VIDEOMD: GENERATING PROTEIN DYNAMICS WITH CONTINUOUS MOLECULAR VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding continuous protein dynamics is fundamental to elucidating biological function. Molecular dynamics (MD) simulations capture such dynamics at atomistic resolution, but long-timescale simulations remain computationally expensive. Deep generative models offer a promising alternative, yet most rely on discrete structural abstractions, where continuous motion is mediated by predefined nodes, tokens, or geometric features. This can create bottlenecks for smooth conformational modeling and increase computational overhead for large proteins. Here, we propose VideoMD, a conditional molecular video generation framework that represents protein dynamics as visual molecular videos and models them directly in continuous visual space. Unlike discrete structural abstractions, molecular videos provide a natural dynamic representation: molecular shape, spatial distance, and relative motion are captured as observable spatiotemporal cues along the evolving trajectory. To evaluate generated dynamics in this space, we introduce Molecular Video Evaluation (MoViE), a unified structure-aware protocol assessing structural validity, ensemble statistics, and temporal dynamics. Under MoViE, extensive experiments show that VideoMD achieves the best performance among the evaluated generative baselines on several ensemble and temporal metrics, including the strongest pairwise-RMSD correlation and a 73.82% relative improvement in latent autocorrelation MAE, while remaining competitive on selected structural-validity metrics. Using a structure-conditioned decoder, generated videos can be converted into coordinate trajectories that retain informative conformational and dynamical statistics. Further analysis shows nearly constant video-generation time across the tested protein sizes at fixed resolution and clip length, highlighting molecular videos as a scalable representation for efficient protein dynamics modeling.

1 INTRODUCTION

Proteins are not static objects. Many biological processes, including ligand binding, allosteric regulation, conformational selection, and enzymatic catalysis, are governed by continuous transitions across metastable conformational states (Berendsen & Hayward, 2000; Hollingsworth & Dror, 2018; Henzler-Wildman & Kern, 2007; Meller et al., 2023). Molecular dynamics (MD) simulation provides an atomistic view of these motions, but observing functionally relevant transitions often requires long-timescale simulations, which remain computationally expensive despite advances in hardware and simulation algorithms (Dror et al., 2012; Hollingsworth & Dror, 2018; Noé et al., 2020). As a result, there is growing interest in learned generative models that can emulate protein conformational dynamics more efficiently (Noé et al., 2019; Schreiner et al., 2023; Klein et al., 2023; Costa et al., 2024; Jing et al., 2024b; Shen et al., 2025; Xu et al., 2025b; Shi et al., 2026; Yu et al., 2026).

Most existing learning-based approaches generate protein dynamics through *discrete structural representations* (Figure 1a). In these formulations, continuous molecular motion is parameterized by predefined atom-, residue-, graph- Schreiner et al. (2023); Costa et al. (2024); Yu et al. (2026), or token-indexed variables Jing et al. (2024b); Cheng et al. (2025); Jing et al. (2024a), and dynamics are modeled by evolving these structural units over time with geometric or equivariant architectures (Figure 1b). These methods are physically expressive and effective for enforcing local structural constraints, but their representation of motion is mediated by a discrete molecular support: atoms,

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.

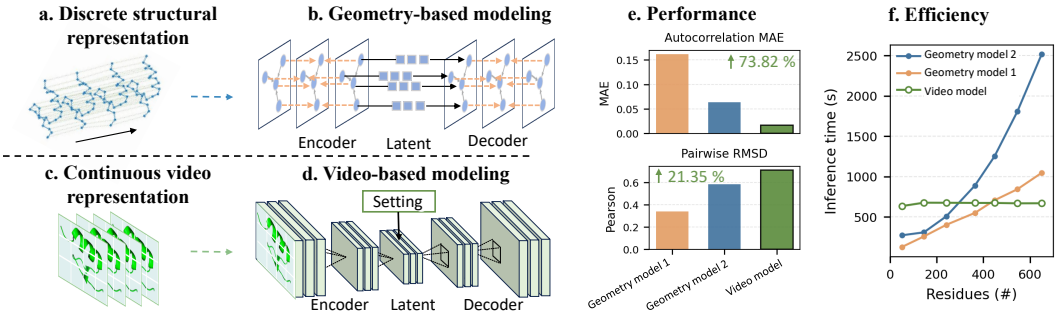


Figure 1: **(a)** Protein trajectories with discrete structural representations. **(b)** Geometry-based trajectory generation pipeline. **(c)** Our continuous molecular video representation of protein trajectories. **(d)** Molecular video generation pipeline. **(e)** Performance in MAE for autocorrelation and pairwise RMSD Pearson correlation across different methods. **(f)** Efficiency analysis across different protein lengths.

residues, edges, or tokens must first be chosen, and temporal evolution is then learned through updates on these predefined units.

This discrete-support representation introduces two practical challenges for long-horizon trajectory generation: **(i)** ***motion discretization bottleneck***. Graph- and geometry-based models must first choose the atoms, residues, edges, or tokens that carry motion, thereby expressing continuous conformational change as updates over selected structural units. This makes relational modeling tractable, but fine-grained cues, including gradual local displacement, smooth deformation, and motion propagation across structural regions, become less explicit and harder to preserve over long horizons. **(ii)** ***discrete-unit scaling***. Since message passing, attention, and pairwise geometric reasoning are performed over discrete molecular units, training and inference costs typically grow with the number of atoms, residues, or interactions, limiting scalability to larger proteins and biomolecular systems. These challenges motivate a complementary continuous-field view, where molecular motion is learned from dense spatiotemporal observations rather than from updates over predefined structural units.

To address these limitations, we propose VideoMD, a protein video representation and generation framework for molecular dynamics trajectories. VideoMD maps each trajectory to a fixed-resolution, fixed-length video tensor $V \in \mathbb{R}^{T\times C\times H\times W}$, forming a dense visual field for learning conformational motion (Figure 1c). Through this video representation, VideoMD captures molecular shape, domain motion, secondary-structure rearrangement, loop flexibility, and relative movement as spatiotemporal cues across the full rendered protein. It uses a conditional video generation model to synthesize full trajectories from the initial-conformation image and simulation configuration (Figure 1d), achieving improved temporal fidelity and cross-system scale preservation over geometry-based baselines (Figure 1e). Since generation operates on fixed-resolution videos, inference cost depends mainly on video resolution and clip length rather than protein size, yielding nearly constant cost across proteins of different lengths (Figure 1f). To evaluate generated dynamics in this continuous video representation, we further introduce Molecular Video Evaluation (MoViE), a structure-aware protocol that measures structural validity, ensemble statistics, and temporal dynamics within a unified framework. Generated molecular videos can also be decoded into coordinate trajectories conditioned on the initial structure, enabling downstream structural and dynamical analysis.

Overall, our main contributions are as follows:

1. We introduce a continuous-video formulation of protein MD trajectory generation, representing each trajectory as a fixed-resolution, fixed-length tensor in a unified visual space, rather than relying on predefined structural representations such as graphs or geometric tokens.
2. We propose VideoMD, a conditional molecular video generation framework that synthesizes full protein trajectory videos from an initial conformation and simulation configuration, with inference cost largely decoupled from protein length.

3. We introduce Molecular Video Evaluation (MoViE), a structure-aware evaluation protocol for assessing generated protein videos in terms of structural plausibility, temporal coherence, and ensemble statistics.
4. Experiments on ATLAS demonstrate competitive structural validity and strong ensemble and temporal performance. Generated videos can be decoded into coordinate trajectories under structural conditioning, enabling further analysis of protein conformations and dynamics.

2 Related Work

2.1 Protein Dynamics Generation in Predefined Structural Spaces

Existing learning-based approaches to protein conformational modeling and dynamics largely operate in predefined structural spaces, where atoms, residues, or geometric tokens serve as the basic variables for generation. Prior work can be broadly divided into two settings: *equilibrium ensemble modeling*, which aims to sample conformational distributions without explicit temporal modeling (Del Alamo et al., 2022; Wayment-Steele et al., 2024; Noé et al., 2019; Jing et al., 2024a; Zheng et al., 2024; Wang et al., 2024; Lu et al., 2024; 2025; Lewis et al., 2025), and *trajectory modeling*, which seeks to generate temporally consistent transitions or full trajectories (Schreiner et al., 2023; Klein et al., 2023; Costa et al., 2024; Jing et al., 2024b; Xu et al., 2025b; Shen et al., 2025; Shoghi et al., 2026; Cheng et al., 2025; Yu et al., 2026; Shi et al., 2026; Xu et al., 2025a; Feng et al., 2025).

2.2 Video Generative Models

Recent video generative models, particularly diffusion- and transformer-based approaches, have substantially improved temporal coherence and motion quality in high-dimensional video synthesis (Chen et al., 2024; Wang et al., 2023; Kong et al., 2024; OpenAI, 2024; Google DeepMind, 2026). Our setting differs fundamentally from natural video: molecular identity and topology are fixed, camera geometry can be standardized, and the relevant variation is conformational rather than semantic. These properties make protein trajectories a controlled but domain-specific regime for video modeling. In our setting, the goal is not open-ended scene synthesis, but conditional trajectory generation from the initial conformation and simulation configuration in a standardized visual space.

3 Methods

3.1 Continuous MD as Conditional Video Generation

Let $s(t) \in \mathcal{S}$ denote the continuous molecular state at physical time t, and let $c \in \mathcal{C}$ denote the simulation context, such as force field, solvent condition, and requested time horizon. A rendering operator $\mathcal{R} : \mathcal{S} \to \mathcal{X}$ maps a molecular state to a standardized image. Given sampled times $0 = t_0 < t_1 < \cdots < t_{T-1}$, the observed MD video is

$$V = (x_1, \ldots, x_{T-1}), \qquad x_i = \mathcal{R}(s(t_i)). \tag{1}$$

We denote the full rendered clip by $\tilde{V} = (x_0, x_1, \ldots, x_{T-1})$, where $x_i = \mathcal{R}(s(t_i))$. Here T counts all frames, including the conditioning frame x_0; continuous molecular motion is observed at sampled times. VideoMD conditions on the first frame x_0 and simulation context c to model the future trajectory $x_{1:T-1}$, i.e.,

$$p_\theta(x_{1:T-1} \mid x_0, c). \tag{2}$$

This is a conditional distribution over future rendered trajectories.

This formulation differs from direct graph or coordinate generation. Coordinate/graph models learn from a representation $D(s(t))$, where D encodes design choices such as atom or residue nodes, graph edges, and update rules. VideoMD instead learns from $x_t = \mathcal{R}(s(t))$, a dense visual observation after pose and rendering standardization.

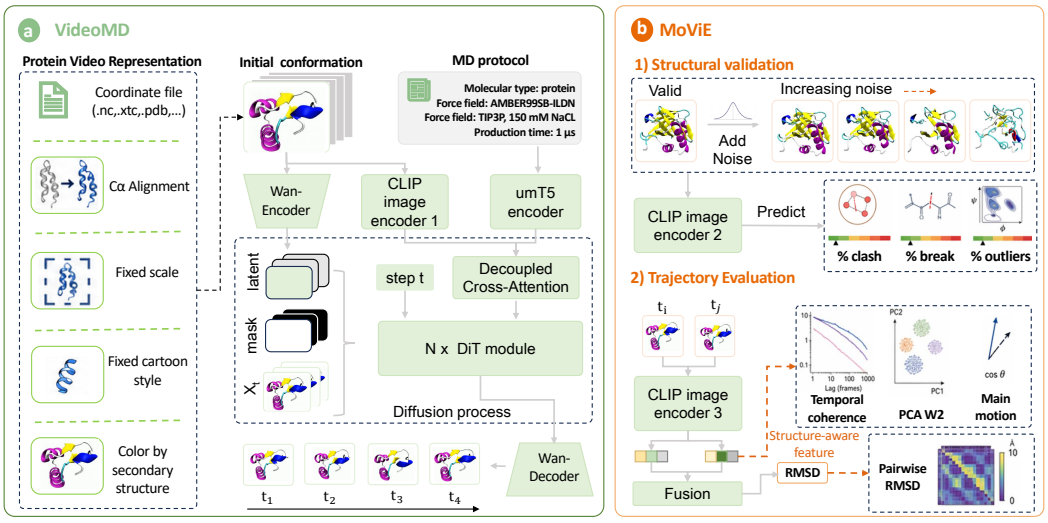


Figure 2: Schematic overview of VideoMD.

3.2 VIDEOMD

Protein Video Representation The rendering pipeline is designed so that pixel variation primarily reflects conformational dynamics (Figure 2a). We use VMD Humphrey et al. (1996) to render each trajectory after aligning all frames to the first frame by C_α atoms using the Kabsch algorithm Kabsch (1976). All frames are rendered with a fixed `NewCartoon` representation, secondary-structure coloring, and fixed material (`AOShiny`), background, view scale, and resolution. This standardized setup reduces nuisance variation from rigid-body motion, so that pixel changes more directly reflect internal conformational dynamics. Secondary-structure coloring further provides consistent visual cues that make local and global structural rearrangements easier to discern. Rendering details are provided in Appendix A.1.

Simulation-Protocol Conditioning Because molecular dynamics depends strongly on simulation conditions, we encode these factors as simulation-setting prompts to condition generation. Each prompt contains the simulated system type, force field, solvent model, ionic condition, and requested production length. Full prompt templates are listed in Appendix B. To evaluate the response to protocol conditioning, we vary the simulation protocol while keeping the initial frame fixed, and assess the resulting changes in generated dynamics in Appendix C. We fix the number of generated frames T across all samples and encode the requested physical horizon in the prompt c. Therefore, different prompts correspond to different physical durations represented over the same number of video frames.

Conditional video generation architecture. We instantiate VideoMD with the open-source Wan image-to-video backbone at 1280×720 (Wan et al., 2025) (Figure 2a) and adapt it to MD generation by conditioning on both structure and simulation protocol. The model takes the first rendered frame x_0, depicting the protein and its initial conformation, and a text prompt $c \in \mathcal{C}$ describing the simulation protocol. These are provided through the image and text conditioning pathways of the video diffusion transformer, respectively. To match the backbone resolution, we convert pre-rendered 2048×2048 frames to 1280×720 using a standardized content-aware crop rather than direct resizing; Appendix A.2 provides details. We fine-tune the transformer with low-rank adaptation (LoRA) (Hu et al., 2022), inserting low-rank adapters into attention and feed-forward projections while freezing the variational autoencoder (VAE) and conditioning encoders. This parameter-efficient setup adapts the pre-trained video model to molecular video generation with limited MD-rendered training data. Because the model operates on fixed-resolution rendered videos rather than variable-size molecular graphs, its inference cost depends primarily on video resolution and clip length, not directly on the number of residues.

Training objective. We train VideoMD using a latent flow-matching objective (Lipman et al., 2023; Wan et al., 2025). Let $V = (x_1, \ldots, x_{T-1})$ denote the target standardized video, and let $z_1 = E_{\mathrm{VAE}}(\tilde{V})$ be the latent representation of the full clip. We sample Gaussian noise $z_0 \sim \mathcal{N}(0, I)$ and a training time $\tau \in [0, 1]$, and define

$$z_\tau = \tau z_1 + (1 - \tau) z_0, \quad v^\star = z_1 - z_0. \tag{3}$$

Conditioned on the prompt embedding $e(c)$, the first-frame image condition $g(x_0)$, and the diffusion time τ, the model predicts a velocity field $u_\theta(z_\tau, \tau, e(c), g(x_0))$. The training loss is

$$\mathcal{L}_{\mathrm{gen}} = \mathbb{E}_{x_0, V, c, z_0, \tau}\left[\left\| m \odot \left(u_\theta(z_\tau, \tau, e(c), g(x_0)) - v^\star\right)\right\|_2^2\right], \tag{4}$$

where m selects the latent tokens corresponding to future frames $x_{1:T-1}$ and excludes the fixed first-frame conditioning tokens. Optimizing this objective trains the model to generate trajectory videos that are consistent with both the initial molecular structure and the simulation protocol.

Trajectory sampling. To accelerate sampling, all reported VideoMD results use an inference pipeline following FastVideo (Hao AI Lab, 2025). We adopt the `FlowUniPCMultistepScheduler`, based on *UniPC* (Zhao et al., 2023), together with PyTorch scaled dot-product attention (SDPA) (PyTorch, 2023) and FP16 VAE decoding. Full implementation details are provided in Appendix D.

3.3 Molecular Video Evaluation

Evaluating molecular videos requires structure-aware measurements alongside visual appearance. We construct MoViE from two learned visual predictors trained against structure-derived supervision (Figure 2b): one estimates structural invalidity ratios from a single rendered frame, and the other predicts the globally aligned structural RMSD between a pair of rendered frames. Both predictors adopt the same Contrastive Language–Image Pre-training (CLIP) visual architecture (Radford et al., 2021) but are trained independently with separate parameters. At evaluation time, all MoViE metrics are computed from rendered videos alone; direct structure-level measurements are used only for predictor training and validation.

Validation predictor. For a single rendered frame x_i, the visual encoder $f_{\phi_{\mathrm{val}}}$ produces an image representation $f_{\phi_{\mathrm{val}}}(x_i)$, which is passed to a quality head $u_{\omega_{\mathrm{val}}}$ to predict three structure-level invalidity ratios, $\widehat{\mathbf{q}}_i = u_{\omega_{\mathrm{val}}}(f_{\phi_{\mathrm{val}}}(x_i))$, corresponding to residue clashes, chain breaks, and Ramachandran outliers. The predictor is trained with mean-squared error against the corresponding targets $\mathbf{q}_i$ computed from the underlying structure, i.e., $\mathcal{L}_{\mathrm{val}} = \mathbb{E}_i \|\widehat{\mathbf{q}}_i - \mathbf{q}_i\|_2^2$.

Corruption-based supervision. To train the validation predictor over a broad range of structural quality, we augment valid structures with synthetic corruptions. Starting from a clean structure s, we perturb residue frames with translational noise and IGSO(3) rotational noise at varying severity levels η, yielding corrupted samples $\tilde{s} \sim \mathcal{C}_\eta(s)$. We then compute the corresponding invalidity targets $\mathbf{q}(\tilde{s})$ from the corrupted geometry, providing continuous supervision over structural degradation rather than a binary label; additional validation details are given in Appendix K.

RMSD predictor. Given rendered frames x_i and x_j, a CLIP-based visual encoder f_{ϕ_r} extracts frame representations, which are combined by a fusion head g_{ψ_r} to predict the aligned structural RMSD, $\widehat{r}_{ij} = g_{\psi_r}([f_{\phi_r}(x_i), f_{\phi_r}(x_j), |f_{\phi_r}(x_i) - f_{\phi_r}(x_j)|])$. The model is trained with mean-squared error, $\mathcal{L}_{\mathrm{RMSD}} = \mathbb{E}_{(i,j)}[(\widehat{r}_{ij} - r_{ij})^2]$, where r_{ij} is the RMSD computed from the underlying aligned structures. This objective encourages f_{ϕ_r} to learn structure-aware visual representations for trajectory quality assessment.

Metric use in MoViE. Structural-validity metrics are obtained by thresholding the validation predictor outputs and reporting frame-pass rates, which provide video-based estimates of local structural plausibility: Clash Pass, Break Pass, Rama Pass, and Overall Pass. We additionally design a set of metrics in the learned representation space, following prior work in Jing et al. (2024b) and Shoghi et al. (2026). Ensemble-statistics metrics are computed from the RMSD predictor and characterize conformational diversity, system-level consistency, and distributional agreement between generated and reference trajectories: Pairwise RMSD, RMSD Pearson r, MD

PCA W_2, JOINT PCA W_2, and PC SIM > 0.3. Temporal-dynamics metrics are evaluated in the learned representation space and measure agreement in lag-dependent motion and memory: LATENT RMSD MAE and LATENT AUTOCORR. MAE. Exact metric definitions and calibrated thresholds are provided in Appendix E.

Coordinate recovery analysis. To assess whether generated videos support downstream structural and dynamical analysis, we use a deterministic, structure-conditioned decoder to recover ordered C_α trajectories and evaluate their conformational and dynamical summaries. A real-video oracle applies the same decoder to rendered reference MD trajectories to assess readout capacity. Decoder training, data splits, and conditioning details are provided in Appendix F; coordinate-recovery results and input controls are reported in Section 4.5 and Appendix G.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. The main experiments use ATLAS, which contains triplicated 100 ns atomistic MD trajectories for 1390 proteins Vander Meersche et al. (2024). We follow the PDB-release-date-based train/validation/test split used by prior protein ensemble and trajectory work Jing et al. (2024a;b); Shen et al. (2025). Dataset and simulation details are in Appendix H.

Implementation. Unless otherwise noted, all reported results use the standard ATLAS train/validation/test split. Full hyperparameters and implementation details are provided in Appendix J. Supplementary protocol-conditioning experiments, including the CATH2 dataset and VideoMD-MIX configuration, are presented in Appendix C.

Baselines. We compare VideoMD with AlphaFolding Cheng et al. (2025), MDGen Jing et al. (2024b), ConfRover Shen et al. (2025), and PVB Yu et al. (2026). We use the official checkpoints for AlphaFolding, MDGen, and ConfRover, and retrain PVB on ATLAS using the same data split as VideoMD. All coordinate-based baseline trajectories are rendered using the same pipeline and evaluated under MoViE. We additionally include the pretrained Wan backbone without LoRA adaptation as a qualitative baseline to examine the role of molecular-domain fine-tuning (Appendix J.1). Further baseline details are provided in Appendix I.

Metrics. We report metrics covering structural validity, ensemble statistics, and temporal dynamics. Exact definitions of all trajectory evaluation metrics are provided in Appendix E. For predictor validation, including the RMSD predictor and validation predictor, we report mean absolute error (MAE) and root-mean-square error (RMSE). Predictor error and cross-view statistics are reported in Appendix K.3. The coordinate-recovery analysis uses conformational and dynamical metrics computed from recovered trajectories (Section 4.5); post-processed structural outputs are reported separately in Appendix G.2.

4.2 BENCHMARK ON ATLAS

We first evaluate VideoMD on the standard 100 ns ATLAS benchmark. Table 1 reports results across three metric groups: structural validity, ensemble statistics, and temporal dynamics. Among generative methods, VideoMD performs best on several key ensemble-level and dynamical metrics under MoViE. In particular, it achieves the highest RMSD PEARSON r (0.7110), suggesting stronger agreement in system-level exploration scale, as well as the best JOINT PCA W_2 (8.9622) and PC SIM > 0.3 (36.59), indicating improved alignment between generated and reference ensembles in representation space. On temporal dynamics, VideoMD attains the best AUTOCORR. MAE (0.0167), reducing error by 73.82% relative to the second-best generative method, and remains competitive on LATENT RMSD MAE (0.1852), where it ranks second and closely matches the best method.

VideoMD remains competitive with most evaluated generative baselines on MoViE-estimated structural validity, ranking second on RAMA PASS (71.85) and OVERALL PASS (45.00). Although MDGen leads all four validity metrics, VideoMD achieves stronger ensemble and temporal agreement

Table 1: MoViE results for the ATLAS 100 ns setting. All scores use rendered videos; temporal and PCA metrics use learned visual features. **Bold** and underlined entries denote the best and second-best generative results. Δ is the relative improvement over the next-best method where VideoMD ranks first. Methods with OOM failures are averaged over successful systems, so their evaluation sets differ.

	Structural validity				Ensemble statistics					Temporal dynamics	
Model	**Clash pass ↑**	**Break pass ↑**	**Rama pass ↑**	**Overall pass ↑**	**Pairwise RMSD**	**RMSD r ↑**	**MD PCA W2 ↓**	**Joint PCA W2 ↓**	**PC sim > 0.3 ↑**	**Latent RMSD MAE ↓**	**Autocorr. MAE ↓**
MD oracle	95.62	97.51	97.15	94.26	1.8547	0.8324	3.1384	7.7361	51.22	0.0288	0.0071
AlphaFolding†	0.45	0.00	0.00	0.00	–	–	–	–	–	–	–
MDGen	**72.46**	**91.00**	**76.93**	**69.22**	1.1537	0.4823	4.7587	15.1752	9.76	0.8832	0.0662
ConfRover‡	50.98	68.54	49.02	42.07	1.7433	0.5859	4.3534	10.9783	35.44	**0.1778**	0.0638
PVB	52.90	88.44	30.61	20.85	4.3013	0.3423	**4.0249**	16.7462	1.22	0.6765	0.1616
VideoMD	50.12	86.26	71.85	45.00	1.5680	**0.7110**	4.3437	**8.9622**	**36.59**	0.1852	**0.0167**
Δ (%)	–	–	–	–	–	+21.35	–	+18.36	+3.25	–	+73.82

OOM: out of memory. † AlphaFolding: 5 OOM cases. ‡ ConfRover: 3 OOM cases.

under MoViE. These results show that the molecular-video formulation combines competitive structural plausibility with advantages in modeling conformational distributions and temporal behavior.

Figure 3 further examines temporal metrics and runtime scaling. Panels (a) and (b) show that VideoMD closely tracks the MD reference across lags in both latent RMSD and latent autocorrelation, indicating closer agreement in feature displacement and autocorrelation within the MoViE representation. These feature-space results do not directly establish physical dynamical fidelity. By contrast, the baselines show larger deviations at longer lags, with some methods overestimating latent RMSD or exhibiting faster autocorrelation decay.

Panel (c) examines scaling with sequence length. At fixed resolution and clip length, measured VideoMD generation time spans 635–677 s across the tested 51–724-residue systems. Visual Joint PCA W_2 is 10.13 for $L \leq 200$, 10.03 for $200 < L \leq 400$, and 9.66 for $L > 400$. These measurements characterize the fixed-size video generator; the plotted baseline timings use different hardware and are not a speed comparison with MD or a matched end-to-end benchmark. The numerical values used to plot Figure 3 are provided in Appendix L.

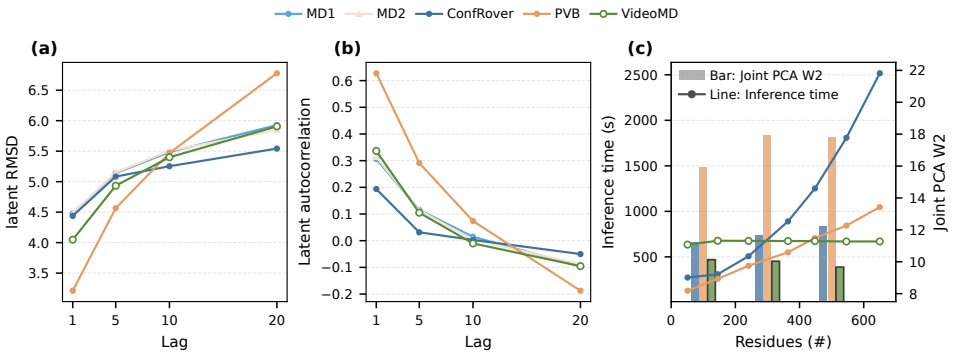


Figure 3: MoViE temporal metrics and runtime scaling. (a) Latent RMSD and (b) autocorrelation versus lag. (c) Joint PCA W_2 by protein-length bin (bars) and video-generation time excluding coordinate processing (lines); baseline hardware differs.

4.3 Qualitative Case Study

Figure 4 provides a qualitative comparison between a reference MD trajectory and a generated trajectory on a representative ATLAS example (PDB ID 6d7y_A). This case study complements the quantitative results by illustrating whether the model preserves large-scale structural organization while producing nontrivial temporal motion. Panel A compares reference MD frames and generated

frames at matched time points. The generated trajectory maintains the overall secondary-structure layout and the relative organization of the central helix, beta sheet, and flexible loop regions, while still exhibiting visible conformational change over time. Panel B shows that the generated trajectory also reproduces an RMSD fluctuation range comparable to that of the reference trajectory. The timing of individual peaks is not expected to match exactly, however, because conformational dynamics are inherently stochastic. Overall, this example suggests that the model captures plausible coarse-grained structural evolution rather than merely reproducing static appearance. Additional qualitative examples are provided in Appendix L.

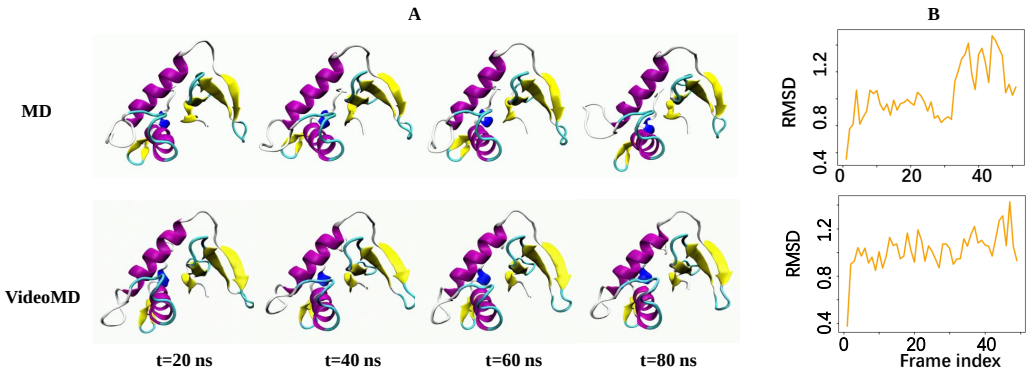


Figure 4: Qualitative case study for PDB ID 6d7y_A. (A) Reference MD and VideoMD frames at matched time points. (B) RMSD to the initial frame as a function of frame index for the reference and generated trajectories. VideoMD preserves the global secondary-structure layout and produces a comparable RMSD fluctuation range.

4.4 Consistency Between MoViE and Coordinate-Space Metrics

To assess whether MoViE preserves method-level trends observed in coordinate space, we compare Pairwise RMSD r, AutoCor MAE, and RMSD-curve MAE for the three methods with values available in both evaluations: MDGen, ConfRover, and PVB (Figure 5). The associations are positive for all three metrics, with Pearson $r = 0.9996$, 0.9690, and 0.7843, respectively. These results provide preliminary evidence of method-level trend agreement in this limited comparison. Because each comparison contains only three methods and the coordinate-space values come from an earlier evaluation setting, this analysis is descriptive rather than a general calibration test. It does not assess evaluator bias on native VideoMD artifacts.

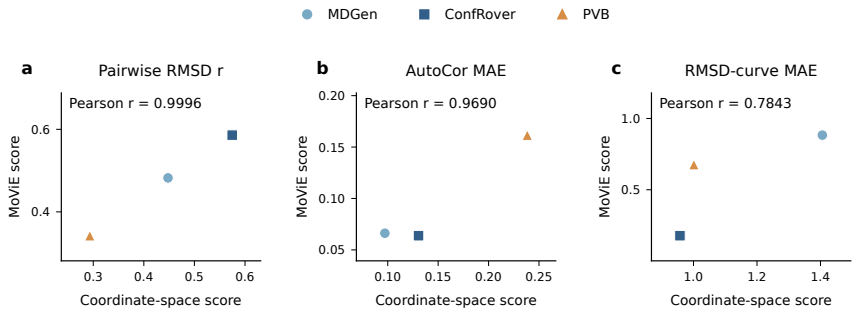


Figure 5: MoViE versus coordinate-space scores for MDGen, ConfRover, and PVB. Pearson correlations are descriptive ($n = 3$ methods per panel).

Table 2: Coordinate metrics against reference MD. Baseline cohorts are not matched to current VideoMD results. Bold compares generated-video inputs only; the oracle uses reference MD videos.

		Conformational statistics			Dynamical statistics	
Group	**Method / input**	Pairwise RMSD $r \uparrow$	MD-PCA $W_2 \downarrow$	Joint-PCA $W_2 \downarrow$	RMSD-curve MAE $\downarrow$	AutoCor-curve MAE $\downarrow$
Coordinate-based baselines	MDGen	0.4478	1.7645	2.3645	1.4058	0.09716
	ConfRover	0.5747	1.6354	2.5363	0.9573	0.13054
	PVB	0.2932	1.8307	6.2665	1.0012	0.23838
Generated-video inputs	VideoMD	**0.6920**	**1.6842**	**2.3382**	**0.6721**	**0.07224**
	Repeated first frame	0.2252	1.9818	2.5780	2.3073	0.12714
	Shuffled input	0.5004	1.8102	2.7972	2.3318	0.13474
Oracle	Reference MD video	0.8808	1.5120	1.9887	0.2788	0.03664

4.5 Coordinate Recovery and Dynamical Fidelity

We examine whether VideoMD's visual outputs support coordinate-based analysis of protein motion. Using the same deterministic, structure-conditioned decoder, we compare original generated videos with repeated-first-frame and shuffled-input controls. Decoded trajectories are evaluated against reference MD statistics independently of MoViE, with decoded reference MD videos serving as an oracle.

Coordinate recovery beyond visual appearance. Table 2 shows that reference-video decoding achieves a Pairwise RMSD correlation of 0.8808 and an RMSD-curve MAE of 0.2788, establishing a useful reference for the decoder's recovery performance. With generated videos, the corresponding scores are 0.6920 and 0.6721. Although a gap to the oracle remains, VideoMD outperforms both input controls across all five metrics, covering conformational distributions and temporal behavior. These results extend the evidence beyond visual evaluation: generated molecular videos support the recovery of coordinate trajectories with measurable agreement to reference MD statistics.

Coordinate-space evidence beyond MoViE. Directly computed from recovered coordinates, VideoMD's Pairwise RMSD r (0.6920), RMSD-curve MAE (0.6721), and autocorrelation-curve MAE (0.07224) provide evidence beyond MoViE features. These scores and Joint-PCA W_2 are numerically favorable relative to the historical baselines, while ConfRover has lower MD-PCA W_2; unmatched cohorts limit this comparison to descriptive evidence.

Generated video content contributes to coordinate recovery. With the decoder and structural conditions fixed, repeating the first frame and shuffling future frames increase RMSD-curve MAE from 0.6721 to 2.3073 and 2.3318, respectively, and worsen autocorrelation and distributional scores. Thus, generated conformational changes and their temporal ordering contribute to recovery beyond the shared structural conditions and decoder priors, supporting molecular videos as inputs for structural and dynamical analysis.

5 Conclusion

We introduced VideoMD, a framework for generating protein dynamics as molecular videos conditioned on an initial frame and simulation protocol. By representing trajectories in a standardized visual space, VideoMD adapts pretrained video generation to molecular motion. On ATLAS, it achieves strong ensemble and temporal performance under MoViE, while structural validity remains an area for improvement. At fixed resolution and clip length, generation time varies little across the tested protein sizes. Beyond visual evaluation, structure-conditioned decoding recovers coordinate trajectories for conformational and dynamical analysis. These findings support molecular videos as a useful representation for learning protein motion and connecting visual generation with coordinate-based analysis.

AI USE STATEMENT

For this manuscript revision, generative AI was used solely to assist with manuscript writing, including language editing, grammatical correction, and stylistic polishing. All content was reviewed and approved by the authors.

REPRODUCIBILITY STATEMENT

Rendering/crop and prompt specifications are given in Appendices A–B; protocol-conditioning experiments appear in Appendix C; and sampling and MoViE metric definitions appear in Appendices D–E. Appendix F documents the coordinate decoder, and Appendix G reports coordinate recovery, input controls, structural repair, and stage-wise costs. Dataset, baseline, and generation configurations are documented in Appendices H–J. Predictor error and cross-view statistics are documented in Appendix K.3, with additional benchmark measurements in Appendix L.

REFERENCES

A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, 2001. doi: 10.1016/S0006-3495(01)76033-X.

Herman JC Berendsen and Steven Hayward. Collective protein dynamics in relation to function. *Current opinion in structural biology*, 10(2):165–169, 2000.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7310–7320, 2024.

Kaihui Cheng, Ce Liu, Qingkun Su, Jun Wang, Liwei Zhang, Yining Tang, Yao Yao, Siyu Zhu, and Yuan Qi. 4d diffusion for dynamic protein structure prediction with reference and motion guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 93–101, 2025.

Allan Dos Santos Costa, Ilan Mitnikov, Franco Pellegrini, Ameya Daigavane, Mario Geiger, Zhonglin Cao, Karsten Kreis, Tess Smidt, Emine Kucukbenli, and Joseph Jacobson. Equijump: Protein dynamics simulation via so (3)-equivariant stochastic interpolants. *arXiv preprint arXiv:2410.09667*, 2024.

David Del Alamo, Davide Sala, Hassane S. Mchaourab, and Jens Meiler. Sampling alternative conformational states of transporters and receptors with alphafold2. *eLife*, 11:e75751, 2022.

Ron O Dror, Robert M Dirks, JP Grossman, Huafeng Xu, and David E Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41:429–452, 2012.

Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, 13(7):e1005659, 2017. doi: 10.1371/journal.pcbi.1005659.

Bin Feng, Jiying Zhang, Xinni Zhang, Zijing Liu, and Yu Li. Biomd: All-atom generative model for biomolecular dynamics simulation. *arXiv preprint arXiv:2509.02642*, 2025.

FFmpeg Developers. ffmpeg tool. `http://ffmpeg.org/`, 2021. Version 4.4.2.

Google DeepMind. Veo, 2026. URL `https://deepmind.google/models/veo/`. Accessed: 2026-04-15.

Hao AI Lab. Fastvideo: A unified inference and post-training framework for accelerated video generation. `https://github.com/hao-ai-lab/FastVideo`, 2025. Accessed: 2026-04-27.

Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

Lim Heo and Michael Feig. One bead per residue can describe all-atom protein structures. *Structure*, 32(1):97–111.e6, 2024. doi: 10.1016/j.str.2023.10.013.

Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024a.

Bowen Jing, Hannes Stärk, Tommi Jaakkola, and Bonnie Berger. Generative modeling of molecular dynamics trajectories. *Advances in Neural Information Processing Systems*, 37:40534–40564, 2024b.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2307.07635.

Leon Klein, Andrew Foong, Tor Fjelde, Bruno Mlodozeniec, Marc Brockschmidt, Sebastian Nowozin, Frank Noé, and Ryota Tomioka. Timewarp: Transferable acceleration of molecular dynamics by learning time-coarsened dynamics. *Advances in Neural Information Processing Systems*, 36:52863–52883, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761):eadv9817, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jiarui Lu, Bozitao Zhong, Zuobai Zhang, and Jian Tang. Str2str: A score-based framework for zero-shot protein conformation sampling. In *International Conference on Learning Representations*, volume 2024, pp. 10663–10694, 2024.

Jiarui Lu, Xiaoyin Chen, Stephen Zhewen Lu, Chence Shi, Hongyu Guo, Yoshua Bengio, and Jian Tang. Structure language models for protein conformation generation. In *International Conference on Learning Representations*, volume 2025, pp. 98645–98678, 2025.

Artur Meller, Michael D. Ward, Jonathan Borowsky, Meghana Kshirsagar, Jeffrey M. Lotthammer, Felipe Oviedo, Juan M. Lavista Ferres, and Gregory R. Bowman. Predicting locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Nature Communications*, 14, 2023. URL https://api.semanticscholar.org/CorpusID:257281992.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual review of physical chemistry*, 71(1):361–390, 2020.

OpenAI. Video generation models as world simulators, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/. Accessed: 2026-04-15.

PyTorch. torch.nn.functional.scaled_dot_product_attention. https://docs.pytorch.org/docs/stable/generated/torch.nn.functional.scaled_dot_product_attention.html, 2023. Accessed: 2026-04-27.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Mathias Schreiner, Ole Winther, and Simon Olsson. Implicit transfer operator learning: Multiple time-resolution models for molecular dynamics. *Advances in Neural Information Processing Systems*, 36:36449–36462, 2023.

Yuning Shen, Lihao Wang, Huizhuo Yuan, Yan Wang, Bangji Yang, and Quanquan Gu. Simultaneous modeling of protein conformation and dynamics via autoregression. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Nima Shoghi, Yuxuan Liu, Yuning Shen, Rob Brekelmans, Pan Li, and Quanquan Gu. Scalable spatio-temporal se (3) diffusion for long-horizon protein dynamics. *arXiv preprint arXiv:2602.02128*, 2026.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2003.12039.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research*, 52 (D1):D384–D392, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Yan Wang, Lihao Wang, Yuning Shen, Yiqun Wang, Huizhuo Yuan, Yue Wu, and Quanquan Gu. Protein conformation generation via force-guided SE(3) diffusion models. *arXiv preprint arXiv:2403.14088*, 2024.

Hannah K Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell, and Dorothee Kern. Predicting multiple conformations via sequence clustering and alphafold2. *Nature*, 625(7996):832–839, 2024.

Kai Xu, Jianmin Wang, Mingquan Liu, Kewei Zhou, Shaolong Lin, Weihong Li, Lin Shi, Peng Zhou, Huanxiang Liu, and Xiaojun Yao. Efficient generation of protein and protein–protein complex dynamics via se (3)-parameterized diffusion models. *Journal of Chemical Information and Modeling*, 65(22):12366–12376, 2025a.

Yaoyao Xu, Di Wang, Zihan Zhou, Tianshu Yu, and Mingchen Chen. Tempo: Temporal multi-scale autoregressive generation of protein conformational ensembles. *arXiv preprint arXiv:2511.05510*, 2025b.

Ziyang Yu, Wenbing Huang, and Yang Liu. Unified biomolecular trajectory generation via pretrained variational bridge. *arXiv preprint arXiv:2602.07588*, 2026.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36: 49842–49869, 2023.

Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*, 6(5):558–567, 2024.

A VIDEO RENDERING DETAILS

A.1 VMD RENDERING.

MD trajectories are rendered into videos using VMD 1.9.4. The default representation is removed and replaced with a NewCartoon representation colored by secondary structure and rendered with the AOShiny material. The display uses a white background, axes are disabled, and GLSL rendering is enabled. To emphasize intra-protein conformational dynamics rather than global translation and rotation, each sampled frame is rigidly aligned to the first frame using protein C_α atoms. Specifically, if x_i denotes frame i, a least-squares rigid transform is estimated with `measure fit` between selections `protein and name CA` in frame i and frame 0, and the resulting transformation is applied to all atoms in the frame. Frames are rendered with Tachyon at 2048×2048 resolution with 12 anti-aliasing samples and exported as BMP images. The image sequence is assembled into an H.264 video using ffmpeg (FFmpeg Developers, 2021) at 15 frames per second with `libx264`, `preset=slow`, `crf=18`, `pix_fmt=yuv420p`, and `+faststart`.

Table 3: Protein Video Rendering settings.

Parameter	Value	Purpose
Representation	NewCartoon	Cartoon-style protein rendering
Coloring	Structure	Secondary-structure coloring
Material	AOShiny	Fixed material and shading
Axes	Off	Remove coordinate-axis artifacts
Background	White	Standardized background
Rendering mode	GLSL	Fixed OpenGL shading
Alignment	protein and name CA, frame 0	Remove global rigid-body motion
View scaling	scale by 0.9	Avoid edge clipping

A.2 VIDEO CROPPING AND STANDARDIZATION.

Rendered videos are post-processed to obtain a uniform input format. For video-generation training, we compute a single global crop for each clip by scanning all frames, thresholding non-white content with a white threshold of 240, expanding the union content box with a 30-pixel safety margin, padding with color `0xfcfffc` to the 16:9 target aspect ratio, and resizing to 1280×720. For RMSD-predictor and validation-model training, the same content-aware crop-and-pad procedure is applied with the final target changed to 224×224, followed by CLIP-style normalization.

B PROMPT TEMPLATES

Table 4: Full MD-setting prompts used for conditional generation.

Source	Prompt
ATLAS	Simulations were performed on proteins only, using the all-atom CHARMM36m force field (July 2020) in explicit TIP3P water within a periodic triclinic box, with Na^+/Cl^- ions added to neutralise the system at 150 mM, and a production run length of 100 ns. Keep the same chain connectivity and backbone topology as the input frame.
CATH2	Simulations were performed on proteins only, using the all-atom ff99SB-ILDN force field in explicit TIP3P water within a periodic triclinic box, with Na^+/Cl^- ions added to neutralise the system at 100 mM, and a production run length of 1000 ns. Keep the same chain connectivity and backbone topology as the input frame.

C Simulation Protocol-Conditioned Exploration

We study how generated dynamics respond to simulation-protocol prompts while the initial conformation is held fixed.

Supplementary data and model. We train VideoMD-MIX on the union of ATLAS and the BioEMU-derived CATH2 dataset (Lewis et al., 2025). CATH2 contains 871 proteins with three replicate 1000 ns simulations per protein and uses a similar PDB-release-date-based split. The BioEMU domain-motion set supplies 22 large-scale hinge-motion examples for testing. VideoMD-MIX uses the same architecture, training objective, optimization, and sampling settings as the base model; its validation and RMSD predictors are trained on the same mixed training data. Dataset and implementation details are provided in Appendices H and J.

Protocol comparison. We prompt the mixed model with either ATLAS or CATH2 settings, which differ in production length (100 ns versus 1000 ns), force field, and ionic condition. Both protocols are represented during training. We assess changes in exploration while retaining dependence on the starting state.

Table 5: Conformational diversity under simulation-protocol conditioning. Values are predicted pairwise RMSD in Å aggregated over all intra-video frame pairs for each prompt; larger central and upper-quantile values indicate broader generated exploration.

Prompt	Mean	Median	Std.	P10	P90
ATLAS	2.421	1.350	3.483	0.858	5.226
CATH2	2.896	1.703	3.512	0.987	6.276

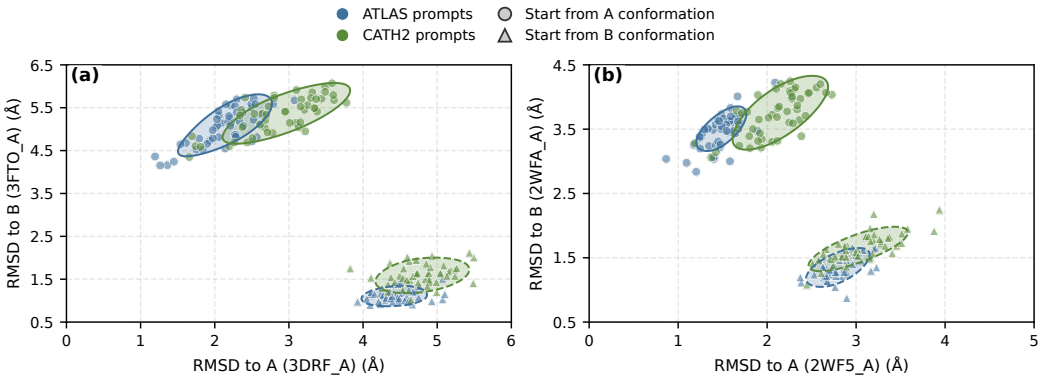


Figure 6: Two-state generation examples under simulation-protocol conditioning on ATLAS and CATH2. Samples initialized from conformations A or B are shown by their RMSDs to the two endpoint states.

Table 5 reports larger predicted pairwise RMSD under the CATH2 prompt than under ATLAS: mean 2.896 versus 2.421, median 1.703 versus 1.350, and P90 6.276 versus 5.226. Figure 6 illustrates prompt-dependent exploration for two systems and both initial conformations. The table summarizes intra-video frame-pair statistics; the two displayed systems are qualitative examples, not independent estimates of population-level controllability. Per-system and sampling uncertainty are not quantified.

Interpretation and scope. Because production length, force field, and ionic conditions change together, the experiment does not isolate the causal contribution of any single setting or establish generalization to unseen protocols. The reported changes in predicted exploration support a response to protocol conditioning, not independently calibrated physical control.

D TRAJECTORY SAMPLING DETAILS

Accelerated inference configuration. All reported VideoMD samples are generated with an accelerated Wan inference pipeline following FastVideo (Hao AI Lab, 2025). Specifically, we use FastVideo's `FlowUniPCMultistepScheduler`, which applies *UniPC* (Zhao et al., 2023), a training-free higher-order predictor–corrector solver for diffusion and flow-based models. Compared with simpler first-order updates, UniPC provides a more accurate numerical integration of the reverse dynamics, improving sample quality per denoising step and enabling efficient sampling with 20 denoising steps in our experiments. To further reduce wall-clock latency, attention is computed with PyTorch scaled dot-product attention (SDPA) (PyTorch, 2023), and VAE decoding is performed in FP16. Unless otherwise noted, this accelerated configuration is used for all quantitative and qualitative results in the paper.

E MOVIE METRIC IMPLEMENTATION DETAILS

This appendix defines the video-native MoViE metrics. The coordinate-space measurements in Table 2 are distinct from these visual metrics; their evaluation provenance is described in Appendix G. Consistent with Section 3.3, we group the reported metrics into three categories: *structural validity*, *ensemble statistics*, and *temporal dynamics*. Structural-validity metrics are derived from the validation predictor, while ensemble-statistics and temporal-dynamics metrics are derived from the RMSD predictor through either its pairwise RMSD estimates $\widehat{r}_{ij}$ or its learned frame representations.

Structural validity metrics and thresholds. For coordinate-bearing structures used in predictor supervision and validation, we compute three structure-level invalidity ratios: chain discontinuity, steric self-overlap, and unfavorable backbone torsions. These are supervision targets, not coordinates reconstructed from generated videos; at video evaluation time, the corresponding ratios are estimated by the visual predictor.

Let N_{adj} denote the number of chemically adjacent residue pairs with C_α atoms. The chain-break ratio measures local backbone discontinuity:

$$q_{\mathrm{break}} = \frac{\#\{(i,j) \in \mathrm{Adj} : \|C_{\alpha,i} - C_{\alpha,j}\|_2 > d_{\mathrm{break}}\}}{N_{\mathrm{adj}}}, \qquad d_{\mathrm{break}} = 4.5 \text{ Å}.$$

For non-adjacent residues, we define close-contact violations using $d_{\mathrm{clash}} = 3.0$ Å. The residue-clash ratio measures steric crowding and is defined as the fraction of residues that participate in at least one such violation:

$$q_{\mathrm{clash}} = \frac{\#\{i : \exists j \notin \mathrm{Adj}(i), \|C_{\alpha,i} - C_{\alpha,j}\|_2 < d_{\mathrm{clash}}\}}{N_{\mathrm{res}}}.$$

The Ramachandran-outlier ratio measures local backbone-torsion plausibility:

$$q_{\mathrm{rama}} = \frac{N_{\mathrm{outlier}}}{N_{\mathrm{evaluated}}},$$

computed with Phenix `ramalyze` when available and an internal ϕ/ψ-based fallback otherwise.

A frame is counted as passing each structural criterion when

$$q_{\mathrm{clash}} < \tau_{\mathrm{clash}}, \qquad q_{\mathrm{break}} < \tau_{\mathrm{break}}, \qquad q_{\mathrm{rama}} < \tau_{\mathrm{rama}},$$

with fixed thresholds

$$\tau_{\mathrm{clash}} = 0.002, \qquad \tau_{\mathrm{break}} = 0.005, \qquad \tau_{\mathrm{rama}} = 0.051471.$$

Let T denote the number of evaluated frames, and let $\widehat{q}_{\mathrm{clash},t}$, $\widehat{q}_{\mathrm{break},t}$, and $\widehat{q}_{\mathrm{rama},t}$ be the three invalidity ratios predicted for frame t by the validation predictor. The four MoViE structural-validity metrics are then

$$\textsc{Clash Pass} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[\widehat{q}_{\mathrm{clash},t} < \tau_{\mathrm{clash}}],$$

$$\textsc{Break Pass} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[\widehat{q}_{\mathrm{break},t} < \tau_{\mathrm{break}}]\,,$$

$$\textsc{Rama Pass} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[\widehat{q}_{\mathrm{rama},t} < \tau_{\mathrm{rama}}]\,,$$

and

$$\textsc{Overall Pass} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[\widehat{q}_{\mathrm{clash},t} < \tau_{\mathrm{clash}} \ \wedge\ \widehat{q}_{\mathrm{break},t} < \tau_{\mathrm{break}} \ \wedge\ \widehat{q}_{\mathrm{rama},t} < \tau_{\mathrm{rama}}]\,.$$

Thus, CLASH PASS, BREAK PASS, and RAMA PASS each measure the fraction of frames that satisfy one structural criterion, while OVERALL PASS measures the fraction that satisfy all three simultaneously. In other words, the first three metrics isolate specific failure modes, whereas OVERALL PASS serves as a stricter summary of overall frame-level plausibility.

When direct structure access is available, the same pass indicators can be computed from the true ratios $q_{\mathrm{clash}}, q_{\mathrm{break}}, q_{\mathrm{rama}}$. In MoViE, however, we report these quantities from rendered videos alone by applying the thresholds to the predictor outputs $\widehat{\mathbf{q}}_t$.

Table 6 summarizes threshold calibration. The structural-validity thresholds for clash and break are set manually to small tolerance levels: 0.2% for clash residue ratio and 0.5% for break pair ratio. The Ramachandran threshold is the reported 99th-percentile reference-MD value. The table records the calibration value; its exact calibration split should be retained with the evaluation configuration.

Table 6: Calibration of structural validity thresholds. Clash and break use fixed tolerance thresholds, while the Ramachandran threshold is calibrated from the 99th percentile of the reference-MD distribution (ATLAS and CATH2).

Metric	Reference-MD percentile	Reference-MD value	Threshold source	Threshold used
Clash residue ratio	99	0.000000	Fixed tolerance	0.002000
Break pair ratio	99	0.000000	Fixed tolerance	0.005000
Ramachandran outlier ratio	99	0.051471	Reference MD, 99th percentile	0.051471

The OVERALL PASS criterion is intentionally conservative: a frame is counted as reasonable only if all three predicted or measured invalidity ratios fall below their respective thresholds.

Ensemble statistics. The RMSD predictor maps a pair of rendered frames (x_i, x_j) to a predicted pairwise RMSD value $\widehat{r}_{ij}$, and its visual encoder also produces frame representations used for feature-based metrics. We report PAIRWISE RMSD as the mean predicted pairwise RMSD over all frame pairs (i, j) with $i < j$, which serves as a proxy for conformational diversity. Across systems, we report RMSD r, the Pearson correlation between generated and reference PAIRWISE RMSD summaries, measuring whether the method preserves system-level differences in exploration scale.

We additionally compute feature-distribution metrics from the RMSD-predictor representations. Let Z^{ref} and Z^{gen} denote the resulting feature matrices. In MD PCA, the PCA basis is fit on Z^{ref} alone, and both reference and generated features are projected into that basis; in the main results, we use the first 2 principal components. In JOINT PCA, PCA is fit on the concatenated features $[Z^{\mathrm{ref}}; Z^{\mathrm{gen}}]$; unless otherwise noted, we again report results using the first 2 principal components. In both cases, we compute a Gaussian 2-Wasserstein distance between the projected feature distributions:

$$W_2^2 = \|\mu_{\mathrm{gen}} - \mu_{\mathrm{ref}}\|_2^2 + \mathrm{Tr}\Big(\Sigma_{\mathrm{gen}} + \Sigma_{\mathrm{ref}} - 2(\Sigma_{\mathrm{gen}}^{1/2}\Sigma_{\mathrm{ref}}\Sigma_{\mathrm{gen}}^{1/2})^{1/2}\Big)\,.$$

Lower values indicate closer agreement between generated and reference feature distributions. We also report PC SIM > 0.3, the percentage of systems whose first principal components have absolute cosine similarity greater than 0.3.

Temporal dynamics. Temporal-dynamics metrics are computed from RMSD-predictor representations after PCA projection. Let $u_t \in \mathbb{R}^D$ denote the projected latent trajectory for a video of length

T, where the PCA basis is fit on the reference features Z^{ref} and, in the main results, $D = 32$. The lagged latent RMSD curve is defined as

$$L(k) = \frac{1}{T-k} \sum_{t=1}^{T-k} \|u_{t+k} - u_t\|_2,$$

and the mean autocorrelation curve is

$$A(k) = \frac{1}{D} \sum_{d=1}^{D} \frac{\frac{1}{T-k} \sum_{t=1}^{T-k} (u_{t,d} - \bar{u}_d)(u_{t+k,d} - \bar{u}_d)}{\frac{1}{T} \sum_{t=1}^{T} (u_{t,d} - \bar{u}_d)^2 + \epsilon}.$$

We report LATENT RMSD MAE as the mean absolute error between generated and reference $L(k)$ curves evaluated at lags $k \in \{1, 5, 10, 20\}$, and LATENT AUTOCORR. MAE as the mean absolute error between generated and reference $A(k)$ curves evaluated over the same set of lags. Lower values indicate better agreement with the temporal scale and memory of the reference dynamics.

F COORDINATE DECODING FOR ASSESSING GENERATED MOLECULAR VIDEOS

F.1 INPUTS, OUTPUTS, AND THE ROLE OF THE READOUT

VideoMD learns a distribution over molecular videos. The coordinate-decoding procedure is a downstream readout of the geometry accessible from a generated observation sequence; it does not change the generator's output space or introduce a second stochastic trajectory generator. Its inputs are a 49-frame RGB video $\widehat{V}$, the initial C_α coordinates $X_0 \in \mathbb{R}^{N\times 3}$, the amino-acid sequence σ, and a known orthographic camera Π. Its output is an ordered C_α trajectory $\widehat{X}_{0:T-1}$ with $T = 49$ and $\widehat{X}_0 = X_0$. All indexing here includes frame 0 in the total of 49 frames.

The structure-conditioned readout used for this analysis is

$$\widehat{X}_{0:T-1} = D_\psi(\widehat{V}; X_0, \sigma, \Pi), \qquad \widehat{X}_0 = X_0, \tag{5}$$

where D_ψ denotes the deterministic decoding procedure described below.

Initial geometry supplies residue correspondence, molecular topology, and a structural prior. The camera relates the three-dimensional initial state to its projected appearance. This turns decoding into registered non-rigid deformation estimation. The decoder receives image-motion observations and does not access the Wan VAE or diffusion-transformer features. Evaluation uses its deterministic mean output and a single inference configuration for all targets.

F.2 RESIDUE-INDEXED IMAGE MOTION

Let $p_i^0 = \Pi(X_{0,i})$ be the image location associated with residue i in the initial frame. Frozen RAFT (Teed & Deng, 2020) supplies a flow estimate $F_{0\to t}$, and frozen CoTracker (Karaev et al., 2024) supplies long-range tracks p_{it}^C. Their residue-indexed displacement observations are

$$u_{it}^R = F_{0\to t}(p_i^0), \qquad u_{it}^C = p_{it}^C - p_i^0. \tag{6}$$

RAFT confidence c_{it}^R uses forward–backward consistency and foreground/in-frame validity; CoTracker provides visibility-based confidence c_{it}^C. We fuse supported observations as

$$u_{it} = \frac{c_{it}^R u_{it}^R + c_{it}^C u_{it}^C}{c_{it}^R + c_{it}^C}, \qquad c_{it}^R + c_{it}^C > 0, \tag{7}$$

and use the combined confidence

$$c_{it} = \left[1 - (1 - c_{it}^R)(1 - c_{it}^C)\right] \exp\left(-\frac{\|u_{it}^R - u_{it}^C\|_2}{\tau}\right). \tag{8}$$

Here τ controls the penalty for disagreement. Observations with no support have zero weight. Correspondence remains anchored to X_0, while occlusion and inconsistent image motion reduce the influence of a residue's observations on lifting. The projected sites are an indexing device; a cartoon rendering need not expose every C_α atom as a separately visible point.

F.3 Structure-Constrained Lifting

From X_0, we construct K non-rigid, low-frequency ANM modes $B_k \in \mathbb{R}^{N\times 3}$ (Atilgan et al., 2001). Let $P \in \mathbb{R}^{2\times 3}$ denote the linear part of the orthographic projection. The image-plane displacement observations initialize each frame's mode amplitudes through

$$a_t^{(0)} = \arg\min_{a\in\mathbb{R}^K} \sum_i c_{it} \left\| P\left(\sum_{k=1}^{K} B_{k,i} a_k\right) - u_{it} \right\|_2^2 + \lambda_{\text{ridge}} \|a\|_2^2. \tag{9}$$

A temporal encoder predicts a correction Δa_t, giving

$$a_t = a_t^{(0)} + \Delta a_t, \qquad X_t^{\text{coarse}} = X_0 + \sum_{k=1}^{K} B_k a_{kt}. \tag{10}$$

This low-frequency representation supports collective deformations consistent with the initial topology, and regularizes directions that are poorly constrained by the projection. It is a structural prior, not a claim that a single view uniquely observes depth.

F.4 Equivariant Residual Decoding and Training

Low-frequency modes do not describe all local motion. An SE(3)-equivariant graph–temporal decoder therefore predicts bounded residuals r_{it} and uncertainty. Its covalent/contact graph combines sequence, initial geometry, tracker observations and confidences, coarse displacement, and temporal mode context. Final future-frame coordinates are

$$\widehat{X}_{t,i} = X_{t,i}^{\text{coarse}} + r_{it}, \qquad 1 \le t < T, \tag{11}$$

with the initial frame fixed to X_0 and its residual fixed to zero. The equivariance claim concerns the geometric residual decoder; the observation interface additionally depends on the supplied camera. Reported evaluations use the deterministic mean rather than uncertainty sampling.

Paired supervision comes from rendered MD trajectories. Decoder training follows the same protein-level ATLAS split used for VideoMD: rendered trajectories from the training partition are used for fitting, the validation partition is used for checkpoint selection, and the selected checkpoint is frozen before evaluation on the 82-protein test partition. The structural condition is the initial C_α geometry associated with the retained first video frame. All four evaluation conditions use this frozen checkpoint. Both input controls retain the first frame and its structural condition.

The decoder is trained with

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}}. \tag{12}$$

Here $\mathcal{L}_{\text{rec}}$ combines coordinate and reprojection supervision. The geometric term $\mathcal{L}_{\text{geo}}$ penalizes deviations in distances, bonds, contacts, and excluded volume. The dynamic term $\mathcal{L}_{\text{dyn}}$ matches derivatives, RMSF, and lag-dependent behavior. These are training objectives, not hard guarantees that every decoded frame satisfies local stereochemistry; the raw/refined evaluation explicitly tests that distinction.

F.5 Optional Repair and All-Atom Completion

Restrained OpenMM minimization (Eastman et al., 2017) repairs local geometry of the recovered C_α trajectory. This is local structural optimization of an existing ordered sequence, not an MD rollout generating a new motion sequence. Structural-quality measurements before and after refinement are reported in Table 7.

Optional cg2all backmapping (Heo & Feig, 2024) produces all-atom completions of the recovered C_α conformers for downstream structure-based analysis. The Ramachandran favored rate assesses those completed structures. All-atom completion adds a learned structural prior; it is not evidence that the single-view video directly resolved side-chain coordinates or their true dynamics.

Table 7: Structural quality before and after restrained OpenMM refinement in a separate paired evaluation.

Metric	Raw recovery	After refinement
C_α validity (%) ↑	26.815	**97.443**
Break rate (%) ↓	1.9582	**0.0250**
Bond-length MAE (Å) ↓	0.1625	**0.0633**

G Coordinate Recovery, Structural Repair, and Cost

G.1 Evaluation Provenance and Metric Interpretation

Table 2 reports measurements on the 82-protein ATLAS test partition for original, repeated-first-frame, and shuffled generated-video input, together with a real-video oracle. All conditions use the same decoder checkpoint selected on the validation partition and frozen before test evaluation. The oracle renders reference MD trajectories and applies the same structure-conditioned decoder and coordinate metrics, providing a direct check of readout capacity rather than another generative baseline.

Repeated-first-frame input removes visible conformational change while preserving the initial appearance. Shuffled input retains the first frame and frame set but changes temporal order. These controls assess the contribution of generated-video motion and ordering, whereas the real-video oracle tests the decoder with faithful reference motion. Structural-repair results are retained as a separate paired analysis in Appendix G.2.

Pairwise-RMSD correlation summarizes agreement in motion scale across targets. Coordinate-space MD-PCA and Joint-PCA W_2 are distinct from the visual-feature metrics in Appendix E; reference-fitted MD-PCA and the separately defined Joint-PCA diagnostic should not be conflated. RMSD-curve and autocorrelation-curve MAEs compare lag-dependent summaries rather than pointwise agreement with a uniquely determined MD future. The real-video oracle performs best across the five metrics; original generated videos remain below this decoder reference but outperform both controls.

G.2 Structural Refinement and All-Atom Completion

Restrained OpenMM refinement improves the local geometry of decoded C_α trajectories, while optional cg2all backmapping provides all-atom structures. Table 7 reports a paired refinement analysis conducted separately from the input-ablation experiment in Table 2.

Refinement increases C_α validity from 26.815% to 97.443%, while reducing chain breaks and bond-length errors. Subsequent cg2all completion achieves a Ramachandran favored rate of 96.733%. These results support the use of refinement and backmapping to improve the structural quality of decoded outputs; they characterize the processed structures rather than raw video predictions.

G.3 Pipeline Runtime and Stage-Wise Accounting

Table 8 reports the complete stage-wise timing breakdown, including the illustrative production budget. It distinguishes generator time from coordinate processing and optional backmapping. The video-to-refined-C_α interval includes the OpenMM interval: adding both would double-count refinement. The coordinate-processing cohort contains 82 proteins, whereas the backmapping timing cohort contains 79; these cohorts must not be treated as identical.

Table 8: Pipeline costs for 49-frame videos and recovered 49-conformer trajectories. All times are in seconds. Cohort-normalized processing costs and the backmapping median are distinct statistics; the illustrative total is a stage-wise sum, not measured end-to-end latency.

Stage	Cohort wall time	Per-target statistic
VideoMD generation (49 frames)	Manuscript timing	635–677
Video to refined C_α	571.978 / 82 proteins	6.975 (normalized)
OpenMM (included above)	245.092 / 82 proteins	2.989 (normalized)
Optional cg2all completion	3,968.600 / 79 proteins	36.781 (median)
Illustrative production total	Stage-wise sum	679–721

An illustrative stage-wise budget using the backmapping median is

$$[635, 677] + 6.975 + 36.781 = [678.756, 720.756]\ \text{s}, \tag{13}$$

or approximately 679–721 s. This mixes a generation range, a cohort-normalized recovery time, and a median completion time; it is not a measured end-to-end mean, median, or latency interval. The backmapping cohort-normalized cost is instead $3968.600/79 \approx 50.235$ s, which would give a different illustrative budget of approximately 692–734 s. Neither sum is a hardware- and cohort-matched throughput comparison. Local refinement is inexpensive in the reported accounting, but size-independent generation should not be extrapolated to the coordinate-processing stages.

G.4 Input-Video Controls and Interpretation

Table 2 compares the original generated video with repeated-first-frame and shuffled-input controls, and includes a real-video oracle decoded by the same readout. The interventions precede coordinate decoding, and input shuffling retains the first frame.

Real-video oracle. The oracle uses rendered reference MD videos rather than generated videos. Its stronger coordinate and dynamical metrics show that the decoder can recover informative trajectories when provided faithful molecular motion; the remaining gap between generated-video input and the oracle reflects the combined effects of video generation and recovery.

Visible motion and the static control. Repeating the first frame removes visible conformational variation while preserving the initial appearance and structural conditions. Differences from original-video input therefore characterize the contribution of video motion to the full structure-conditioned readout, not unconditioned recovery from pixels alone.

Input order and recovered distributions. Let P permute future frames while retaining frame 0, and let C denote the structural and camera conditions. Temporal tracking and decoding mean that $D_\psi(PV; C)$ need not equal $PD_\psi(V; C)$: the recovered conformations can change despite an unchanged image set. Thus this intervention diagnoses the full readout, rather than isolating tracking, temporal priors, or reordering a fixed coordinate ensemble.

H Dataset Details

ATLAS. ATLAS simulations were performed with GROMACS v2019.6 and the CHARMM36m force field. Proteins were solvated in explicit TIP3P water, neutralized with ions, energy-minimized, equilibrated in NVT and NPT stages, and simulated at 300 K and 1 bar. The dataset contains triplicated 100 ns trajectories for proteins with diverse structures and dynamics (Vander Meersche et al., 2024). We follow the PDB-release-date split used by prior work, where proteins released before May 1, 2018 are used for training and proteins released after May 1, 2019 are used for testing (Berman et al., 2000; Jing et al., 2024a;b; Shen et al., 2025).

CATH2. We additionally use a CATH-derived dataset from BioEMU (Lewis et al., 2025), denoted as CATH2, to study conditioning on multiple simulation protocols. This dataset contains 871 proteins, each with three replicate 1000 ns MD simulations. We follow a time-based split by PDB release date, with an approximate ratio of 8:1:1: training structures range from 1900-01-01 to 2014-01-08, validation structures from 2014-01-22 to 2015-10-28, and test structures from 2015-10-28 to 2019-05-29. Simulations are run with OpenMM using the AMBER ff99SB-ILDN force field, with explicit solvent and the TIP3P water model. Structures are solvated in a cubic box with 1 nm padding and 0.1 M NaCl. Simulations are conducted at 300 K and 1 bar. These results assess the mixed-data model under protocol conditioning, rather than establishing zero-shot transfer from ATLAS to CATH2.

Domain Motion. To examine protocol-conditioned large-scale conformational changes, we use the BioEMU domain-motion set (Lewis et al., 2025), which contains 22 examples of large-scale hinge motions. Each example is defined by a pair of experimentally observed conformations: `7UBJ_A/7UBL_A`, `6PWJ_A/6PWK_A`, `6PQN_A/6PQX_A`, `6JQ4_A/6KML_B`, `7E8R_A/7EDB_A`, `6ON4_A/6WFQ_A`, `1M74_A/1TF2_A`, `3ZSF_A/2YLN_A`, `6HNK_A/6HNI_A`, `3FTO_A/3DRF_A`, `6MKA_A/6MKJ_A`, `1URP_A/2DRI_A`, `1K23_C/1WPM_B`, `1BPD_A/2BPG_E`, `6ML0_A/6MLP_A`, `1AKE_A/4AKE_A`, `1DV1_A/1DV2_A`, `1WDN_A/1GGG_A`, `1DPE_A/1DPP_A`, `1CLL_A/5GGM_A`, `3O80_A/3O8M_A`, and `2WFA_A/2WF5_A`. We use this set only for supplementary testing of prompt-conditioned motion generation.

I Baseline details

This section provides additional implementation details for the baselines used in our experiments, including the MD oracle, MDGen (Jing et al., 2024b), AlphaFolding (Cheng et al., 2025), PVB (Yu et al., 2026), and ConfRover (Shen et al., 2025). For the MoViE benchmark, coordinate-based baselines are rendered and scored in video space. Their native sampling settings use 50-frame trajectories where supported, whereas VideoMD uses 49-frame clips. The recorded baseline hardware is an NVIDIA A800 with 80 GB memory, whereas VideoMD sampling lists one H800 (Table 10); runtime ratios therefore describe these configurations rather than hardware-matched throughput.

MD oracle. For the 100 ns setting, we use one of the triplicated trajectories in the ATLAS dataset as the oracle trajectory to estimate the expected performance of an independent simulation run.

MDGen. For MDGen (Jing et al., 2024b), we follow the official data preparation and inference procedures and use the released checkpoint `atlas.ckpt`. The model is trained on ATLAS trajectories preprocessed at 400 ps intervals (i.e., stride 40 from the base 10 ps interval). During inference, we use the official forward-simulation script `sim_inference.py` with `num_frames=250`, `num_rollouts=1`, and `suffix=_R1`. The generated trajectories are post-processed for the video-based evaluation; the native generation length is specified above.

AlphaFolding. For AlphaFolding (Cheng et al., 2025), we follow the official data preparation and inference procedures and use the recommended checkpoint `frame32_step_95000.pth`. Trajectories are generated using the recommended extrapolation script `run_eval_extrapolation.sh` with `sample_step=1`, `n_motion=2`, and `n_frames=32`. The generated trajectories are post-processed for the video-based evaluation; the native generation length is specified above.

Table 9: Hyperparameter and training settings.

	VideoMD	Validation	RMSD
Model details			
Backbone	Wan2.1-I2V-14B-720P	ViT-L/14	ViT-L/14
Input size	$1280 \times 720 \times 49$	224×224	224×224
LoRA rank r	32	32	32
LoRA α	—	32	32
LoRA dropout	—	0.1	0.1
Head hidden dim	—	512	512
Head dropout	—	0.1	0.1
Training			
Learning rate	10^{-4}	10^{-4}	10^{-4}
Weight decay	10^{-2}	10^{-4}	10^{-4}
Batch size	1	64	32
Gradient accumulation	1	—	—
Epochs	5	50	50
Random seed	—	42	42
Data-loader workers	0	8	8
LR schedule	—	Plateau	Plateau
Plateau factor	—	0.5	0.5
Plateau patience	—	5	5
Checkpoint interval (epochs)	1	10	10
Hardware			
GPU device	6× A100	1× H800	1× H800

PVB. For PVB (Yu et al., 2026), we retrain the model on the ATLAS dataset using the same data split as in our experiments. We follow the original model design and training protocol, while adapting data preprocessing and trajectory formatting to our benchmark setup. During inference, we use `inf_step=1000`. The generated trajectories are post-processed for the video-based evaluation; the native generation length is specified above.

ConfRover. For ConfRover (Shen et al., 2025), we use the official codebase and released model weights. We use the forward-simulation mode to generate trajectories, with `n_frames=50` and `stride_in_10ps=200`.

J Implementation Details

Table 9 summarizes the main model configurations and training hyperparameters used in our experiments. For VideoMD, we fine-tune Wan2.1-I2V-14B-720P on molecular video trajectories at a spatial resolution of 1280×720 with 49 frames per clip. Unless otherwise specified, VideoMD is trained for 5 epochs with a learning rate of 10^{-4}, weight decay of 10^{-2}, and batch size 1 on 6×A100 GPUs. For the validation and RMSD prediction models, we use ViT-L/14 backbones with LoRA adaptation (rank 32) and a two-layer prediction head of hidden dimension 512. Both models are trained for 50 epochs using a learning rate of 10^{-4} and the ReduceLROnPlateau scheduler. Additional optimization details, including LoRA settings, checkpoint intervals, and hardware configurations, are reported in Table 9.

For trajectory generation, we use the sampling configuration listed in Table 10. Unless otherwise noted, VideoMD samples trajectories at 1280×720 resolution with 49 frames, 20 denoising steps, and guidance scale 5.0. We use an empty negative prompt, fp16 VAE precision, and the `TORCH_SDPA` attention backend, without CPU offloading for the DiT, text encoder, image encoder, or VAE.

Table 10: VideoMD trajectory sampling

VideoMD sampling	
Spatial size	1280×720
Clip length (frames)	49 $(4n{+}1)$
Denoising steps	20
Guidance scale	5.0
RNG seed	0
Negative prompt	" "
VAE precision	`fp16`
Attention backend	`TORCH_SDPA`
DiT CPU offload	off
Text encoder CPU offload	off
Image encoder CPU offload	off
VAE CPU offload	off (default)
GPU device	$1\times$ H800

Evaluation coverage. A qualitative no-LoRA control is reported below. A quantitative, noise-paired generator-adaptation ablation is not included in the present measurements; the main benchmark therefore describes the adapted model under the recorded configuration without isolating the contribution of each adaptation component.

J.1 Qualitative No-LoRA Adaptation Control

We run the official Wan2.1-I2V-14B-720P backbone with LoRA disabled on the ATLAS evaluation set. Generation uses the same 1280×720 spatial resolution, 49-frame clip length, 20 denoising steps, guidance scale 5.0, and FlowUniPC multistep scheduler used by VideoMD. Figure 7 shows a representative example. Without molecular-domain adaptation, the sequence rapidly departs from the standardized white rendering and develops pronounced pose, scale, and appearance drift. The LoRA-adapted VideoMD sequence remains more consistently anchored to the initial molecular representation. Because the archived adapted sample and the no-LoRA sample are not noise-paired, this comparison is qualitative and does not provide a matched-sample estimate of the LoRA effect.

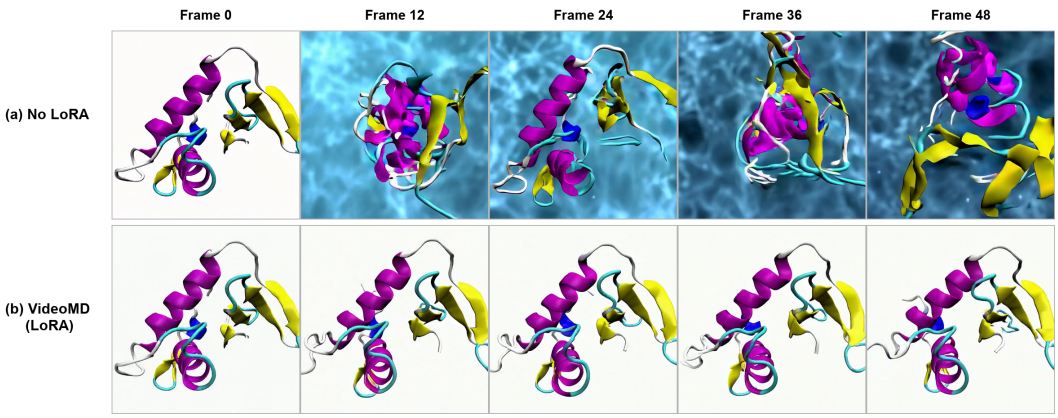


Figure 7: Qualitative domain-adaptation control for PDB ID 6d7y_A. The unadapted Wan backbone (top) develops background, pose, scale, and appearance drift after the conditioning frame. LoRA-adapted VideoMD (bottom) retains the standardized molecular-video appearance and more stable spatial framing. Frames 0, 12, 24, 36, and 48 are shown. The runs are not noise-paired, so this figure illustrates a representative failure mode rather than a quantitative matched-sample comparison.

K MoViE Details and Validation

K.1 Predictor Architecture and Training

Validation predictor. For a single rendered frame x_i, the visual encoder $f_{\phi_{\text{val}}}$ produces an image representation $f_{\phi_{\text{val}}}(x_i)$, which is passed to a quality head $u_{\omega_{\text{val}}}$ to predict three structure-level invalidity ratios, $\widehat{\mathbf{q}}_i = u_{\omega_{\text{val}}}(f_{\phi_{\text{val}}}(x_i))$, corresponding to residue clashes, chain breaks, and Ramachandran outliers. The predictor is trained with mean-squared error against the corresponding targets $\mathbf{q}_i$ computed from the underlying structure, i.e., $\mathcal{L}_{\text{val}} = \mathbb{E}_i \|\widehat{\mathbf{q}}_i - \mathbf{q}_i\|_2^2$.

Corruption-based supervision. To train the validation predictor over a broad range of structural quality, we augment valid structures with synthetic corruptions. Starting from a clean structure s, we perturb residue frames with translational noise and IGSO(3) rotational noise at varying severity levels η, yielding corrupted samples $\tilde{s} \sim \mathcal{C}_\eta(s)$. We then compute the corresponding invalidity targets $\mathbf{q}(\tilde{s})$ from the corrupted geometry, providing continuous supervision over structural degradation rather than a binary label; additional validation details are given in Appendix K.

RMSD predictor. Given rendered frames x_i and x_j, a CLIP-based visual encoder f_{ϕ_r} extracts frame representations, which are combined by a fusion head g_{ψ_r} to predict the aligned structural RMSD, $\widehat{r}_{ij} = g_{\psi_r}([f_{\phi_r}(x_i), f_{\phi_r}(x_j), |f_{\phi_r}(x_i) - f_{\phi_r}(x_j)|])$. The model is trained with mean-squared error, $\mathcal{L}_{\text{RMSD}} = \mathbb{E}_{(i,j)}[(\widehat{r}_{ij} - r_{ij})^2]$, where r_{ij} is the RMSD computed from the underlying aligned structures. This objective encourages f_{ϕ_r} to learn structure-aware visual representations for trajectory quality assessment.

Predictor error and cross-view statistics are presented in Appendix K.3. The definitions and controlled-corruption tests below document the evaluated settings.

K.2 Metric Definitions

For a fixed evaluation view, let y_i be the ground-truth RMSD and $\hat{y}_i$ be the predicted RMSD for test pair i. We report

$$\text{MSE} = \frac{1}{N}\sum_i (\hat{y}_i - y_i)^2, \quad \text{MAE} = \frac{1}{N}\sum_i |\hat{y}_i - y_i|, \quad \text{RMSE} = \sqrt{\text{MSE}},$$

For view sensitivity, each pair is rendered from six views $v \in \{\text{front}, \text{back}, \text{left}, \text{right}, \text{top}, \text{bottom}\}$. For sample i, we compute

$$\sigma_i = \text{std}_v(\hat{y}_{i,v}), \qquad r_i = \max_v \hat{y}_{i,v} - \min_v \hat{y}_{i,v}, \qquad e_i^{\text{view}} = \frac{1}{6}\sum_v |\hat{y}_{i,v} - y_i|.$$

K.3 Validation of MoViE Predictors

Table 11 summarizes prediction error and cross-view robustness on the evaluated rendered validation data. The structural-validity predictors exhibit low cross-view variation, especially for the clash-residue and break-pair ratios, indicating that these signals are stable under changes in rendering view. By contrast, RMSD prediction is more view-sensitive, which is consistent with global conformational deviation depending more strongly on motion patterns visible in a particular projection. We additionally observe that the structural-quality signals remain useful under thresholded classification, with details in Appendix K.4.

K.4 Structural-Quality Predictor Validation

Synthetic-noise validation data. To calibrate the structural-quality predictor across a controlled range of degradation levels, we use clean ATLAS structures and RFdiffusion-style residue-frame perturbations. The supplementary validation set contains five example proteins evaluated at seven noise levels: steps 0, 1, 2, 3, 5, 10, and 20, corresponding to noise levels 0, 0.002, 0.004, 0.006, 0.010, 0.020, and 0.040. For each rendered structure, the validation table records the system identifier, view, noise level, PDB identifier, per-target pass rates, and the target ratios used by the validation model:

Table 11: Prediction error and cross-view robustness of MoViE-based predictors. Lower is better for MAE, RMSE, and robustness statistics. For each sample, we compute the standard deviation and range across rendering views, and report the mean and 95th percentile. The evaluated data have coordinate-derived targets; RMSD quantities are in Å, and invalidity ratios are unitless.

Predictor	Target	MAE	RMSE
RMSD predictor	RMSD	0.478	0.713
Validation predictor	Clash residue ratio	0.0111	0.0280
Validation predictor	Break pair ratio	0.0103	0.0160
Validation predictor	Invalid Rama ratio	0.0402	0.0557

	View standard deviation		View range	
Target	Mean	P95	Mean	P95
RMSD	0.4250	1.114	1.234	3.131
Clash residue ratio	0.0046	0.0245	0.0133	0.0691
Break pair ratio	0.0039	0.0099	0.0112	0.0282
Invalid Rama ratio	0.0188	0.0421	0.0543	0.1224

clash residue ratio, break-pair ratio, and invalid Ramachandran ratio. Figure 8 shows the visual examples for steps 0–10, and Figure 9 summarizes the corresponding pass-rate trend.

Table 12: Summary of the synthetic-noise validation set. Local-quality percentages are averaged over five example systems and are distinct from the all-three thresholded criterion. The all-three column reports the fraction of examples satisfying all three validation thresholds: $q_{\mathrm{clash}} < 0.002$, $q_{\mathrm{break}} < 0.005$, and $q_{\mathrm{rama}} < 0.051471$.

Step	Noise level	Clash pass (%)	Break pass (%)	Rama pass (%)	Invalid Rama ratio mean	All-three reasonable (%)
0	0.000	100.00	100.00	99.52	0.0048	100
1	0.002	100.00	100.00	97.49	0.0251	80
2	0.004	100.00	99.53	94.45	0.0555	40
3	0.006	100.00	98.49	89.65	0.1035	0
5	0.010	100.00	95.59	84.30	0.1570	0
10	0.020	97.93	93.11	64.45	0.3555	0
20	0.040	74.09	92.70	51.82	0.4818	0

The local-quality percentages in Table 12 should be distinguished from its all-three, example-level threshold criterion. For instance, the listed Rama percentage equals $100(1-q_{\mathrm{rama}})$ for the displayed mean invalidity ratio; it is not the fraction of the five examples passing the Rama threshold.

For threshold classification, the positive class is "reasonable." For a single target k, a sample is GT reasonable if $q_k < \tau_k$ and predicted reasonable if $\hat{q}_k < \tau_k$. For the all-three metric, both GT and prediction must satisfy all three thresholds. Accuracy, precision, recall, F1, specificity, balanced accuracy, Matthews correlation coefficient (MCC), and ROC-AUC are reported with "reasonable" as the positive class.

Table 13: Structural-quality predictor threshold-classification metrics. Positive class is "reasonable."

Target	Accuracy	Precision	Recall	F1	Specificity	Balanced acc.	MCC	AUC
Clash	0.944	0.989	0.937	0.962	0.966	0.951	0.859	0.984
Break	0.881	0.862	0.892	0.877	0.872	0.882	0.763	0.952
Rama	0.849	0.744	0.846	0.792	0.850	0.848	0.678	0.934
All three	0.841	0.711	0.853	0.776	0.835	0.844	0.661	0.909

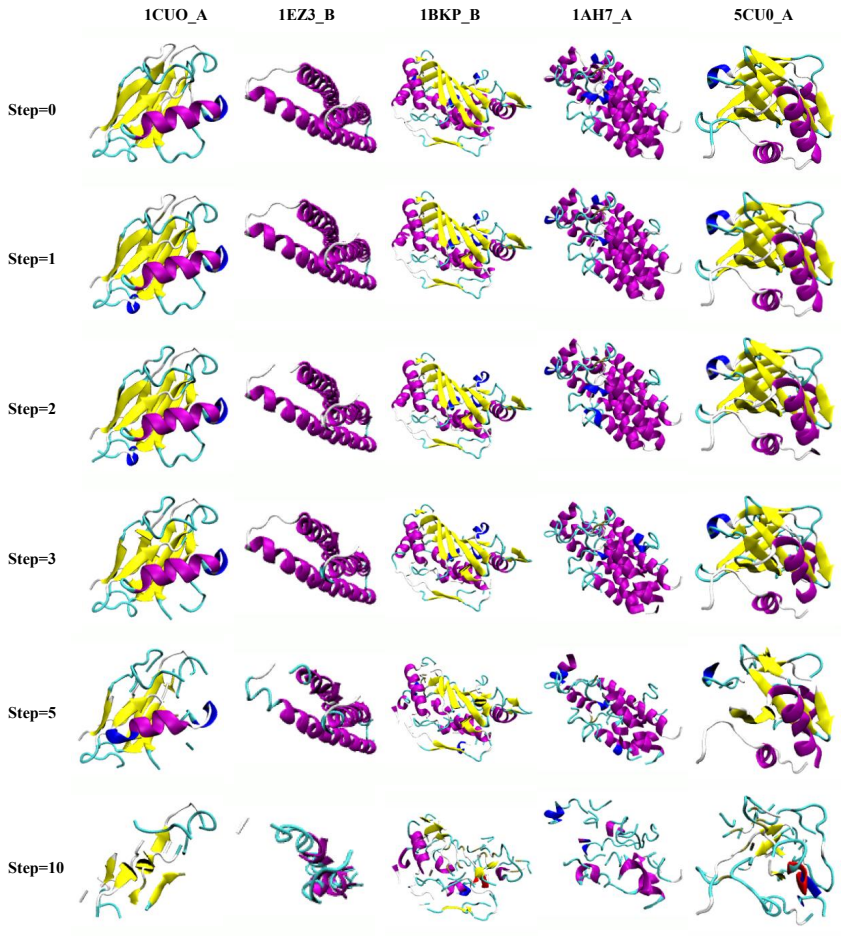


Figure 8: Synthetic corruption examples used to illustrate validation-model supervision. Columns are example ATLAS proteins and rows are increasing perturbation steps. The step-10 row shows severe disruptions in global shape and local geometry, consistent with the lower pass rates in Table 12.

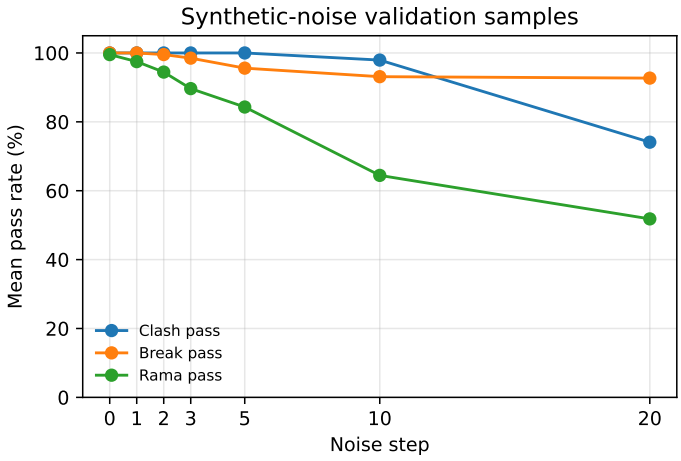


Figure 9: Local-quality percentages across five synthetic-noise examples, using the original plotted values. The Ramachandran percentage decreases fastest with noise. These percentages are not the example-level, all-three thresholded pass rate (Table 12).

L ADDITIONAL RESULTS

The following tables give the detailed temporal, runtime, and length-stratified measurements underlying the main video-space benchmark. These MoViE comparisons retain their native sampling settings and success cohorts. Additional visual examples complement Section 4.3.

The complete MoViE benchmark is reported in Table 1 in Section 4.2.

At 650 residues, the recorded times are 669.8 s for VideoMD and 2516.6 s for ConfRover; ConfRover runs out of memory on the 724-residue example. These are configuration-specific measurements on different hardware. VideoMD's visual Joint PCA W_2 values are 10.131, 10.033, and 9.663 across the three length bins. These measurements characterize the fixed-size visual generator, not the full coordinate-recovery pipeline.

Table 14: Inference times on representative systems of increasing sequence length, used to plot the runtime scaling curves in Figure 3(c). Values are measured in seconds; lower is better. The original values are transposed for readability; hardware settings differ between implementations.

System	Residues	ConfRover	PVB	VideoMD
6okd_C	51	273.2	128.4	635.0
6l4p_B	145	309.3	258.6	677.0
6xb3_H	241	507.6	402.6	676.2
7aqx_A	364	889.9	550.4	674.0
7p41_D	448	1251.8	707.2	673.8
7ned_A	546	1807.9	843.8	668.9
6l8s_A	650	2516.6	1046.7	669.8
6sms_A	724	OOM	1136.9	672.0

Table 15: Detailed lag values used to draw Figure 3. Each MAE averages absolute errors over the four lags relative to MD1. The two panels retain the same reported values while separating the feature-displacement and autocorrelation readouts.

	Latent feature RMSD				
Model	Lag 1	Lag 5	Lag 10	Lag 20	MAE
MD1	4.4780	5.1380	5.4850	5.9270	–
MD2	4.4770	5.1500	5.5070	5.8470	0.0288
MDGen	4.084869	4.356147	4.464480	4.589556	0.8832
ConfRover	4.438518	5.082976	5.252969	5.542511	0.1778
PVB	3.210845	4.564502	5.469905	6.777428	0.6765
VideoMD	4.0476	4.9330	5.4000	5.906414	0.1852

	Latent autocorrelation				
Model	Lag 1	Lag 5	Lag 10	Lag 20	MAE
MD1	0.306400	0.115600	0.014800	-0.096600	–
MD2	0.314139	0.113740	0.003981	-0.088788	0.0071
MDGen	0.163163	0.049445	-0.006486	-0.062583	0.0662
ConfRover	0.193722	0.031430	0.002945	-0.050078	0.0638
PVB	0.627547	0.291103	0.073730	-0.187538	0.1616
VideoMD	0.336400	0.104800	-0.010100	-0.095370	0.0167

Table 16: Joint PCA W_2 by sequence length bin, used to plot the distributional-fidelity curves in Figure 3(c). Lower values indicate better agreement with the MD reference.

Method	$L \leq 200$	$200 < L \leq 400$	$L > 400$
ConfRover	11.248	11.638	12.238
VideoMD	10.131	10.033	9.663
PVB	15.903	17.910	17.818

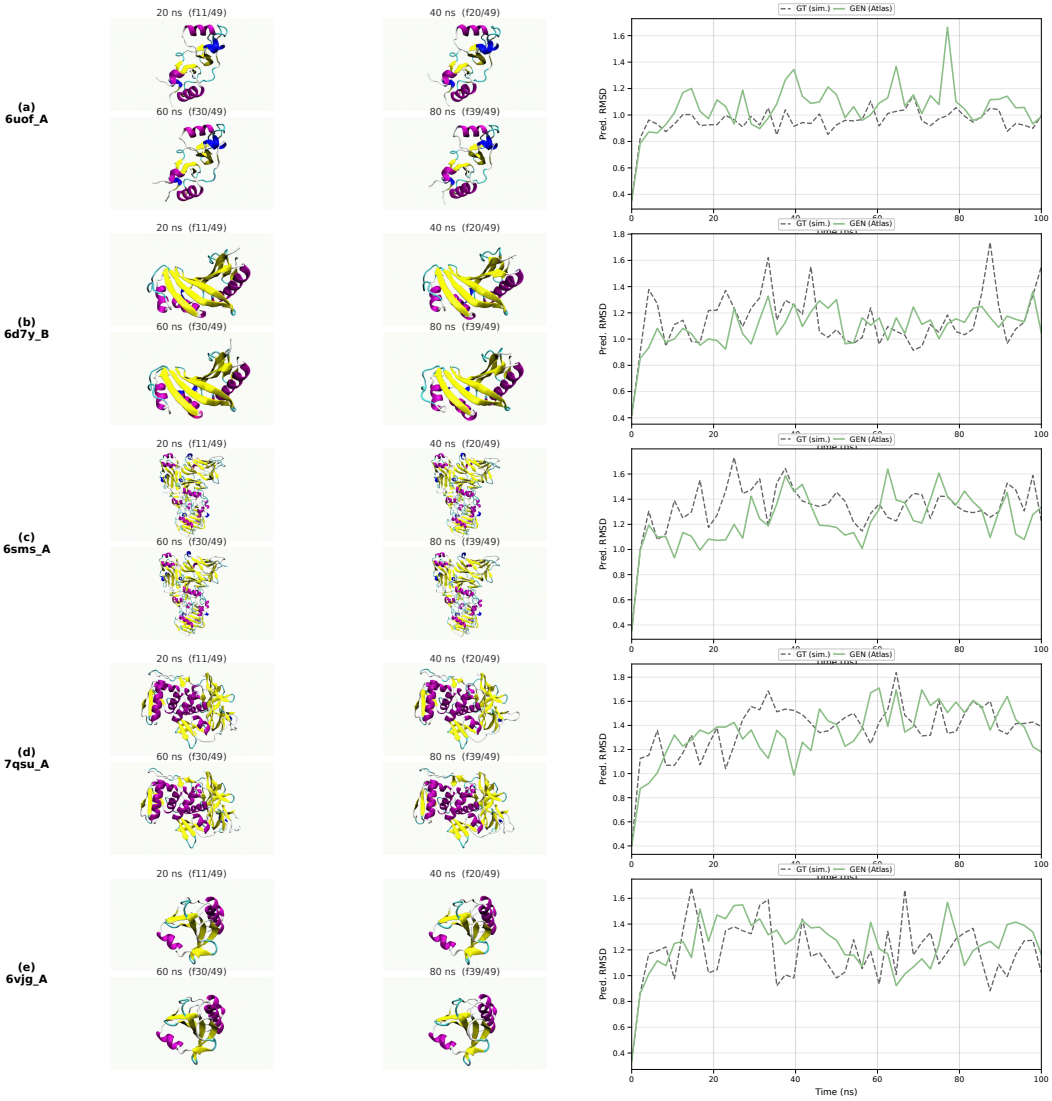


Figure 10: Additional qualitative case studies.