment improves both generation and the semantic organization of hidden states.

Ablation. Table 4 isolates the two alignment components on the angles–moments pair. Removing the contrastive term weakens phase correspondence and returns the $X \mid Y$ direction to the unaligned error level. Removing covariance matching is less severe but still reduces the gain. The full objective is therefore not simply a generic representation penalty: it combines temporal correspondence with local geometric regularity.

6. Discussion

The proposed locally aligned bidirectional cross-modal diffusion framework offers a practical health motivation: dense force or moment measurements may be unavailable outside specialized laboratories, while cheaper kinematic or wearable signals are easier to obtain but less directly informative. A bidirectional model can be used either for sensor imputation or for generating physically plausible complementary channels, provided that the generated sequences remain phase-consistent with the observed signal. More broadly, the same formulation applies when two structured clinical streams are time-locked but differ in invasiveness, cost, or sampling reliability.

Table 4. Ablation on angles–moments generation measured by MSE. Lower is better.

Variant	$X \mid Y$	$Y \mid X$
w/o $\mathcal{L}_{\text{con}}$	0.18 ± 0.03	0.08 ± 0.03
w/o $\mathcal{L}_{\text{cov}}$	0.17 ± 0.03	0.07 ± 0.02
LAMBDA	$\mathbf{0.14 \pm 0.02}$	$\mathbf{0.07 \pm 0.01}$

7. Conclusion and Future Work

We studied paired conditional diffusion models for cross-modal time-series generation and asked whether coupling their denoising trajectories via a lightweight *diffusion-step* regularizer improves phase-consistent synthesis. We introduced LAMBDA, which aligns local latent neighborhoods at matched diffusion steps using a first-order windowed sequence-contrastive term and a second-order covariance-matching term without changing the sampling procedure. Across biomechanical modalities and canonical dynamical systems, this stepwise alignment consistently improves conditional generation quality relative to the same diffusion backbones trained without alignment and strengthens the learned representations for downstream probes.

The next step is to move from synchronized laboratory-style pairs toward the partial, irregular, and device-dependent measurements that arise in clinical and wearable settings. Therefore, future work includes evaluating robustness when the shared-process assumption is weakened (e.g., partial coupling or misalignment) and extending the approach beyond two modalities via centroid-based or coordinated pairwise alignment schemes.

References

Bardes, A., Ponce, J., and LeCun, Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv preprint arXiv:2105.04906*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

S. Dey, T. Yoshida, R. H. Foerster, M. Ernst, T. Schmalz, R. M. Carnier, and A. F. Schilling, "A hybrid approach for dynamically training a torque prediction model for devising a human-machine interface control strategy," *arXiv preprint* arXiv:2110.03085, 2021.

Embry, K. R., Villarreal, D. J., Macaluso, R. L., and Gregg, R. D. The effect of walking incline and speed on human leg kinematics, kinetics, and EMG, 2018. URL https://dx.doi.org/10.21227/gk32-e868.

Gondur, R., Sikandar, U. B., Schaffer, E., Aoi, M. C., and Keeley, S. L. Multi-modal Gaussian process variational autoencoders for neural and behavioral data. *arXiv preprint arXiv:2310.03111*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Oord, A. van den, Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Sauer, T., Yorke, J. A., and Casdagli, M. Embedology. *Journal of Statistical Physics*, 65:579–616, 1991.

Shen, L. and Kwok, J. Non-autoregressive conditional diffusion models for time series prediction. In *International Conference on Machine Learning*, pp. 31016–31029. PMLR, 2023.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, pp. 366–381. Springer, 2006.

Yoon, J., Jarrett, D., and Van der Schaar, M. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yu, Y., Zhang, W., and Deng, Y. Frechet inception distance (FID) for evaluating GANs. China University of Mining Technology Beijing Graduate School, 2021.

Yuan, X. and Qiao, Y. Diffusion-TS: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*, 2024.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow Twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. DiffWave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., and Arora, R. Deep generalized canonical correlation analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pp. 1–6, 2019.

Casale, F. P., Dalca, A., Saglietti, L., Listgarten, J., and Fusi, N. Gaussian process prior variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018.