OBSERVATION-ADAPTIVE SPATIOTEMPORAL POSITIONAL ENCODING: A STATE-SPACE PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning to perceive and predict the evolving physical world is central to physical intelligence, but developing intrinsic spatiotemporal perception remains a key challenge. Transformers represent position through positional encoding (PE). However, temporal coordinates in PE typically use raw time indices, which specify when observations are sampled but do not reflect how much the system state changes between them. To address this limitation, we propose **SpatiotemporalPE**, which combines spatial structure and temporal evolution. Using a state-space formulation, we relate changes in the underlying state to elapsed time. Specifically, we estimate local state variation from observation differences through averaging and noise correction to construct **Observation-difference Coordinate**. We instantiate SpatiotemporalPE for molecular dynamics simulation and video understanding. SpatiotemporalPE achieves the best MAE on 8 of 10 molecular dynamics tasks and reduces average MAE by 14.2% relative to the best baselines. SpatiotemporalPE improves overall accuracy from 43.85% to 44.63% in Video experiment. These results support observation-adaptive coordinates as a cross-domain inductive bias for spatiotemporal modeling.

1 INTRODUCTION

Physical processes unfold across spatial and temporal dimensions, making four-dimensional spatiotemporal perception essential for physical intelligence (Lake et al., 2017; LeCun, 2022; Kipf et al., 2018; Hafner et al., 2019). At the microscopic scale, molecular dynamics (MD) describes how molecular configurations evolve in space and time (Chodera & Noé, 2014; Mardt et al., 2018), with applications to biomolecular interactions and drug discovery (Siebenmorgen et al., 2024; Janson & Feig, 2025; Libouban et al., 2025; Noé et al., 2020). In video, image observations change over time, providing a setting for learning latent representations of spatiotemporal dynamics (Ha & Schmidhuber, 2018; Bardes et al., 2024; Assran et al., 2025; Finn et al., 2016). These settings highlight the need for an intrinsic spatiotemporal perception component that enables models to understand the physical world.

To model spatiotemporal sequences and dynamic processes, prior work has developed geometric and equivariant architectures for spatial structure (Bronstein et al., 2021; Satorras et al., 2021; Fuchs et al., 2020; Liao & Smidt, 2023; Batzner et al., 2022; Batatia et al., 2022), as well as neural differential equations and learned simulators for state evolution (Chen et al., 2018; Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021). Task-specific models further address molecular trajectories (Jing et al., 2024; Liu et al., 2026; Feng et al., 2026; Shi et al., 2026) and video dynamics (Bertasius et al., 2021; Arnab et al., 2021; Tong et al., 2022; Wang et al., 2023; Liu et al., 2022; Fan et al., 2021). For Transformers used in these settings, positional encoding (PE) is the core component that supplies spatial and temporal position information to attention (Vaswani et al., 2017; Shaw et al., 2018; Wei et al., 2025; Wu et al., 2021b).

Existing PE methods typically use raw time or frame indices as temporal coordinates. These indices specify when observations occur, but not how much the system changes between them: equally spaced MD snapshots may contain small vibrations or large conformational changes, while video frames may show slow or rapid motion (Lucas & Kanade, 1981; Teed & Deng, 2020). Thus, raw temporal coordinates alone cannot make PE reflect the evolution of spatial structure. Related work captures temporal patterns through autocorrelation or spectral attention (Wu et al., 2021a; Zhou et al.,

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

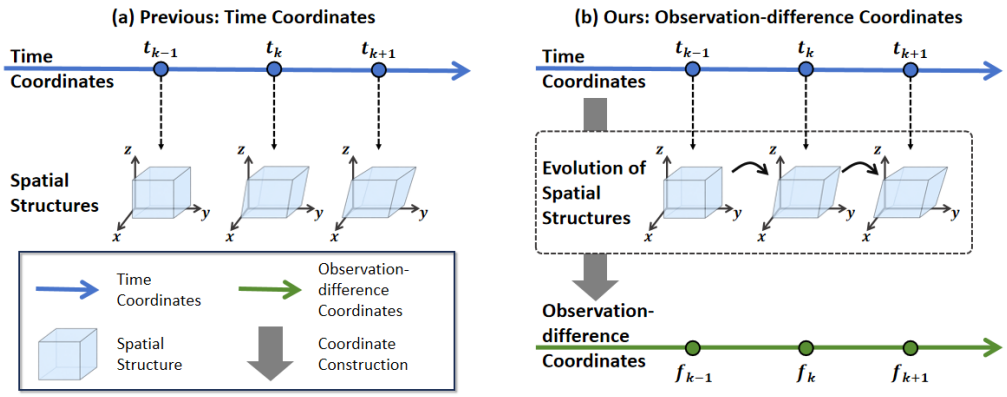


Figure 1: Comparison of temporal coordinate representations. (a) Each time point on the original time axis corresponds to a spatial structure. (b) The proposed Observation-difference Coordinate integrates spatial and temporal information by incorporating the evolution of spatial structures into the temporal coordinate.

2022), while signal-aware, contextual, and data-adaptive PE incorporate input content into positional representations (Irani & Metsis, 2025; Golovneva et al., 2024; Zheng et al., 2024). We therefore seek a spatiotemporal coordinate that combines temporal and spatial information, whose increments relate to the magnitude of system-state variation. However, the system state is generally latent, and observations contain noise, making direct estimation of state evolution challenging.

Our Contributions. We propose **SpatiotemporalPE**, a PE framework that integrates temporal and spatial information (Figure 1). We adopt a state-space perspective, where each observation O_k is viewed as a noisy measurement of an underlying state S_k (Kalman, 1960; Revach et al., 2022), and construct a temporal coordinate that reflects estimated state variation. Specifically, we construct the **Observation-difference Coordinate** $f_k = f(\mathcal{O}, t_k)$ from the observation sequence $\mathcal{O}$ and raw time t_k. We then combine f_k with the task-dependent spatial coordinate g_i to form the **Spatiotemporal Coordinate** $\tau_{i,k} = \tau(g_i, f_k)$, and apply PE on this coordinate.

The goal is therefore not to introduce another temporal embedding, but to construct a coordinate system in which positional distance reflects observation-dependent spatiotemporal state evolution while remaining compatible with existing Transformer architectures.

To verify the empirical effect, we instantiate SpatiotemporalPE for MD simulation and video understanding. In MD experiment, average MAE decreases by 14.2% relative to BioMD (Feng et al., 2026); controlled video evaluation improves overall accuracy from 43.85% to 44.63% (Fu et al., 2025). Ablations examine coordinate adaptation, velocity features, and noise correction.

2 Related Work

Rotary Position Embedding. Rotary Position Embedding (RoPE) encodes position by applying position-dependent rotations to query and key vectors (Su et al., 2024). Its key property is that the inner product between two rotated representations depends explicitly on their relative positional displacement (Liutkus et al., 2021), making RoPE particularly suitable for attention over structured sequences. Beyond fixed positional indices, recent work has explored learned, context-dependent, data-adaptive, and signal-aware positional representations (Kazemi et al., 2019; Golovneva et al., 2024; Zheng et al., 2024; Irani & Metsis, 2025; 2026).

Molecular Dynamics Modeling. Equivariant architectures such as Equiformer incorporate three-dimensional symmetry into atomistic representations (Liao & Smidt, 2023), while MDGen formulates molecular trajectory prediction as conditional generative modeling (Jing et al., 2024). For protein-ligand binding dynamics, NeuralMD introduces a physics-informed, multi-grained group-symmetric model together with second-order neural ODE and SDE solvers (Liu et al., 2026). More recently, BioMD adopts a hierarchical forecasting-and-interpolation framework to generate long-timescale

all-atom protein-ligand trajectories (Feng et al., 2026). These methods primarily advance the representation or prediction of three-dimensional molecular states, whereas SpatiotemporalPE complements them by introducing an observation-adaptive temporal coordinate for spatiotemporal attention.

Video Modeling. Video Transformers commonly organize visual tokens along spatial and temporal axes. TimeSformer factorizes space-time attention (Bertasius et al., 2021), while ViViT studies different spatiotemporal tokenization and attention structures for video representation learning (Arnab et al., 2021). VideoRoPE (Wei et al., 2025) extends RoPE to video with low-frequency temporal allocation, a spatially symmetric diagonal layout, and adjustable temporal spacing to encode spatiotemporal relationships.

3 Problem Formulation: A State-Space View of Spatiotemporal Modeling

3.1 Observation Model

We consider spatiotemporal sequence modeling from a state-space perspective. Let $\{t_k\}_{k=1}^K$ denote the raw time indices, where t_k may represent a simulation step, physical timestamp, or video frame index depending on the task. At raw time t_k, $S_k \in \mathbb{R}^{d_o}$ denotes the underlying system state, while $O_k \in \mathbb{R}^{d_o}$ denotes the task-dependent observation. Following the standard observation model in state-space modeling (Kalman, 1960; Särkkä, 2013; Fraccaro et al., 2017; Krishnan et al., 2017), we define the observation model by

$$O_k = S_k + \epsilon_k, \qquad \epsilon_k \sim \mathcal{N}(0, r_k I_{d_o}), \tag{1}$$

where ϵ_k is the observation noise, r_k is its local per-dimension variance, and I_{d_o} is the $d_o \times d_o$ identity matrix.

3.2 Spatiotemporal Task Formulation

Each observation O_k at raw time t_k is represented by one or more tokens. We associate each token $i \in \{1, \dots, N\}$ with a task-dependent spatial coordinate g_i, such as an atom coordinate or a video-patch grid position, and collect all spatial coordinates as $\mathcal{G} = \{g_i\}_{i=1}^N$.

Formally, we consider spatiotemporal predictors of the form

$$Y = F_\theta\left(\mathcal{O}, \mathcal{G}, \{t_k\}_{k=1}^K; \mathcal{C}\right), \tag{2}$$

where $\mathcal{C}$ denotes optional task-specific conditioning information and Y is the task target. We mainly consider two spatiotemporal tasks, as illustrated in **??**.

Molecular Dynamics Simulation. The observation sequence $\mathcal{O}$ contains historical ligand snapshots, and g_i is the three-dimensional coordinate of an atom-level token. The raw timestamps $\{t_k\}_{k=1}^K$ are the corresponding simulation steps, while $\mathcal{C}$ contains the fixed protein context. The task target Y is the future ligand trajectory.

Video Understanding. An input video is divided into spatiotemporal tokens. The observation sequence $\mathcal{O}$ contains the video-token features, and g_i is the spatial patch-grid coordinate of a token. The raw timestamps $\{t_k\}_{k=1}^K$ specify the corresponding frame indices. The task target Y is the answer to multiple choice questions.

Section 4 introduces the general framework of SpatiotemporalPE and the core component Observation-difference Coordinate. The detailed task formulations and task-specific instantiations of SpatiotemporalPE are presented in Section 5.

4 SpatiotemporalPE

Motivated by 4D-aware sequence modeling, we introduce SpatiotemporalPE, a positional encoding that jointly represents spatial structure and temporal evolution. SpatiotemporalPE operates on a Spatiotemporal Coordinate τ, which combines a task-dependent spatial coordinate with the

Observation-difference Coordinate instead of treating space and raw time independently. We introduce SpatiotemporalPE framework and Observation-difference Coordinate in Sections 4.1 and 4.2, respectively.

4.1 SPATIOTEMPORALPE FRAMEWORK

We define SpatiotemporalPE as a rotary positional encoding (RoPE) applied to a Spatiotemporal Coordinate τ. For a token i at raw time t_k, let $\tau_{i,k}$ denote its Spatiotemporal Coordinate. We apply RoPE to $\tau_{i,k}$ axis-wise:

$$R(\tau_{i,k}) = \operatorname{diag}\left(R(\tau_{i,k}^{(1)}), R(\tau_{i,k}^{(2)}), \ldots, R(\tau_{i,k}^{(M_{\text{pos}})})\right), \tag{3}$$

where M_{pos} is the number of positional axes and $R(\tau_{i,k}^{(m)})$ denotes the rotary transformation on the m-th positional axis. For a query token (i,k) and a key token (j,ℓ), the query and key vectors are rotated as

$$\widehat{q}_{i,k} = R(\tau_{i,k})q_{i,k}, \qquad \widehat{k}_{j,\ell} = R(\tau_{j,\ell})k_{j,\ell}. \tag{4}$$

The resulting attention score is

$$\widehat{q}_{i,k}^{\top}\widehat{k}_{j,\ell} = q_{i,k}^{\top}R(\tau_{i,k})^{\top}R(\tau_{j,\ell})k_{j,\ell}. \tag{5}$$

Since RoPE preserves the relative-position property on each axis, we have

$$R(\tau_{i,k})^{\top}R(\tau_{j,\ell}) = R(\tau_{j,\ell} - \tau_{i,k}), \tag{6}$$

and therefore the attention score depends on the relative displacement $\tau_{j,\ell} - \tau_{i,k}$ between the two tokens in the Spatiotemporal Coordinate space. Thus, SpatiotemporalPE changes what attention perceives as positional distance: instead of measuring token separation only by raw time or fixed spatial coordinates, it measures their relative displacement in a joint spatiotemporal positional space.

Spatiotemporal Coordinate τ. We now define the Spatiotemporal Coordinate τ used in SpatiotemporalPE. Consider a token i at raw time t_k. Let g_i denote its task-dependent spatial coordinate, and let $\mathcal{O} = \{O_k\}_{k=1}^{K}$ denote the observation sequence defined in Section 3.

We define the Spatiotemporal Coordinate of token (i,k) as

$$\tau_{i,k} = \tau\left(g_i, f(\mathcal{O}, t_k)\right) = \left(\tau_{i,k}^{(1)}, \tau_{i,k}^{(2)}, \ldots, \tau_{i,k}^{(M_{\text{pos}})}\right) \in \mathbb{R}^{M_{\text{pos}}}, \tag{7}$$

where $f(\mathcal{O}, t_k)$ is the Observation-difference Coordinate, which is related to the magnitude of state variation and shared by all tokens at the same raw time t_k. For compactness, we write $f_k = f(\mathcal{O}, t_k)$, such that the Spatiotemporal Coordinate is $\tau_{i,k} = \tau(g_i, f_k)$.

4.2 OBSERVATION-DIFFERENCE COORDINATE f

We define the Observation-difference Coordinate $f_k = f(\mathcal{O}, t_k)$ by combining the raw time coordinate t_k with state variation inferred from the observation sequence $\mathcal{O} = \{O_k\}_{k=1}^{K}$.

For the raw time interval $\Delta t_k = t_{k+1} - t_k$, we define the corresponding Observation-difference Coordinate increment Δf_k as

$$\Delta f_k = f_{k+1} - f_k = \alpha_k \Delta t_k, \quad f_1 = 0, \tag{8}$$

where α_k is the Temporal Scaling Coefficient. Thus, constructing the Observation-difference Coordinate reduces to relate Equation (8) with the observation model Equation (1) and estimating α_k from the observations.

State-Variation Assumption. The latent state S_k is introduced through the observation model $O_k = S_k + \epsilon_k$ in Equation (1), while the Observation-difference Coordinate increment Δf_k is defined through Equation (8). To connect these two definitions, we model the state variation $\Delta S_k = S_{k+1} - S_k$ using Δf_k:

$$\Delta S_k \sim \mathcal{N}\left(0, \lambda \Delta f_k I_{d_o}\right), \tag{9}$$

where λ is a global scale parameter. This core assumption gives $\mathbb{E}[\|\Delta S_k\|_2^2/d_o] = \lambda \Delta f_k$, such that a larger expected magnitude of state variation corresponds to a larger Observation-difference Coordinate increment, as illustrated in Figure 2. This assumption is used to define a temporal coordinate whose increments correspond to expected state variation, rather than to model the complete underlying state transition dynamics.

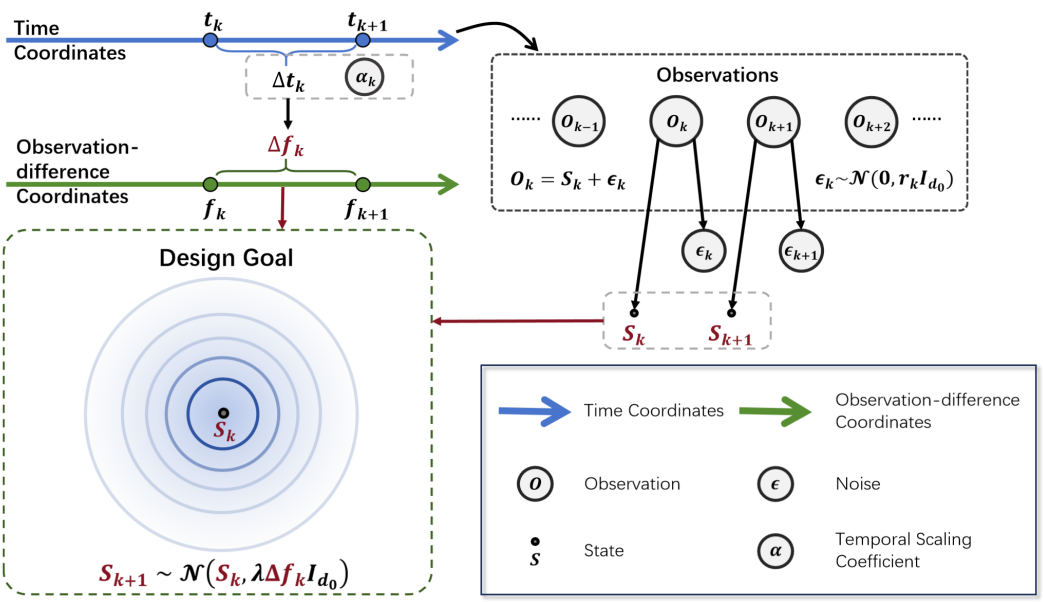


Figure 2: Construction of the Observation-difference Coordinate. The Temporal Scaling Coefficient α_k transforms the raw time interval Δt_k into the coordinate increment Δf_k. The core assumption links Δf_k to the latent state variation under the observation model, allowing α_k to be estimated from observation differences.

Temporal Scaling Coefficient α_k. We now derive the detail algorithm of α_k from the distribution in Equation (9). Taking the difference of the observation model gives $\Delta O_k = \Delta S_k + \epsilon_{k+1} - \epsilon_k$. Substituting $\Delta f_k = \alpha_k \Delta t_k$ from Equation (8) into the state-variation distribution in Equation (9), and assuming independent observation noise with $r_{k+1} \approx r_k$ within a short interval, gives

$$\Delta O_k \sim \mathcal{N}\left(0, \left(\lambda \alpha_k \Delta t_k + 2r_k\right) I_{d_o}\right). \tag{10}$$

Let $v_k = \frac{1}{\sqrt{d_o}} \|\Delta O_k\|_2$ denote the dimension-normalized magnitude of the observation difference. Maximum-likelihood estimation under Equation (10) gives the pointwise estimator

$$\alpha_k^{\mathrm{pt}} = \frac{v_k^2 - 2r_k}{\lambda \Delta t_k}. \tag{11}$$

The detailed derivation is provided in Section B.

To reduce sensitivity to local fluctuations, we pool v_m^2 over a local window W_k centered at t_k and enforce the non-negativity of the state-variation scale. The window-averaged estimator is

$$\alpha_k = \frac{\mathrm{ReLU}\left(\frac{1}{|W_k|} \sum_{m \in W_k} v_m^2 - 2r_k\right)}{\lambda \Delta t_k}. \tag{12}$$

The resulting $\{\alpha_k\}$ determine the increments $\{\Delta f_k\}$ through Equation (8), and hence determine the Observation-difference Coordinate f. The estimation of r_k and λ is described in Section C.

5 Instantiations of Spatiotemporal Coordinate τ

Section 4.1 defines the general interface $\tau_{i,k} = \tau(g_i, f_k)$, whose task-specific form must be specified. And Section 4.2 gives the detail of Observation-difference Coordinate f_k. This section instantiates τ and describes the corresponding computation pipelines for molecular dynamics simulation and video understanding.

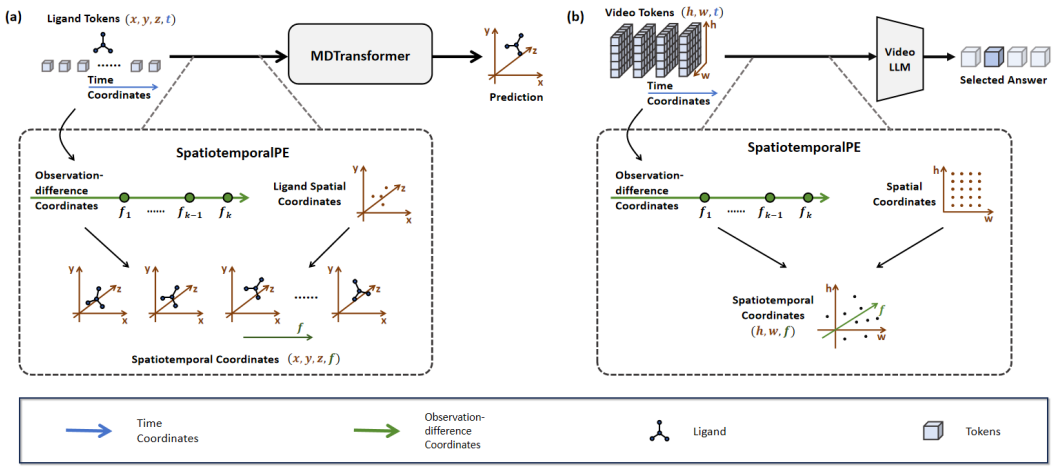


Figure 3: Instantiations and pipeline of molecular dynamics simulation and video understanding. (a) For molecular dynamics, ligand coordinates (x, y, z) and f form (x, y, z, f) for future trajectory prediction. (b) For video understanding, patch coordinates (h, w) and f form (h, w, f), which is mapped into VideoRoPE's diagonal layout before the video LLM selects an answer

5.1 Molecular Dynamics Simulation

As illustrated in Figure 3(a), the molecular dynamics instantiation replaces (x, y, z, t) with the Spatiotemporal Coordinate (x, y, z, f). Each ligand atom in an input snapshot is converted into an atom-level token, and the input trajectory is used as the observation sequence to estimate the Observation-difference Coordinate f_k. For token i, its spatial coordinate $g_i = (x_i, y_i, z_i)$ is combined with f_k to construct

$$\tau_{i,k}^{\mathrm{MD}} = \tau^{\mathrm{MD}}(g_i, f_k) = (x_i, y_i, z_i, f_k) \in \mathbb{R}^4. \tag{13}$$

The encoder outputs corresponding to the last input snapshot are passed to the decoder to predict the ligand coordinates of the next snapshot. The predicted snapshot is appended to the input trajectory, and the procedure is repeated to generate the future trajectory autoregressively.

5.2 Video Understanding

As illustrated in Figure 3(b), we instantiate SpatiotemporalPE for video understanding. For each sampled frame k, its visual token features form the observation O_k used to estimate the Observation-difference Coordinate f_k. We first combine f_k with the patch-grid coordinate $g_i = (h_i, w_i)$ of video token i to construct the spatiotemporal coordinate (h_i, w_i, f_k).

We then map this coordinate into VideoRoPE's diagonal positional layout. Replacing its raw frame index with f_k gives the position indices

$$\begin{aligned} b_k &= T_s + \delta(f_k - f_1), \\ \tau_{i,k}^{\mathrm{VU}} &= \left(b_k,\ b_k + w_i - \frac{W}{2},\ b_k + h_i - \frac{H}{2} \right), \end{aligned} \tag{14}$$

where T_s is the number of preceding text tokens, δ is VideoRoPE's temporal spacing factor. The resulting indices define the positional encoding for the video tokens in the video LLM, which selects an answer to each multiple-choice question.

6 Experiments

6.1 Experimental Setup

Molecular Dynamics setup. We follow the NeuralMD protocol on MISATO, a protein-ligand dynamics dataset in which each complex has a 100-snapshot trajectory. We adopt the semi-flexible

Table 1: Results on ten single-trajectory binding dynamics predictions in the semi-flexible setting. The **best** result in each column is shown in bold and the second-best is underlined.

Method	Metric	5WIJ	4ZX0	3EOV	4K6W	1KTI	1XP6	4YUR	4G3E	6B7F	3B9S
VerletMD	MAE	14.629	21.278	27.960	15.428	18.157	13.753	16.764	5.111	31.934	19.473
	Matching	5.459	7.971	13.588	7.505	7.467	4.672	9.555	3.388	21.691	0.923
	Stability	24.360	19.168	13.067	15.441	19.352	28.129	16.542	31.852	11.050	57.801
GNN-MD	MAE	2.280	2.370	3.512	3.695	6.641	2.378	7.031	2.709	4.136	2.578
	Matching	0.803	0.555	1.216	1.038	0.386	0.966	0.920	0.893	1.194	1.414
	Stability	54.475	68.613	40.984	42.480	81.831	49.239	47.555	61.802	39.067	49.306
DenoisingLD	MAE	2.501	3.138	4.055	3.942	7.051	2.218	7.128	3.588	4.431	2.811
	Matching	0.815	1.072	1.209	0.839	0.268	0.676	0.834	1.069	0.672	0.472
	Stability	52.418	44.228	41.469	53.820	91.986	64.951	49.676	40.823	61.583	71.852
NeuralMD-ODE	MAE	2.252	1.878	3.858	3.656	6.675	1.924	6.957	2.191	3.921	3.039
	Matching	0.464	0.428	1.062	0.928	0.337	0.537	0.584	0.505	0.459	0.659
	Stability	82.046	81.401	47.328	49.438	86.430	75.533	69.775	71.436	75.692	76.065
NeuralMD-SDE	MAE	2.260	2.158	3.395	3.765	6.646	2.061	7.038	2.345	3.842	3.132
	Matching	0.615	0.696	0.962	1.076	0.167	0.615	0.749	0.521	0.741	0.444
	Stability	67.464	59.109	50.108	49.700	98.508	69.423	60.344	68.729	57.917	77.801
BioMD	MAE	3.191	2.200	4.193	3.362	**4.251**	2.273	**5.163**	2.535	5.156	4.243
	Matching	**0.386**	**0.424**	**0.712**	0.778	0.153	0.544	0.477	0.228	**0.189**	0.095
	Stability	**87.15**	**83.08**	**64.43**	57.68	98.32	78.05	78.84	94.42	**96.53**	**100.00**
SpatiotemporalPE (Ours)	MAE	**2.035**	**1.642**	**3.316**	**2.670**	6.313	**1.564**	6.386	**1.648**	**3.700**	**2.073**
	Matching	0.393	0.438	0.946	**0.727**	**0.129**	**0.512**	**0.435**	**0.169**	0.205	**0.071**
	Stability	86.805	81.322	54.635	**60.448**	**99.383**	**79.612**	**79.210**	**97.833**	95.675	**100.00**

setting (rigid protein, flexible ligand) and consider two regimes: *single-trajectory prediction*, splitting each trajectory into the first 80 snapshots for training and the last 20 for testing; and *multi-trajectory prediction* on MISATO-100, MISATO-1000, and MISATO-All, where each data point is one full trajectory and the train/val/test sets contain disjoint complexes. We report MAE for reconstruction; **Matching**, the MSE of pairwise atomic distances, for geometric fidelity; and **Stability**, the fraction of atom pairs whose predicted distance is within $\Delta = 0.5$ Å of ground truth, for long-rollout physical plausibility. Full data construction, splits, and the three-seed configuration are detailed in Section D.

Video Understanding We evaluate SpatiotemporalPE on Video-MME under a controlled VideoWeave-style fine-tuning setting. VideoRoPE and SpatiotemporalPE use the same backbone, training data, optimizer, and LoRA configuration. Both models are fine-tuned for one epoch with $L = 2$, a learning rate of 1×10^{-5}, and rank-16 LoRA updates applied to the query and key projections. The only difference lies in the temporal coordinate used by RoPE: VideoRoPE adopts the raw frame index, whereas SpatiotemporalPE replaces it with the observation-adaptive coordinate f_k. This controlled comparison therefore directly evaluates the effect of temporal-coordinate calibration.

Compute resources. For the molecular dynamics experiments, training on MISATO-All used a single 80GB NVIDIA RTX PRO 6000 GPU and took approximately 12 hours per epoch. For video understanding, fine-tuning on 10,000 VideoWeave-processed samples, each containing 16 frames, used the same GPU and took approximately 12 hours for the single training epoch.

6.2 Molecular Dynamics Simulation

On the ten single-trajectory tasks (Table 1), SpatiotemporalPE achieves the lowest MAE on 8 complexes. It also obtains the best Matching and Stability on several complexes for which another method has a lower MAE, such as 1KTI and 4YUR. On the three multi-trajectory splits (Table 2), SpatiotemporalPE achieves the best reported MAE, RMSE, Matching, and Stability in each split. These results show gains in both trajectory reconstruction and the geometry of predicted molecular structures.

Table 2: Results on three multi-trajectory binding dynamics predictions under the **sliding-window** (in80/out20) protocol. The **best** result in each column is shown in bold, and the second-best is underlined.

Method	MISATO-100				MISATO-1000				MISATO-All			
	Reconstruction		Validity		Reconstruction		Validity		Reconstruction		Validity	
	MAE	RMSE	Matching	Stability	MAE	RMSE	Matching	Stability	MAE	RMSE	Matching	Stability
VerletMD	14.317	12.716	6.927	19.78	27.374	37.649	10.042	20.10	24.242	28.359	10.101	18.76
NeuralMD-ODE	4.210	5.509	0.407	84.76	4.548	4.810	0.446	84.69	4.448	4.658	0.428	84.90
NeuralMD-SDE	4.215	5.512	0.407	84.71	4.555	4.816	0.446	84.59	4.455	4.663	0.430	84.70
GNN-MD	4.198	5.524	0.489	77.67	4.500	4.691	0.553	75.06	4.434	4.622	0.509	77.59
DenoisingLD	4.184	5.499	0.396	85.39	4.467	4.676	0.430	85.42	4.400	4.607	0.412	85.64
BioMD	4.747	5.564	0.471	80.28	4.522	4.715	0.472	82.47	4.390	4.598	0.420	85.15
SpatiotemporalPE	**4.138**	**4.337**	**0.386**	**86.33**	**3.979**	**4.259**	**0.377**	**86.92**	**3.989**	**4.278**	**0.376**	**87.04**

6.3 Ablation Study

We compare SpatiotemporalPE with simple cumulative observation-difference coordinates on all ten single-trajectory complexes. Table 3 reports the results averaged over all ten complexes, while the complete per-complex results are provided in Section E.

Table 3: Comparison with simple cumulative observation-difference coordinates, averaged over all ten single-trajectory complexes. Lower MAE and Matching are better, while higher Stability is better. The best result for each metric is shown in bold.

Method	MAE ↓	Matching ↓	Stability ↑
SimpleV1	4.048	0.504	75.791
SimpleV2	4.043	0.494	77.245
Full SpatiotemporalPE	**3.138**	**0.403**	**83.492**

Simple coordinate baselines. We test whether directly accumulating observation differences is sufficient. SimpleV1 uses $\alpha_k = v_k/(\lambda\Delta t_k)$, while SimpleV2 uses $\alpha_k = v_k^2/(\lambda\Delta t_k)$. Both baselines construct f_k by cumulatively summing $\Delta f_k = \alpha_k \Delta t_k$ without the local-window averaging or noise correction used by SpatiotemporalPE. To isolate coordinate construction, these variants retain the remaining architecture and training configuration of the full model. Relative to the stronger simple baseline for each metric, full SpatiotemporalPE improves MAE by 22.4%, Matching by 18.5%, and Stability by 8.1%. It also achieves better Matching and Stability on all ten complexes and better MAE on nine of the ten complexes. Therefore, the improvement cannot be explained solely by cumulative motion magnitude; the local, noise-aware coordinate estimator provides an additional benefit.

6.4 Video Understanding

As reported in Table 4, SpatiotemporalPE improves the overall Video-MME accuracy from 43.85% to 44.63%, corresponding to a gain of 0.78 percentage points. Improvements are observed across all three video-length subsets: $+1.11$ points on short videos, $+0.11$ points on medium videos, and $+1.11$ points on long videos. The gains are therefore most pronounced on the short- and long-video subsets, while performance on medium-length videos remains nearly unchanged.

Table 4: Video-MME results under the controlled VideoWeave-style fine-tuning setting. Both models are fine-tuned with $L = 2$, learning rate 1×10^{-5}, one training epoch, and rank-16 LoRA updates on q, k. We report multiple-choice accuracy (%, ↑) overall and for short, medium, and long videos. The better result in each column is shown in bold.

Method	Overall	Short	Medium	Long
VideoRoPE	43.85	51.11	40.56	39.89
SpatiotemporalPE	**44.63**	**52.22**	**40.67**	**41.00**
Δ	+0.78	+1.11	+0.11	+1.11

Table 5: Video-MME results by task type. All scores are multiple-choice accuracy (%, ↑). Δ denotes the difference between SpatiotemporalPE and VideoRoPE in percentage points. The better result is shown in bold; ties are bolded for both methods.

Task Type	VideoRoPE	SpatiotemporalPE	Δ
Temporal Perception	**43.64**	40.00	-3.64
Spatial Perception	**55.56**	**55.56**	+0.00
Attribute Perception	46.40	**48.20**	+1.80
Action Recognition	**38.34**	37.70	-0.64
Object Recognition	39.83	**42.66**	+2.83
OCR Problems	39.57	**40.29**	+0.72
Counting Problem	29.85	**30.97**	+1.12
Temporal Reasoning	37.29	**38.98**	+1.69
Spatial Reasoning	**57.14**	**57.14**	+0.00
Action Reasoning	43.86	**44.21**	+0.35
Object Reasoning	42.73	**43.17**	+0.44
Information Synopsis	66.25	**66.56**	+0.31

The task-type breakdown in Table 5 provides a more detailed view of the improvement. Among the 12 task categories, SpatiotemporalPE outperforms VideoRoPE on eight, matches it on two, and underperforms it on two. The largest improvements occur in Object Recognition ($+2.83$), Attribute Perception ($+1.80$), Temporal Reasoning ($+1.69$), and Counting Problem ($+1.12$). Smaller gains are also obtained on OCR Problems, Action Reasoning, Object Reasoning, and Information Synopsis.

Overall, replacing raw frame indices with f_k produces a consistent aggregate improvement, although its effect varies across task types. The gain on Temporal Reasoning supports the utility of encoding observation-dependent state evolution. The decline on Temporal Perception may reflect a loss of raw temporal information; future work could explore a hybrid encoding that preserves both t and f.

7 Conclusion

We propose **SpatiotemporalPE**, which combines spatial coordinates with an observation-adaptive temporal coordinate to encode state evolution in spatiotemporal attention. On ten single-trajectory molecular dynamics tasks, SpatiotemporalPE achieves the best MAE on eight and reduces average MAE by 14.2% relative to BioMD. It also achieves the best reported results across all four metrics on three multi-trajectory splits. Replacing raw time with the Observation-difference Coordinate may weaken the model's awareness of elapsed time. In future work, we will investigate a hybrid encoding that combines f_k and t_k to retain both state-variation and elapsed-time information.

AI use statement

AI tools were used to assist with translation and language editing of the manuscript and to provide feedback on the presentation of the method and experiments. The authors reviewed the AI-assisted text and take responsibility for the final content and claims of this paper.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint*

arXiv:2506.09985, 2025. doi: 10.48550/arXiv.2506.09985. URL https://arxiv.org/abs/2506.09985.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Ilyes Batatia, Dávid P. Kovács, Gregor Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/4a36c3c51af11ed9f34615b81edb5bbc-Abstract-Conference.html.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13:2453, 2022. doi: 10.1038/s41467-022-29939-5. URL https://doi.org/10.1038/s41467-022-29939-5.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824. PMLR, 2021. URL https://proceedings.mlr.press/v139/bertasius21a.html.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. doi: 10.48550/arXiv.2104.13478. URL https://arxiv.org/abs/2104.13478.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6571–6583, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

John D. Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology*, 25:135–144, 2014. doi: 10.1016/j.sbi.2014.04.002. URL https://doi.org/10.1016/j.sbi.2014.04.002.

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. doi: 10.1093/biomet/81.3.425.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Fan_Multiscale_Vision_Transformers_ICCV_2021_paper.html.

Bin Feng, Jiying Zhang, Xinni Zhang, Zijing Liu, and Yu Li. BioMD: All-atom generative model for biomolecular dynamics simulation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LQDeJk6NOr.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/d9d4f495e875a2e075a1a4a6e1b9770f-Abstract.html.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7b7a53e239400a13bd6be6c91c4f6c4e-Abstract.html.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The

first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/15231a7ce4ba789d13b722cc5c955834-Abstract.html.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*, 2024. doi: 10.48550/arXiv.2405.18719. URL https://arxiv.org/abs/2405.18719.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. doi: 10.48550/arXiv.1803.10122. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Habib Irani and Vangelis Metsis. Dywpe: Signal-aware dynamic wavelet positional encoding for time series transformers. *arXiv preprint arXiv:2509.14640*, 2025. URL https://arxiv.org/abs/2509.14640.

Habib Irani and Vangelis Metsis. Positional encoding in transformer-based time series models: A survey. *International Journal of Machine Learning and Cybernetics*, 17:339, 2026. doi: 10.1007/s13042-026-03151-z. URL https://doi.org/10.1007/s13042-026-03151-z.

Giacomo Janson and Michael Feig. Generation of protein dynamics by machine learning. *Current Opinion in Structural Biology*, 93:103115, 2025. doi: 10.1016/j.sbi.2025.103115. URL https://doi.org/10.1016/j.sbi.2025.103115.

Bowen Jing, Hannes Stärk, Tommi Jaakkola, and Bonnie Berger. Generative modeling of molecular dynamics trajectories. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1282. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/478b06f60662d3cdc1d4f15d4587173a-Abstract-Conference.html.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. doi: 10.1115/1.3662552. URL https://doi.org/10.1115/1.3662552.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2688–2697. PMLR, 2018. URL https://proceedings.mlr.press/v80/kipf18a.html.

Rahul Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: 10.1609/aaai.v31i1.10779. URL https://doi.org/10.1609/aaai.v31i1.10779.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.

Yann LeCun. A path towards autonomous machine intelligence, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Version 0.9.2.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KwmPfARgOTD.

Pierre-Yves Libouban, Camille Parisel, Maxime Song, Samia Aci-Sèche, Jose Carlos Gómez-Tamayo, Gary Tresadern, and Pascal Bonnet. Spatio-temporal learning from molecular dynamics simulations for protein–ligand binding affinity prediction. *Bioinformatics*, 41(8):btaf429, 2025. doi: 10.1093/bioinformatics/btaf429. URL https://doi.org/10.1093/bioinformatics/btaf429.

Shengchao Liu, Weitao Du, Hannan Xu, Yanjing Li, Zhuoxinran Li, Vignesh Bhethanabotla, Divin Yan, Christian Borgs, Anima Anandkumar, Hongyu Guo, and Jennifer Chayes. A multi-grained symmetric differential equation model for learning protein-ligand binding dynamics. *Nature Communications*, 17:1049, 2026. doi: 10.1038/s41467-025-67808-z. URL https://www.nature.com/articles/s41467-025-67808-z.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Video_Swin_Transformer_CVPR_2022_paper.html.

Antoine Liutkus, Ondřej Cífka, Shih-Lun Wu, Umut Simsekli, Yi-Hsuan Yang, and Gaël Richard. Relative positional encoding for transformers with linear complexity. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7067–7079. PMLR, 2021. URL https://proceedings.mlr.press/v139/liutkus21a.html.

Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981. URL https://publications.ri.cmu.edu/an-iterative-image-registration-technique-with-an-application-to-stereo-vision-ijc

Stéphane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. doi: 10.1109/34.192463.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9:5, 2018. doi: 10.1038/s41467-017-02388-1. URL https://doi.org/10.1038/s41467-017-02388-1.

Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual Review of Physical Chemistry*, 71:361–390, 2020. doi: 10.1146/annurev-physchem-042018-052331. URL https://doi.org/10.1146/annurev-physchem-042018-052331.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=roNqYL0_XP.

Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adrià López Escoriza, Ruud J. G. van Sloun, and Yonina C. Eldar. KalmanNet: Neural network aided kalman filtering for partially known dynamics. *IEEE Transactions on Signal Processing*, 70:1532–1547, 2022. doi: 10.1109/TSP.2022.3158588. URL https://doi.org/10.1109/TSP.2022.3158588.

Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993. doi: 10.1080/01621459.1993.10476408. URL https://doi.org/10.1080/01621459.1993.10476408.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468. PMLR, 2020. URL https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Simo Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013. URL https://www.cambridge.org/core/books/bayesian-filtering-and-smoothing/C372FB31C5D9A100F8476C1B23721A67.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of NAACL-HLT 2018, Volume 2: Short Papers*, pp. 464–468, 2018. doi: 10.18653/v1/N18-2074.

Liang Shi, Jiarui Lu, Junqi Liu, Chence Shi, Zhi Yang, and Jian Tang. ATMOS: Atomic trajectory modeling with state space models for biomolecular dynamics. *arXiv preprint arXiv:2603.17633*, 2026. URL https://arxiv.org/abs/2603.17633.

Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, Erinc Merdivan, Kieran Didi, André S. Dias Mourão, Radosław Kitel, Pietro Liò, Stefan Kesselheim, Marie Piraud, Fabian J. Theis, Michael Sattler, and Grzegorz M. Popowicz. MISATO: Machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nature Computational Science*, 4(5):367–378, 2024. doi: 10.1038/s43588-024-00627-2. URL https://www.nature.com/articles/s43588-024-00627-2.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pp. 402–419, 2020. doi: 10.1007/978-3-030-58536-5_24. URL https://doi.org/10.1007/978-3-030-58536-5_24.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022. doi: 10.52202/068431-0732. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14560, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_VideoMAE_V2_Scaling_Video_Masked_Autoencoders_With_Dual_Masking_CVPR_2023_paper.html.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. VideoRoPE: What makes for good video rotary position embedding? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66118–66136. PMLR, 2025. URL https://proceedings.mlr.press/v267/wei25h.html.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10033–10041, 2021b. URL https://openaccess.thecvf.com/content/ICCV2021/html/Wu_Rethinking_and_Improving_Relative_Position_Encoding_for_Vision_Transformer_ICCV_2021_paper.html.

Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, and Yu Li. DAPE: Data-adaptive positional encoding for length extrapolation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0838. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f050fa9f0d898e3f265d515f50ae8f9-Abstract-Conference.html.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

A NOTATION

Notation scope. Table 6 summarizes the notation used throughout the paper. Symbols introduced only for a task-specific instantiation or a local derivation are defined where they first appear and are not included in the table.

Table 6: Summary of the main notation used in this paper.

Symbol	Definition
Indices and Dimensions	
i, j	Token indices.
k, ℓ	Raw time indices.
N	Number of tokens.
K	Number of observations in the input sequence.
d_o	Dimensionality of an observation or latent state.
M_{pos}	Number of positional axes in the Spatiotemporal Coordinate τ.
State-space Formulation	
$O_k \in \mathbb{R}^{d_o}$	Observation.
$\mathcal{O} = \{O_k\}_{k=1}^{K}$	Task-dependent observation sequence.
$S_k \in \mathbb{R}^{d_o}$	Latent state.
$\Delta O_k = O_{k+1} - O_k$	Observable difference.
$\Delta S_k = S_{k+1} - S_k$	Latent state variation.
ϵ_k	Observation noise.
r_k	Local per-dimension variance of the observation noise.
I_{d_o}	Identity matrix of size $d_o \times d_o$.
Coordinates	
t_k	Raw temporal coordinate.
$\Delta t_k = t_{k+1} - t_k$	Raw temporal increment.
g_i	Task-dependent spatial coordinate of token i.
$f_k = f(\mathcal{O}, t_k)$	Observation-difference Coordinate at raw time t_k.
$\Delta f_k = f_{k+1} - f_k$	Increment of the Observation-difference Coordinate.
α_k	Temporal Scaling Coefficient.
$\tau_{i,k} = \tau(g_i, f_k)$	Spatiotemporal Coordinate of token i at raw time t_k.
Coordinate Estimation Parameters	
v_k	Dimension-normalized magnitude of the observation difference.
$\mathcal{W}_k$	Local temporal window.
λ	Global state-variation scale.

B DERIVATION OF THE POINTWISE MAXIMUM LIKELIHOOD ESTIMATOR α_k

Here we provide a detailed derivation of the pointwise maximum-likelihood estimator introduced in Section 4.2

We can get the distribution of v_k from Equation (10). Since

$$\Delta O_k / \sigma_k \sim \mathcal{N}(0, I_{d_o}), \tag{15}$$

the squared norm satisfies

$$\frac{\|\Delta O_k\|_2^2}{\sigma_k^2} = \frac{d_o v_k^2}{\sigma_k^2} \sim \chi^2_{d_o}. \tag{16}$$

Equivalently, v_k^2 follows a Gamma distribution,

$$v_k^2 \sim \Gamma\left(\frac{d_o}{2}, \frac{2\sigma_k^2}{d_o}\right). \tag{17}$$

Therefore, the negative log-likelihood of a single observation v_k^2 is

$$\ell_k(\alpha_k) = \frac{d_o}{2}\log\sigma_k^2 + \frac{d_o v_k^2}{2\sigma_k^2}, \qquad \sigma_k^2 = \lambda\alpha_k\Delta t_k + 2r_k. \tag{18}$$

Differentiating with respect to α_k gives

$$\frac{\partial\ell_k(\alpha_k)}{\partial\alpha_k} = \frac{d_o\lambda\Delta t_k}{2}\left(\frac{1}{\sigma_k^2} - \frac{v_k^2}{(\sigma_k^2)^2}\right). \tag{19}$$

Setting the derivative to zero yields

$$\sigma_k^2 = v_k^2. \tag{20}$$

Substituting $\sigma_k^2 = \lambda\alpha_k\Delta t_k + 2r_k$, we obtain the single-point estimator

$$\alpha_k^{pt} = \frac{v_k^2 - 2r_k}{\lambda\Delta t_k}. \tag{21}$$

C Parameter Estimation of α

Here, we provide the detailed parameter estimation methods for Equation (12), which requires three quantities to be specified from data: the task-dependent observation O_k, the observation-noise variance r_k, and the global state-variation scale parameter λ.

Instantiation of the observation O_k. The scalar variation statistic v_k is defined as $v_k = \frac{1}{\sqrt{d_o}}\|\Delta O_k\|_2$. Therefore, the key task-specific choice here is the instantiation of the observation O_k.

In molecular dynamics (MD), the observation at time step k can be instantiated in two ways. The first choice is to use the atomic coordinate vector directly. The second choice is to use a learned state representation as the observation, i.e.,

$$O_k = h_k^{\mathrm{MD}} \in \mathbb{R}^{d_o}, \tag{22}$$

where h_k^{MD} denotes the atom state embedding at time t_k. In this case, v_k measures the effective temporal variation in the learned state-representation space rather than in the raw coordinate space.

In video, the observation is instantiated from visual patch features. Let $F_k^{\mathrm{Vid}} \in \mathbb{R}^{N_p\times d_p}$ denotes the patch-feature tensor extracted from the frame at time t_k, where N_p is the number of visual patches and d_p is the feature dimension of each patch. We define

$$O_k = \mathrm{vec}\big(F_k^{\mathrm{Vid}}\big) \in \mathbb{R}^{d_o}, \qquad d_o = N_p d_p, \tag{23}$$

where $\mathrm{vec}(\cdot)$ denotes vectorization. In the video setting, the temporal coordinate is driven by the evolution of visual token features rather than by raw frame indices alone.

Noise estimation r_k. In the observation model Equation (1), the observation noise is modeled with covariance $r_k I_{d_o}$. Accordingly, r_k should be estimated as a scalar per-dimension noise variance from the vector-valued observation sequence. To this end, we perform a wavelet decomposition along the temporal dimension, following the multiresolution wavelet framework (Mallat, 1989). For each observation dimension $a \in \{1,\dots,d_o\}$, let

$$\{O_k^{(a)}\}_{k=1}^K \longrightarrow \{c_{A,J}^{(a)}, c_{D,J}^{(a)}, c_{D,J-1}^{(a)}, \dots, c_{D,1}^{(a)}\} \tag{24}$$

denote the wavelet decomposition of the a-th scalar component of the observation sequence, where $c_{A,J}^{(a)}$ is the approximation coefficient at level J and $c_{D,j}^{(a)}$ is the detail coefficient at level j.

For each time window centered at t_k, we collect the corresponding first-level detail coefficients into

$$D_k^{(a)} = \{d_m^{(a)} \in c_{D,1}^{(a)} \mid m \in W_k\}. \tag{25}$$

We then estimate the local per-dimension noise variance by the squared median absolute deviation (MAD),

$$r_k^{(a)} = \mathrm{MAD}\Big(D_k^{(a)}\Big)^2, \tag{26}$$

where

$$\mathrm{MAD}\Big(D_k^{(a)}\Big) = \mathrm{median}_{d\in D_k^{(a)}}\left|d - \mathrm{median}_{d'\in D_k^{(a)}}(d')\right|. \tag{27}$$

This MAD-based local scale estimate is inspired by robust wavelet-shrinkage noise estimation (Donoho & Johnstone, 1994; Rousseeuw & Croux, 1993). Finally, we aggregate the component-wise estimates into a single scalar noise variance:

$$r_k = \frac{1}{d_o}\sum_{a=1}^{d_o} r_k^{(a)}. \tag{28}$$

This yields a robust local estimate of the average per-dimension observation-noise variance.

Global calibration of the state-variation scale λ. The parameter λ controls the global second-order scale of latent state variation in the state-variation assumption Equation (9). Following Equation (12), we first define the variation magnitude after noise removal as

$$\gamma_k = \frac{1}{|W_k|}\sum_{m\in W_k} v_m^2 - 2r_k. \tag{29}$$

Since λ is assumed to be globally constant, it should reflect the typical scale of $\gamma_k/\Delta t_k$ over the whole observation sequence. We therefore estimate it by

$$\lambda = \mathrm{median}_k\left(\frac{\gamma_k}{\Delta t_k}\right). \tag{30}$$

With this choice, λ provides a robust global calibration of the state-variation magnitude. When the variation magnitude γ_k at time step k is close to the typical variation magnitude of the whole sequence, the resulting α_k is close to 1, meaning that the Observation-difference Coordinate f evolves at approximately the same rate as the physical timeline t.

D Molecular Dynamics Experimental Details

MISATO is built from experimentally resolved protein–ligand complexes in the Protein Data Bank (PDB), with structural refinement based on semi-empirical quantum mechanics; each complex is associated with a 100-snapshot trajectory spanning 8 nanoseconds under fixed temperature and pressure.

For *single-trajectory prediction*, each 100-snapshot trajectory is split temporally into the first 80 snapshots for training and the last 20 for testing, following NeuralMD. For *multi-trajectory prediction*, MISATO-100 and MISATO-1000 are obtained by randomly sampling 100 and 1,000 complexes and splitting them 80%/10%/10% into train/validation/test. MISATO-All uses the official split shipped with the dataset; after removing complexes with peptide ligands, the resulting split contains 13,066 training, 1,357 validation, and 1,357 test complexes. All experiments use three random seeds (0, 42, 123).

E Additional Molecular Dynamics Ablations

E.1 Full Per-Complex Ablation Results

Simple coordinate baselines. SimpleV1 and SimpleV2 retain the same model architecture as full SpatiotemporalPE but replace its local, noise-aware estimator with direct cumulative observation-difference coordinates. Their complete per-complex results are reported in Table 7.

E.2 Noise-estimation Ablation

To isolate the contribution of noise correction, we compare full SpatiotemporalPE with a controlled variant that retains the same local-window averaging but sets $r_k = 0$. As shown in Table 8, noise estimation reduces MAE by approximately 0.9% and Matching by 9.9%, while improving Stability by 5.7%. This result indicates that noise correction contributes primarily to structural rollout quality

Table 7: Complete comparison with the simple cumulative coordinate baselines. Each entry reports MAE / Matching / Stability. The best result for each metric on each complex is shown in bold.

Complex	SimpleV1	SimpleV2	Full SpatiotemporalPE
5WIJ	3.958 / 0.499 / 77.825	4.045 / 0.477 / 77.836	**2.035 / 0.393 / 86.805**
4ZX0	1.931 / 0.470 / 78.335	2.209 / 0.446 / 80.635	**1.642 / 0.438 / 81.322**
3EOV	3.928 / 0.998 / 52.163	3.615 / 0.964 / 52.882	**3.316 / 0.946 / 54.635**
4K6W	3.309 / 0.856 / 52.991	3.196 / 0.891 / 50.989	**2.702 / 0.727 / 60.448**
1KTI	6.444 / 0.145 / 98.981	6.518 / 0.159 / 98.765	**6.313 / 0.129 / 99.383**
1XP6	2.541 / 0.513 / 76.003	2.090 / 0.527 / 75.614	**1.564 / 0.512 / 79.612**
4YUR	7.117 / 0.484 / 76.138	7.252 / 0.551 / 73.456	**6.386 / 0.435 / 79.210**
4G3E	**1.569** / 0.280 / 90.261	1.668 / 0.220 / 94.979	1.648 / **0.169 / 97.833**
6B7F	3.868 / 0.574 / 60.425	4.025 / 0.480 / 73.200	**3.700 / 0.205 / 95.675**
3B9S	5.818 / 0.225 / 94.792	5.814 / 0.222 / 94.097	**2.073 / 0.071 / 100.000**
Average	4.0483 / 0.5044 / 75.7914	4.0432 / 0.4937 / 77.2453	**3.138 / 0.403 / 83.492**

Table 8: Ablation of noise estimation, averaged over all ten molecular complexes. Both variants use the same local-window averaging and shared experimental configuration.

Method	MAE ↓	Matching ↓	Stability ↑
SpatiotemporalPE without noise estimation	3.928	0.534	74.088
SpatiotemporalPE with noise estimation	**3.893**	**0.481**	**78.337**

in the evaluated setting. The variant without noise estimation remains a lighter-weight alternative because it avoids the wavelet-based noise-estimation step.

The noise-estimation ablation uses its own shared experimental configuration; therefore, comparisons should be made between the two rows in Table 8, rather than between its absolute values and those of the other ablation tables.