PREFIXMOTION: PREFIX-CONSISTENT CONDITIONAL MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive conditional motion generation requires new observations to extend a motion sequence without revising its history. We introduce PrefixMotion, a framework that adapts pretrained motion generators to prefix-consistent streaming inference. Under compatible conditions and coupled randomness, full-sequence and prefix executions agree on their shared outputs. The framework enforces this agreement across attention, shape aggregation, and motion decoding. A clean-motion parameterization of rectified flow preserves compatibility with pretrained weights and geometric supervision while enabling three-evaluation Euler sampling. Instantiated with GENMO, PrefixMotion is evaluated through video-conditioned reconstruction and text-conditioned synthesis. On 120-frame windows from 37 3DPW tracks, four prefix-length comparisons yield maximum tensor deviations below 10^{-5} through the global output, given compatible supplied track and camera prefixes. Cached execution is evaluated on 79 complete sequences and accelerates newest-frame updates by 10.6$\times$ over window recomputation on the same GPU. PrefixMotion improves all three 3DPW spatial metrics and all three EMDB-2 trajectory metrics over our evaluation of pretrained GENMO. It also achieves lower error on 10 of 13 metrics than OnlineHMR under shared finalized tracks and DROID-SLAM cameras, with model-specific EMDB intrinsics. On HumanML3D, five-trial mean FID improves from 10.169 to 8.110 and R@3 from 0.475 to 0.492. These results show that explicit temporal constraints can support efficient streaming while preserving conditional generation and improving selected reconstruction metrics.

1 INTRODUCTION

Conditional motion generation maps a control signal to a plausible human motion sequence. A text prompt specifies what a person should do; video observations constrain the motion to explain a recorded scene. These tasks share a generative representation but differ in conditions and evaluation criteria. Interactive use adds a requirement: when the requested duration grows or new observations arrive, motion already generated should not change. Adapting a pretrained generator to this zero-lookahead setting should also preserve its conditional generation quality.

We call this requirement *prefix consistency*: with identical conditions and coupled noise on a shared prefix, generating a full sequence or only its prefix must yield the same motion on that prefix. Global conditions available before generation, such as a text prompt, are held fixed. A causal attention mask is not enough: full-sequence shape statistics, camera features that look one frame ahead, and postprocessing can still alter past outputs. Every operator that reads across frames must respect the same boundary.

We introduce PrefixMotion, a framework for adapting pretrained conditional motion generators to prefix-consistent streaming inference. Its shared clean-motion predictor combines bounded lower-triangular attention with inclusive running shape aggregation. A weight-compatible x_1 rectified-flow parameterization enables a three-evaluation Euler solve. Training uses per-frame corruption times and randomized clean prefixes to expose the model to mixed-quality histories; we evaluate these choices separately from the structural guarantee. For video conditioning, backward-aligned camera features and causal world-space decoding preserve the prefix contract through the reconstructed body output. The guarantee assumes compatible supplied person-track/bounding-box and camera prefixes.

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.

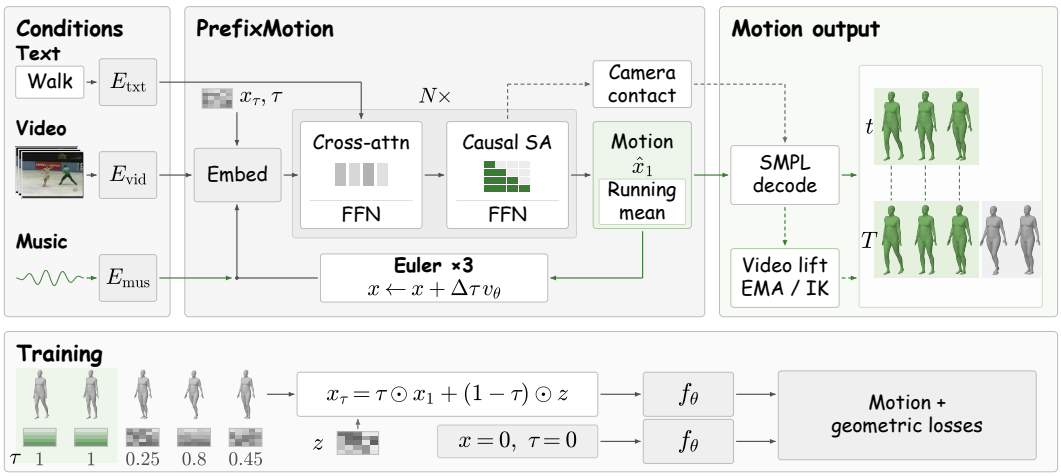

Figure 1: Overview of PrefixMotion. Alternative text, video, or music conditions feed a shared clean-motion predictor. Repeated cross-attention/FFN and causal self-attention (SA)/FFN blocks process the motion state; running shape aggregation preserves prefix-compatible body parameters. Three Euler evaluations use this predictor to construct v_θ. Dashed arrows show the video-specific camera/contact and world-lift path with a root-velocity exponential moving average (EMA) and inverse kinematics (IK). The output rows illustrate separate executions at lengths $t < T$: compatible conditions and coupled noise preserve the first t outputs. Below, independent frame times with an optional clean prefix and a separate zero-state regression pass provide the training inputs. Meshes and latent tiles are schematic; conditioning modes are evaluated separately.

Our implementation uses the pretrained GENMO backbone (Li et al., 2025), retaining its condition encoders, motion representation, and geometric supervision. We evaluate whether imposing prefix consistency preserves generation under both visual and semantic conditions. Video-conditioned generation is assessed as reconstruction on 3DPW and EMDB, including a component-wise consistency analysis and a comparison with OnlineHMR (Zhao et al., 2026) that shares finalized tracks and DROID-SLAM cameras but keeps each model's EMDB intrinsics. Text-conditioned generation is assessed on HumanML3D with fixed captions and a shared variable-length protocol.

Prefix comparisons and a camera-causality check establish why running shape and backward camera alignment are required in addition to causal attention. PrefixMotion improves all three 3DPW spatial and all three EMDB-2 trajectory metrics over our evaluation of pretrained GENMO, and 10 of 13 reconstruction metrics in the stated OnlineHMR comparison. On HumanML3D, it achieves better mean R@3, FID, and matching distance under the shared protocol. Controlled sampling studies distinguish the effects of learned weights from those of the numerical solver. Finally, cached execution reuses past computation while preserving the generated history.

Our contributions are:

- PrefixMotion, a framework that adds prefix-consistent streaming inference to a pretrained conditional motion model, with a composition theorem covering the predictor, shape statistics, and task-specific decoding;
- a pretrained-weight-compatible rectified-flow implementation, an endpoint-safe Euler sampler with three network function evaluations (NFE-3), and geometric supervision; and
- an evaluation of video- and text-conditioned generation with gains on selected metrics, supported by prefix comparisons on 37 tracks, cached streaming on 79 complete sequences, and controlled training/sampling studies.

2 Related Work

Conditional motion generation. Text-to-motion models learn to synthesize human movement from semantic conditions. MDM (Tevet et al., 2023) predicts clean motion within diffusion, allowing

geometric losses on the generated sample. T2M-GPT (Zhang et al., 2023) generates discrete motion codes with an autoregressive transformer. GENMO (Li et al., 2025) unifies multiple conditions, including text, audio, and visual observations, in a generalist motion model. PrefixMotion uses a continuous clean-motion representation and studies agreement between executions at different sequence lengths, including the effects of shape aggregation and downstream decoding. Our implementation reuses GENMO's pretrained backbone; our target is the prefix contract of the complete generator.

Causal generative sequence models. Vid2World (Huang et al., 2026) causalizes pretrained video diffusion models through temporal modules, per-frame diffusion times (Chen et al., 2024), and autoregressive rollout with clean history. PrefixMotion adopts causal attention and per-frame corruption, while its main sampling path jointly integrates the motion prefix. Flow matching (FM) (Lipman et al., 2023) and rectified flow (Liu et al., 2023) provide continuous transport formulations. A clean-sample predictor directly parameterizes an x_1 field, permitting PrefixMotion to retain pretrained prediction weights and geometric losses. We distinguish the operator-level prefix guarantee from empirical effects of corruption design and the chosen numerical solver.

Video-conditioned motion reconstruction. Video HMR is a conditional setting in which visual observations constrain motion. VIBE (Kocabas et al., 2020) uses a temporal motion discriminator, while TCMR (Choi et al., 2021) balances static and temporal evidence. GLAMR (Yuan et al., 2022), SLAHMR (Ye et al., 2023), and WHAM (Shin et al., 2024) jointly reason about body and camera motion. DuoMo (Wang et al., 2026) factorizes camera- and world-space diffusion for offline reconstruction. EMDB (Kaufmann et al., 2023) evaluates global trajectory and jitter. These benchmarks test the geometric and temporal quality of PrefixMotion's generation under visual evidence.

Online recovery and visual body cues. OnlineHMR (Zhao et al., 2026) combines a causal key–value cache with incremental human-centric SLAM. We compare its released body estimator under shared finalized tracks and DROID support, while retaining model-specific EMDB intrinsics. SAM 3D Body (Yang et al., 2026) supplies promptable single-image body cues, and Multi-HMR 2 (Fiche et al., 2026) adds camera-centric localization and tracking. Such upstream systems affect the conditions available to a generator. PrefixMotion's guarantee is conditional on compatible supplied track and camera prefixes; it does not replace incremental tracking or camera-state management.

3 METHOD

3.1 CONDITIONAL GENERATION AND THE PREFIX CONTRACT

PrefixMotion generates normalized parametric motion $m_{1:T} \in \mathbb{R}^{T\times 151}$ from conditions and Gaussian noise. Figure 1 outlines the pipeline. The SMPL-X state (Pavlakos et al., 2019) contains body pose, shape, root orientations, and local root velocity. Write the conditions as $c = (g, u_{1:T})$, where g is available before generation, such as a text prompt, and $u_{1:T}$ contains any frame-aligned inputs. For video conditioning these include visual, supplied person-track/bounding-box, and camera features; text-only generation uses the fixed prompt and the corresponding non-visual condition setting. Let $c|_t = (g, u_{1:t})$ retain g and truncate frame-aligned inputs. A conditional generator F, including motion decoding, is prefix-consistent if, under coupled random tensors,

$$F(c; z_{1:T})_{1:t} = F(c|_t; z_{1:t}) \quad \text{for all } t \le T. \tag{1}$$

Video-conditioned reconstruction adds auxiliary camera/contact heads and world-space decoding to the same generator. We evaluate this complete mapping, conditional on compatible supplied track and camera-trajectory prefixes. Each temporal attention layer uses a backward band of $C = 120$ frames; running shape and causal postprocessing maintain their own prefix statistics. No operator may read frames after the one it outputs.

3.2 THE PREFIXMOTION GENERATOR

Given noisy motion $x \in \mathbb{R}^{T\times 151}$, framewise times $\boldsymbol{\tau} \in [0,1]^T$, and conditions c, the predictor $\hat{x}_1 = f_\theta(x, \boldsymbol{\tau}, c)$ estimates clean motion. We use GENMO's pretrained RoPE (Su et al., 2024) transformer (Li et al., 2025), retaining its condition encoders, output heads, and motion representation.

Let $a_t \in \mathbb{R}^D$ denote the encoded frame-aligned conditions and $G = E_g(g)$ the fixed text memory when present. The input token is $h_t^{(0)} = W_{\text{in}}[a_t + e_\tau(\tau_t); x_t] + b_{\text{in}}$, where e_τ embeds time and $[\,;\,]$ denotes concatenation. The transformer combines cross-attention to G with temporal self-attention. For each temporal head of width d, let Q, K be the RoPE-rotated queries and keys and V the values. We define

$$\text{Attn}_C(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + A_C\right)V, \qquad (A_C)_{ts} = \begin{cases} 0, & 0 \le t-s < C, \\ -\infty, & \text{otherwise}, \end{cases} \tag{2}$$

with row-wise softmax and $C = 120$, including the current frame. Padding keys are excluded. Layer normalization, feed-forward blocks, and output heads operate per frame, while text cross-attention reads only the fixed G.

After L layers, write the motion head as $(q_t, \beta_t) = D_\theta(h_t^{(L)})$, separating SMPL-X shape β_t from the remaining channels q_t. Full-clip shape averaging would couple every output to future tokens. We instead define the predictor output by

$$\hat{x}_{1,t} = (q_t, \bar{\beta}_t), \qquad \bar{\beta}_t = \frac{1}{t}\sum_{s=1}^{t}\beta_s = \bar{\beta}_{t-1} + \frac{\beta_t - \bar{\beta}_{t-1}}{t}, \quad \bar{\beta}_0 = 0. \tag{3}$$

Thus shape is aggregated within each predictor call, before flow integration and body decoding. For prefix-compatible conditions, every output depends only on the corresponding motion and condition prefixes: writing P_t for prefix restriction gives $P_t f_\theta(x, \boldsymbol{\tau}, c) = f_\theta(P_t x, P_t \boldsymbol{\tau}, c|_t)$. Joint inference processes the full retained sequence under the same mask, without chunk resets or overlap blending.

3.3 Rectified Flow from a Clean-Motion Predictor

Given clean motion x_1, we sample valid-frame times $\tau_t \sim \text{Unif}[0,1]$ independently, then with probability 0.5 clamp a random history prefix to $\tau_t = 1$. Padding uses $\tau_t = 1$ with zero loss weight. Training windows are sampled uniformly from 16–120 frames. This exposes the model to mixed-quality histories; Appendix F evaluates this design and finalized-history rollout. Draw $z \sim \mathcal{N}(0, I)$ independently of $(x_1, c, \boldsymbol{\tau})$ and define the straight path, broadcasting time over motion features as $(\boldsymbol{\tau} \odot x)_{t,j} = \tau_t x_{t,j}$:

$$x_{\boldsymbol{\tau}} = \boldsymbol{\tau} \odot x_1 + (1 - \boldsymbol{\tau}) \odot z. \tag{4}$$

The field equations act coordinate-wise at τ_t during training; joint inference uses $\boldsymbol{\tau} = \tau \mathbf{1}_T$. With $\hat{x}_1 = f_\theta(x_\tau, \tau, c)$, the ideal field is

$$v_\theta^0(x_\tau, \tau, c) = \frac{\hat{x}_1 - x_\tau}{1 - \tau}, \qquad \tau < 1. \tag{5}$$

To regularize the denominator near the clean endpoint, we use

$$v_\theta(x_\tau, \tau, c) = \frac{\hat{x}_1 - x_\tau}{\max(1 - \tau, \epsilon)}. \tag{6}$$

We use $\epsilon = 10^{-3}$; if $\hat{x}_1 = x_1$ and $1 - \tau \ge \epsilon$, both fields equal $x_1 - z$. All pretrained tensors are retained. The time embedding e_τ uses the original sinusoidal map and MLP at fractional index $(1 - \tau)(N_{\text{diff}} - 1)$, where N_{diff} is the pretrained diffusion-step count, preserving its noise-to-clean ordering.

Each training update combines two predictions of x_1: an auxiliary regression prediction from an all-zero motion input, labeled $\tau = 0$, and a rectified-path prediction at sampled per-frame times. Predicting clean motion retains the pretrained supervision and permits geometric losses on the decoded body. Both predictions use masked clean-motion MSE and the following geometric objectives:

$$\mathcal{L} = \lambda_x \|M \odot (\hat{x}_1 - x_1)\|_2^2 + \lambda_{\text{3D}}\mathcal{L}_{\text{joint/vert}} + \lambda_{\text{2D}}\mathcal{L}_{\text{reproj}} + \lambda_a \mathcal{L}_{\text{accel}} + \mathcal{L}_{\text{transl/contact}}. \tag{7}$$

For the ideal field, and for the guarded field where $1 - \tau \ge \epsilon$, direct velocity MSE scales the clean-state error by $(1 - \tau)^{-2}$. It has the same coordinate-wise optimum on fully supervised coordinates, but different time weighting and finite-capacity optimization. The acceleration term matches second finite differences of root-relative joints to ground truth rather than penalizing nonzero acceleration itself. For 2D-only examples, the regression path supplies a pseudo-clean path endpoint without MSE weight; valid reprojection terms supervise them.

3.4 ENDPOINT-SAFE NFE-3 EULER SAMPLING

Starting from $x^{(0)} = z$, we use three equal Euler intervals. For $k \in \{0, 1, 2\}$,

$$x^{(k+1)} = x^{(k)} + \tfrac{1}{3} v_\theta\Big(x^{(k)}, \tfrac{k}{3}, c\Big). \tag{8}$$

The three field evaluations occur at $\tau \in \{0, 1/3, 2/3\}$, where the denominator guard is inactive. By Equation 17 in Appendix A, the last interval returns the clean prediction at $\tau = 2/3$ without evaluating the field at $\tau = 1$. Euler is selected on held-out 3DPW by comparing spatial and acceleration errors against NFE-3 Heun at equal latency (Appendix E).

3.5 VIDEO-CONDITIONED WORLD-SPACE DECODING

For video-conditioned reconstruction, the 151-D generated motion state is decoded into body pose, shape, camera-frame and gravity-view orientations, and local root velocity. A separate 3-D camera head, together with the frame bounding box and intrinsics, recovers camera-frame root translation. At inference, camera-velocity conditions encode $t-1 \to t$, initialized by identity/zero motion at the first frame. These quantities form the decoded camera-frame body parameters. Given supplied world-to-camera poses $T_t^{w\to c} \in \mathrm{SE}(3)$ acting on homogeneous column vectors, we lift root parameters to a sequence-global frame anchored at the first camera with

$$T^{c_t \to c_1} = T_1^{w\to c} \left(T_t^{w\to c}\right)^{-1}. \tag{9}$$

Its rotation premultiplies root orientation and the transform maps the SMPL translation, retaining local pose, shape, and within-frame body geometry. Camera-space metrics use the pre-lift prediction. Following benchmark terminology, we call the first-camera-anchored sequence frame "global" or "world"; it is distinct from an externally calibrated absolute frame.

For a one-frame sequence, the lifted root is unchanged. For $T \geq 2$, define the frame displacement $d_t = r_t - r_{t-1}$ for $t = 2, \ldots, T$; this is called root velocity in the implementation. The past-only EMA and reconstruction are

$$\begin{aligned} &\bar{d}_2 = d_2, \quad \bar{d}_t = \alpha \bar{d}_{t-1} + (1-\alpha) d_t, \quad t = 3, \ldots, T, \\ &\bar{r}_1 = r_1, \quad \bar{r}_t = r_1 + \sum_{s=2}^{t} \bar{d}_s, \qquad\qquad t = 2, \ldots, T, \end{aligned} \qquad \alpha = 0.25. \tag{10}$$

Global IK, applied before the EMA, reduces limb/contact inconsistency. Static-joint root-translation correction is disabled based on held-out 3DPW validation; symmetric orientation smoothing is replaced by the rigid camera lift.

3.6 FORMAL PROPERTIES

The central guarantee follows by composing prefix-compatible operators. The proof tracks the same noise realization through the denoiser, solver, body decoder, and causal postprocessing. Supporting clean-MSE, endpoint, and EMA properties are stated and proved in Appendix A.

Theorem 1 (Prefix consistency of PrefixMotion)**.** *Assume every frame-aligned condition constructor Q is prefix-compatible, $P_t Q(y_{1:T}) = Q(y_{1:t})$; the exogenous condition g is identical between executions; and random tensors are coupled on the compared prefix. Then the shared conditional predictor with banded lower-triangular attention, running shape aggregation, and the NFE-3 sampler produces prefix-consistent motion. For the video-conditioned path, compatible supplied person-track/bounding-box and camera-trajectory prefixes, together with coupled flip noise, extend this property through camera lift, flip fusion, the selected global IK, and root-velocity EMA to satisfy Equation 1 for final body outputs.*

4 EXPERIMENTS

Our evaluation addresses three questions: whether the complete generator preserves its history as a sequence grows, whether causal adaptation retains generation quality under visual and textual conditions, and whether prefix consistency enables efficient incremental execution. Component studies explain the roles of temporal conditioning, training objectives, and sampling.

4.1 Datasets, Metrics, and Protocol

For video conditioning, we evaluate on 3DPW (von Marcard et al., 2018) and EMDB (Kaufmann et al., 2023). On 3DPW and EMDB-1, we report PA-MPJPE, MPJPE, PVE, and acceleration error. On EMDB-2, following WHAM (Shin et al., 2024), we report W-MPJPE$_{100}$ (WA2), WA-MPJPE$_{100}$ (WAA), relative trajectory error (RTE), jitter, and foot sliding (FS). Position, vertex, and foot-sliding errors are in millimeters; RTE is a percentage, acceleration error is in m/s^2, and jitter is the released evaluator's m/s^3 value divided by 10. Lower is better for every metric.

For text conditioning, we evaluate on HumanML3D (Guo et al., 2022) with its pretrained text–motion evaluator after converting generated SMPL joints to the HumanML3D representation. Both models use the same 992 test caption/length conditions, spanning 40–196 frames, in five paired generation trials with classifier-free guidance (Ho & Salimans, 2022) scale 2.5. We report R-Precision at top 3 (R@3), FID, matching distance (MM Dist), and diversity. Higher R@3 and lower FID/MM Dist are better; diversity should approach the real-data value. Sampling, batching, and length normalization are specified in Appendix C.

PrefixMotion is fine-tuned with the rectified-flow objective for 200,000 optimization steps in a single training stage. Video reconstruction uses FP32 without TF32, flip testing, a 120-frame context, NFE-3 Euler, released DROID-SLAM (Teed & Deng, 2021) trajectories, root-velocity EMA 0.25, and global IK. Solver and context are selected on held-out 3DPW with camera trajectories derived from ground truth. EMA and IK settings are fixed before test evaluation; static-translation correction is disabled using the same validation split. GENMO denotes our evaluation of the official pretrained weights under each table's stated protocol. For video, GENMO uses DDIM-5 (Song et al., 2021); both models use seed 42, the official NVIDIA benchmark support data, and the released preprocessing. Five-seed robustness and fixed-weight inference ablations appear in Appendix E; controlled training comparisons appear in Appendices F, G, and J.

Training data. Adaptation uses a fixed mixture of AMASS (Mahmood et al., 2019), BEDLAM (Black et al., 2023), Human3.6M (Ionescu et al., 2014), 3DPW, HumanML3D, Motion-X++ (Zhang et al., 2025), and AIST++ (Li et al., 2021). Only the official 3DPW training split is used, with exactly 5% sampling probability. Motion-X++ includes text and 2D supervision, and AIST++ uses 35-D music features. The complete training protocol is reported in Appendix I.

4.2 Prefix Consistency under Sequence Extension

We first test whether causal attention alone preserves prefix agreement through shape aggregation and body decoding. For one selected 120-frame valid window from each of 37 3DPW test tracks, we compare the first t outputs of the full execution with a separate t-frame execution at $t \in \{16, 32, 64, 96\}$. Main and flipped random tensors are coupled on the shared prefix. Table 4 reports the component decomposition, separating output agreement from camera-feature causality.

PrefixMotion satisfies both numerical agreement and camera-input validity for every tested track and prefix length. Attention-only conversion leaves full-clip shape aggregation in place, causing decoded-tensor differences of up to 6.28. Running shape reduces the maxima below 10^{-5}; backward camera alignment removes the one-frame lookahead of a $t \to t+1$ feature. Causal EMA and IK preserve the result, and the rigid lift replaces symmetric orientation smoothing. This decomposition separates two requirements: prefix-compatible computation within the generator and prefix-compatible construction of its conditions. Numerical agreement on precomputed inputs alone tests only the former. The comparisons extend through final body parameters under Theorem 1's input assumptions; tensor scopes are defined in Appendix B.

4.3 Video-Conditioned Generation: Reconstruction Quality

Prefix-consistent conversion and reconstruction quality. PrefixMotion combines prefix-consistent inference with gains in 3DPW local geometry and EMDB-2 global trajectory (Table 1). Relative to GENMO, 3DPW PA-MPJPE, MPJPE, and PVE decrease by 1.181, 1.023, and 1.827 mm. On EMDB-2, W-MPJPE$_{100}$, WA-MPJPE$_{100}$, and RTE improve by 12.9%, 24.2%, and 19.7%. It is worse on 3DPW acceleration, all four EMDB-1 metrics, and EMDB-2 jitter and foot sliding, motivating the component analyses below. These results evaluate the complete adaptation, including training

Table 1: Video-conditioned reconstruction with GENMO and causal methods. Lower is better. GENMO is our reevaluation of the official pretrained weights with DDIM-5; PrefixMotion uses NFE-3 Euler. OnlineHMR shares finalized tracks and DROID-SLAM cameras but uses calibrated EMDB intrinsics, whereas PrefixMotion uses the released upstream preprocessing intrinsics (Appendix K). Bold compares OnlineHMR and PrefixMotion. PA denotes PA-MPJPE; position and foot-sliding errors are in millimeters, and RTE is a percentage.

	3DPW				EMDB-1			
Method	PA ↓	MPJPE ↓	PVE ↓	Accel ↓	PA ↓	MPJPE ↓	PVE ↓	Accel ↓
GENMO	36.067	56.098	68.194	4.990	42.424	73.509	85.170	3.590
OnlineHMR	39.134	67.483	79.036	8.279	50.205	81.040	94.711	9.030
PrefixMotion	**34.886**	**55.075**	**66.367**	**6.361**	**44.756**	**75.899**	**87.475**	**5.683**

	EMDB-2				
Method	W-MPJPE$_{100}$ ↓	WA-MPJPE$_{100}$ ↓	RTE ↓	Jitter ↓	Foot slide ↓
GENMO	305.342	122.772	2.290	19.826	4.584
OnlineHMR	**240.077**	**86.736**	**1.393**	84.257	21.753
PrefixMotion	265.948	93.015	1.838	**46.279**	**11.425**



Figure 2: Video-conditioned generation on two manually selected EMDB-1 sequences: outdoor dancing (top) and outdoor workout (bottom). Five instants are uniform quantiles from 0.1 to 0.9 of each valid sequence, cropped independently from ground-truth joints. Dashed white is ground truth; solid teal is PrefixMotion. Both are drawn with calibrated EMDB intrinsics and the same 22-joint SMPL-24 subset. Predictions are translated in 3D to the ground-truth mean hip for local pose comparison, without rotation or scale alignment.

and sampling; isolating the effect of removing future context requires a matched bidirectional-FM control.

Comparison with OnlineHMR. Under shared finalized tracks and DROID-SLAM cameras but model-specific EMDB intrinsics, PrefixMotion has lower aggregate error on 10 of 13 metrics: all eight 3DPW/EMDB-1 metrics, EMDB-2 jitter, and foot sliding. Jitter and foot sliding improve by 45.1% and 47.5%, respectively. PrefixMotion uses flip testing and running-prefix shape; OnlineHMR uses no flip testing and direct per-frame shape. Evaluating PrefixMotion without flip testing preserves the same ordering on all 13 metrics. The sequence-balanced intervals for these ten metrics exclude zero with and without flip testing (Table 25).

Table 2: Text-conditioned generation on HumanML3D using the same 992 test conditions, variable lengths of 40–196 frames, and five paired trials. Model entries are mean (SD); ground-truth entries are means under the same evaluation protocol. Bold compares model rows and marks diversity closer to ground truth.

Method	R@3 ↑	FID ↓	MM Dist ↓	Diversity →
Ground truth	0.774	0.000	3.193	9.278
GENMO	0.475 (0.046)	10.169 (2.494)	5.444 (0.372)	**7.547 (0.523)**
PrefixMotion	**0.492 (0.033)**	**8.110 (1.552)**	**5.291 (0.286)**	6.904 (0.102)

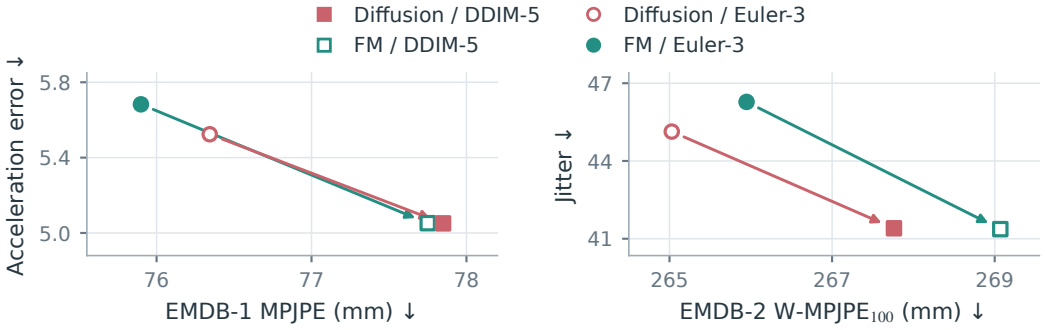

Figure 3: Effects of training objective and sampler at seed 42. Color identifies FM or diffusion-objective weights; circles and squares identify Euler-3 and DDIM-5. Filled markers are the corresponding training–sampler pairings; open markers exchange the samplers between trained models. Arrows change the sampler at fixed weights. Lower is better; full results are in Table 24.

Figure 2 illustrates the recovered articulation along two motion sequences. Translation-only alignment uses the same mean-hip-relative frame as MPJPE and PVE, separating articulated pose from global position. Unaligned camera-space overlays and comparisons of root trajectory, jitter, and contact appear in Figure 5. These examples show full-sequence predictions; the independent full-versus-prefix comparisons in Section 4.2 establish prefix agreement.

4.4 Text-Conditioned Generation

Text conditioning tests whether the causal adaptation preserves the pretrained model's semantic control beyond video reconstruction. Table 2 shows lower five-trial mean FID and matching distance together with higher R@3. R@3 and matching distance both indicate retained text–motion alignment, and the lower FID indicates closer distributional fidelity. Diversity falls further below the real-data value than GENMO's. These results support conditional synthesis under the prefix constraint. On AIST++ music-to-dance generation, PrefixMotion lowers kinetic FID from 78.903 to 68.957 and raises beat alignment from 0.213 to 0.248, while geometric FID and the foot-contact score are worse; Appendix D reports these results and examines continuity across overlapping generation windows.

4.5 Training Choices and Inference Paths

At fixed weights, replacing Euler-3 with DDIM-5 reduces acceleration error, jitter, and foot sliding while increasing local spatial errors and W-MPJPE$_{100}$ (Figure 3). Changing weights at a fixed sampler produces smaller, mixed differences. FM/Euler is lowest on all six local spatial errors among the four pairings. The exchanged samplers expose each model to states absent from its training, so this comparison characterizes inference sensitivity; the training and loss ablations are analyzed separately in Appendix C.

Table 3: Per-update inference latency on the same L40S GPU in FP32, with 120-frame context, three Euler evaluations, and two flip views. Timings include condition encoding, prediction, decoding, flip fusion, EMA, and IK. Speedup is relative to the median latency of window recomputation.

Execution	Median (ms) ↓	p95 (ms) ↓	Speedup ↑
Window recomputation	603	606	1.0×
Cached streaming	**57**	**57**	**10.6**×

4.6 Cached Streaming Inference

Prefix consistency enables stateful, frame-by-frame inference. For each of the three Euler evaluations, we maintain a 120-frame cache of attention keys and values, together with running shape, root-velocity EMA, and IK state. Each update encodes, predicts, and decodes the newest frame while preserving the completed history.

We compare streamed and full-sequence execution on 79 complete test sequences and 107,747 frames, with sequences up to 3,280 frames, in FP32 without TF32 and with coupled seed-42 noise. Raw predictions and both pre-fusion views meet tolerance at every frame. Decoded outputs meet the same tolerances except at one ill-conditioned flip-fusion frame; Appendix B.1 reports the complete checks and numerical analysis.

Table 3 shows a 10.6× median speedup for the cached implementation, which combines state reuse, CUDA Graph replay, and batched IK. Recomputing the window is also exact only for the first 120 frames. In the full pass, stacked attention layers and the running shape mean reach further back, and the cached streamer keeps this history. Condition encoding and geometric processing account for about 70% of the remaining latency (Appendix B.1).

On 3,588 frames from three sequences per dataset, streamed and full mixed-precision outputs differ by 0.26 mm mean camera-frame joint displacement, comparable to the 0.28 mm difference between full mixed-precision and FP32 inference. Caches and graph buffers add 189 MiB of GPU memory. Cached keys keep absolute RoPE positions, and decoder state carries history past the context.

5 Limitations

We apply PrefixMotion only to GENMO; transfer to other pretrained motion generators is untested. The video guarantee assumes compatible supplied person-track and camera-trajectory prefixes, so the benchmark does not establish end-to-end online tracking or camera estimation (Appendix L). Text-conditioned generation is evaluated only on HumanML3D, and combined video-and-text conditioning is untested.

6 Conclusion

PrefixMotion makes preserving generated history an explicit property of conditional motion generation. A shared temporal boundary across prediction, shape aggregation, sampling, and decoding yields prefix consistency under compatible conditions and coupled randomness. The clean-motion rectified-flow formulation retains compatibility with a pretrained generator, while attention caches and causal decoder state enable efficient frame-by-frame execution. Cached inference is evaluated on 79 complete sequences and achieves a 10.6× update speedup over window recomputation on the same GPU.

In experiments, PrefixMotion retains text-conditioned synthesis and improves selected video-reconstruction metrics: all three 3DPW spatial and all three EMDB-2 trajectory measures improve over GENMO, and 10 of 13 errors improve in the stated OnlineHMR comparison. Controlled studies further show how sampling and temporal state affect geometry, dynamics, and contact. Further work can integrate prefix-compatible tracking and camera estimation into a complete online system and evaluate music-conditioned synthesis over naturally long sequences.

ETHICS STATEMENT

Conditional human motion generation can support animation, accessibility, and interactive perception; video-conditioned reconstruction can also enable surveillance. Performance may vary with body shape, clothing, motion, camera motion, and occlusion represented in the training data. Deployment should include consent, data governance, and subgroup performance analysis. The evaluated uses are text- and music-conditioned motion synthesis and video-conditioned reconstruction; biomechanical measurement requires separate validation.

REPRODUCIBILITY STATEMENT

Appendices I–N document model initialization, training and evaluation protocols, evaluator verification, exploratory results, and the reproducibility checklist. Code, configurations, and trained weights will be released upon publication; Appendix N lists what the release includes. The supplementary material contains video reconstructions on benchmark sequences, comparing GENMO with frame-by-frame streamed PrefixMotion outputs.

AI USE STATEMENT

Large language models were used for literature retrieval and summarization; research ideation and experimental design; software implementation, experiment management, and interpretation of results; assistance in formulating and checking mathematical claims and proof drafts; and drafting, revising, and polishing portions of the manuscript. They were not used to generate datasets for training or evaluation. All AI-assisted citations, code, analyses, mathematical arguments, and text were independently reviewed and verified by the authors, who retain full responsibility for the work.

REFERENCES

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, 2024.

Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Guénolé Fiche, Philippe Weinzaepfel, Romain Brégier, and Fabien Baradel. Multi-HMR 2: Multi-person camera-centric human detection, mesh recovery and tracking. In *Proceedings of the European Conference on Computer Vision*, 2026.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. In *International Conference on Learning Representations*, 2026.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The electromagnetic database of global 3D human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. GENMO: A GENeralist model for human MOtion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI Choreographer: Music conditioned 3D dance generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

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*, 2022.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems*, 2021.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. EDGE: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Proceedings of the European Conference on Computer Vision*, 2018.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yufu Wang, Evonne Ng, Soyong Shin, Rawal Khirodkar, Yuan Dong, Zhaoen Su, Jinhyung Park, Kris Kitani, Alexander Richard, Fabian Prada, and Michael Zollhöfer. DuoMo: Dual motion diffusion for world-space human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xitong Yang, Devansh Kukreja, Don Pinkus, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jia-Wei Liu, Nicolás Ugrinovic, Anushka Sagar, Jitendra Malik, Matt Feiszli, Piotr Dollár, and Kris Kitani. SAM 3D Body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X++: A large-scale multimodal 3D whole-body human motion dataset. *arXiv preprint arXiv:2501.05098*, 2025.

Yiwen Zhao, Ce Zheng, Yufu Wang, Hsueh-Han Daniel Yang, Liting Wen, and László A. Jeni. OnlineHMR: Video-based online world-grounded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

A Formal Guarantees and Proofs

Proposition 2 (Clean-MSE and velocity-regression equivalence)**.** *For a scalar feature j at frame t, let $\mathcal{I}_{t,j}$ be the sigma-algebra generated by exactly the variables visible to that output under the causal architecture; it contains $X_{\boldsymbol{\tau},t,j}$ and τ_t but no future-frame input. Suppose the coordinate has positive clean-MSE weight, the same $X_{1,t,j}$ defines both corruption and supervision, $\tau_t < 1$, and $X_{1,t,j}$ has finite second moment. Then the unrestricted pointwise population minimizer of the clean-motion MSE term among $\mathcal{I}_{t,j}$-measurable predictors is, almost surely,*

$$f^*_{t,j} = \mathbb{E}[X_{1,t,j} \mid \mathcal{I}_{t,j}]. \tag{11}$$

Its induced ideal field is the coordinate-wise conditional velocity-regression optimum:

$$\frac{f^*_{t,j} - X_{\boldsymbol{\tau},t,j}}{1-\tau_t} = \mathbb{E}[X_{1,t,j} - Z_{t,j} \mid \mathcal{I}_{t,j}]. \tag{12}$$

Proof of Proposition 2. Fix a fully supervised scalar coordinate (t, j) satisfying the proposition, and abbreviate $\mathcal{I}_{t,j}$ by $\mathcal{I}$. For any $\mathcal{I}$-measurable scalar prediction a, the conditional squared-error risk decomposes as

$$\mathbb{E}\big[(a - X_{1,t,j})^2 \mid \mathcal{I}\big] = (a - \mathbb{E}[X_{1,t,j} \mid \mathcal{I}])^2 + \mathbb{E}\Big[(X_{1,t,j} - \mathbb{E}[X_{1,t,j} \mid \mathcal{I}])^2 \mid \mathcal{I}\Big]. \tag{13}$$

The second term is independent of a, so the conditional risk is minimized almost surely by $f^*_{t,j} = \mathbb{E}[X_{1,t,j} \mid \mathcal{I}]$.

For $\tau_t < 1$, Equation 4 gives

$$X_{1,t,j} - Z_{t,j} = \frac{X_{1,t,j} - X_{\boldsymbol{\tau},t,j}}{1 - \tau_t}. \tag{14}$$

Because $X_{\boldsymbol{\tau},t,j}$ and τ_t are $\mathcal{I}$-measurable, taking the conditional expectation yields

$$\mathbb{E}[X_{1,t,j} - Z_{t,j} \mid \mathcal{I}] = \frac{\mathbb{E}[X_{1,t,j} \mid \mathcal{I}] - X_{\boldsymbol{\tau},t,j}}{1 - \tau_t} \tag{15}$$

$$= \frac{f^*_{t,j} - X_{\boldsymbol{\tau},t,j}}{1 - \tau_t}, \tag{16}$$

which is Equation 12. Therefore clean-motion MSE and coordinate-wise direct velocity regression share the same almost-sure population optimum on these coordinates. Their objectives are not numerically identical: for the ideal field, or where the guard is inactive, substituting the induced field into velocity MSE scales clean-prediction error by $(1 - \tau_t)^{-2}$. Their time weighting and finite-capacity optimization can therefore differ.

This parameterization identity is not a support theorem. Independent continuous per-frame times do not identify the zero-probability shared-time manifold $\{s\mathbf{1}_T : s \in (0, 1)\}$ without additional regularity or generalization. The retained regression pass is labeled $\tau = 0$ but uses zero motion state rather than the Gaussian initial state used at inference, so it does not directly identify that solver state either. The argument also does not apply unchanged to detached pseudo-targets or auxiliary geometric objectives. □

Proposition 3 (Endpoint projection and oracle consistency)**.** *For the ideal field, and equivalently for the guarded implementation whenever $1 - \tau \geq \epsilon$, one Euler interval ending at one satisfies*

$$x + (1 - \tau)v^0_\theta(x, \tau, c) = f_\theta(x, \tau, c). \tag{17}$$

Consequently, the final NFE-3 interval needs no field evaluation at the endpoint. If f_θ is an exact clean-motion oracle along the shared-time straight path, the complete three-step Euler solve returns x_1 exactly.

Proof of Proposition 3. Substituting Equation 5 with $\tau < 1$ into an Euler step of length $1 - \tau$ gives

$$x + (1 - \tau)v^0_\theta(x, \tau, c) = x + (1 - \tau)\frac{f_\theta(x, \tau, c) - x}{1 - \tau} \tag{18}$$

$$= f_\theta(x, \tau, c). \tag{19}$$

Thus the endpoint update is an algebraic projection to the last clean-motion prediction and never evaluates the ideal field at $\tau = 1$. When $1 - \tau \geq \epsilon$, Equation 6 equals the ideal field. This includes every selected NFE-3 call at times zero, one third, and two thirds.

For oracle consistency, let $f_\theta = x_1$ along the straight path and start at $x^{(0)} = z$. At every point on the straight path, the oracle-induced velocity is the constant $x_1 - z$. Inductively applying Equation 8 gives

$$x^{(k)} = z + \tfrac{k}{3}(x_1 - z), \qquad k \in \{0, 1, 2, 3\}. \tag{20}$$

Thus every solver state lies exactly on the straight path, and $x^{(3)} = x_1$. Equivalently, Equation 17 maps the state at two thirds to the oracle's final clean prediction. Model error remains possible, but endpoint discretization introduces no additional error for an exact oracle. □

Proof of Theorem 1. Let P_t truncate a sequence after frame t. A sequence operator A is prefix-causal when $P_t A(y) = P_t A(P_t y)$ for every t. Tokenwise maps, residual sums, and compositions preserve this property. In a lower-triangular attention layer, the query at frame s uses keys and values only from frames at most s; the context band removes additional old frames but introduces no future edge. The transformer's normalization and MLP maps are tokenwise, and the inclusive running shape mean at frame s averages only frames $1{:}s$. Hence each denoiser call is prefix-causal.

By assumption, every frame-aligned condition tensor agrees through t when constructed from the full sequence or its prefix; the exogenous condition g is identical. Text cross-attention therefore receives identical keys and values for matching motion queries. Consider two sampler executions whose main and, when used, flipped random tensors agree through t. Their initial states agree through t. If their states agree through t before a learned-field evaluation, denoiser causality makes the predicted clean motions and induced velocities agree through t. Euler updates and endpoint projection are framewise algebraic operations, so their next states also agree through t. Induction over the three field evaluations proves prefix equality of the sampled motion. The same argument covers flip testing, because mirroring and fusion are framewise.

For video-conditioned generation, the output heads, motion decoder, camera-translation recovery, and SMPL forward kinematics are framewise given their prefix-compatible inputs. For $s > 1$, camera velocity uses the backward transition $s - 1 \to s$, with fixed identity/zero motion at $s = 1$. Given a supplied camera-pose prefix, the rigid lift at s uses only T_1 and T_s. Equation 10 is a forward recurrence. The IK target at frame s depends only on the previous target, the current joints, and the previous frame's contact confidence, and the cyclic-coordinate-descent (CCD) IK solve runs independently at each frame. Each selected postprocessor is therefore prefix-causal, and closure under composition proves Equation 1.

The video-conditioned extension is conditional on supplied person-track/bounding-box and camera prefixes. It does not prove that an offline detector, tracker, or SLAM backend used to construct those prefixes avoided later revisions. The theorem is an exact-arithmetic statement; the empirical audit in Section 4 measures floating-point agreement. □

Proposition 4 (EMA attenuation)**.** *For $0 \le \alpha < 1$ and temporally independent, zero-mean frame-displacement noise with common per-coordinate variance σ^2, the steady-state noise variance after Equation 10 is*

$$\lim_{t\to\infty} \mathrm{Var}(\bar{\epsilon}_t) = \frac{1-\alpha}{1+\alpha}\sigma^2. \tag{21}$$

The filter has unit gain at constant frame displacement and multiplies the Nyquist-frequency amplitude by $(1-\alpha)/(1+\alpha)$. Thus $\alpha = 0.25$ retains constant (zero-frequency) displacement while reducing idealized displacement-noise variance to $0.6\sigma^2$ and the Nyquist amplitude to 0.6 of its input value. Its nonzero phase response and finite-sequence initialization can introduce lag, consistent with the observed smoothness/foot-sliding tradeoff.

Proof of Proposition 4. Re-index the observed frame displacements from $n = 1$ and write $d_n = u_n + \epsilon_n$, where the ϵ_n are independent over time. The initialization in Equation 10 gives $\bar{\epsilon}_1 = \epsilon_1$. For $n \ge 2$, unrolling the recurrence gives

$$\bar{\epsilon}_n = \alpha^{n-1}\epsilon_1 + (1-\alpha)\sum_{k=0}^{n-2} \alpha^k \epsilon_{n-k}. \tag{22}$$

Temporal independence eliminates cross-covariance terms, so

$$\mathrm{Var}(\bar{\epsilon}_n) = \alpha^{2(n-1)}\sigma^2 + (1-\alpha)^2\sigma^2 \sum_{k=0}^{n-2} \alpha^{2k} \tag{23}$$

$$\xrightarrow[n\to\infty]{} (1-\alpha)^2\sigma^2 \sum_{k=0}^{\infty} \alpha^{2k} \tag{24}$$

$$= \frac{(1-\alpha)^2}{1-\alpha^2}\sigma^2 = \frac{1-\alpha}{1+\alpha}\sigma^2. \tag{25}$$

Table 4: Prefix-causality decomposition for one 120-frame valid window from each of 37 3DPW test tracks. "Numerical pass" requires prefixes 16, 32, 64, and 96 to meet tolerance: absolute 10^{-5} for the raw and camera scopes, 5×10^{-4} for the global scope, and relative 10^{-5}, combined as $|\Delta| \le \mathrm{atol} + \mathrm{rtol}\,|\mathrm{reference}|$. "Valid?" marks structural validity: all camera features are prefix-compatible. Bold compares numerical columns across the four configurations; numerical agreement and structural validity are assessed separately.

Configuration	Valid?	Numerical pass ↑	Raw max ↓	Camera max ↓	Global max ↓
Attention only	No	0/37	1.425	6.281	6.280
+ Running-prefix shape	No	**37/37**	9.54×10^{-6}	$\mathbf{2.38\times10^{-6}}$	$\mathbf{2.86\times10^{-6}}$
+ Causal camera features	Yes	**37/37**	$\mathbf{8.58\times10^{-6}}$	2.86×10^{-6}	$\mathbf{2.86\times10^{-6}}$
Full + EMA/IK	Yes	**37/37**	$\mathbf{8.58\times10^{-6}}$	2.86×10^{-6}	8.49×10^{-6}

The transfer function is

$$H(\zeta) = \frac{1-\alpha}{1-\alpha\zeta^{-1}}. \tag{26}$$

It satisfies $H(1) = 1$, so constant frame displacement is preserved, while $|H(-1)| = (1-\alpha)/(1+\alpha)$ at the Nyquist frequency. Integrating the filtered displacements from the original root preserves the first position. For $\alpha = 0.25$, both the stationary variance ratio and the Nyquist amplitude ratio are 0.6. The filter has nonzero phase delay, so this attenuation does not imply monotonic improvement in spatial or contact metrics. □

Corollary 5 (Rigid camera-trajectory lift preserves articulated geometry)**.** *At each frame, an* $\mathrm{SE}(3)$ *trajectory lift preserves every within-body Euclidean distance.*

Proof. For joints j, k transformed by the same framewise rigid motion, $p'_{t,j} = R_t p_{t,j} + b_t$ with $R_t \in \mathrm{SO}(3)$. Therefore

$$\|p'_{t,j} - p'_{t,k}\|_2 = \|R_t(p_{t,j} - p_{t,k})\|_2 = \|p_{t,j} - p_{t,k}\|_2. \tag{27}$$

The lift can change global trajectory error and orientation, but it cannot stretch the body or alter intrinsic, body-relative articulation. □

B Empirical Verification of Prefix Consistency

In appendix tables, ↓ means lower is better and ↑ means higher is better; → denotes closeness to a stated target. Bold marks the best value within the comparison specified in each caption, including ties at the reported precision.

We evaluate how each temporal operator affects finite-precision agreement between a sequence and its prefixes. The comparison uses one selected 120-frame valid window from each of the 37 3DPW test tracks, with prefix lengths 16, 32, 64, and 96. Main and flipped noise tensors are coupled on each compared prefix. The raw scope contains motion-state, camera-head, and contact-logit predictions; the camera and global scopes contain the respective decoded body parameter tensors. Maxima are computed over all entries in each scope and retain the units of their underlying representations. They measure agreement between executions; physical joint-distance errors are evaluated separately in Table 1. Camera-feature validity additionally records whether the conditioning itself respects the prefix boundary. This comparison covers prefixes up to 96 frames within a fixed window. Appendix B.1 extends the evaluation to every frame of the complete test sequences, preserving the condition prefixes and operator state required by the theorem.

B.1 Key–Value-Cached Streaming: Implementation and Numerical Analysis

Incremental execution must preserve both attention history and the state of the body decoder. For each of the three field evaluations, sixteen attention layers, and two flip views, the streamer caches the RoPE-rotated keys and the values of the most recent 120 frames (90 MiB in FP32). Cached keys keep their absolute RoPE positions, as in the full pass. It also carries the running shape sum, the IK

Table 5: Key–value-cached streaming versus the full-sequence pass on every test frame, in FP32 with coupled seed-42 noise. "Strict pass" requires every frame to meet the tolerances of Table 4. Views are the two camera-frame predictions entering flip fusion. Error columns are maximum absolute differences over all frames.

Dataset	Sequences	Frames	Strict pass ↑	Raw max ↓	Views max ↓	Camera max ↓	Global max ↓
3DPW	37	40,219	36/37*	1.14×10^{-5}	8.29×10^{-6}	7.72×10^{-4} *	7.71×10^{-4} *
EMDB-1	17	24,323	17/17	1.34×10^{-5}	2.86×10^{-6}	2.32×10^{-6}	3.27×10^{-5}
EMDB-2	25	43,205	25/25	1.24×10^{-5}	4.29×10^{-6}	2.03×10^{-6}	4.36×10^{-4}

*One ill-conditioned flip-fusion frame; see text.

static-joint targets and previous contact confidence, the root-velocity EMA state, and the first camera pose. Condition encoding, decoding, lift, and flip fusion apply the benchmark operators to one-frame tensors.

We test equivalence to full-sequence inference on all 79 test sequences from 3DPW, EMDB-1, and EMDB-2, totaling 107,747 frames; three EMDB sequences belong to both EMDB protocols and are counted in each. Every sequence exceeds the 120-frame context, and the longest contains 3,280 frames. Each streamed frame is compared with its full-sequence counterpart using coupled main and flipped noise with seed 42, in FP32 without TF32 (Table 5). Since the incremental computation receives only the current prefix, this comparison tests agreement at every sequence length, including beyond the attention-cache capacity.

Raw predictions and both pre-fusion views meet tolerance at every frame. The sole exception after decoding is localized to rotation averaging in flip fusion. At that frame, the main and unflipped root orientations differ by 179.997°, so their mean rotation matrix has singular values $(1, 2.3\times10^{-5}, 2.3\times10^{-5})$, and the orthogonal projection used by flip fusion amplifies a 3.6×10^{-7} input difference to 5.2×10^{-4} rad. No other frame has fusion conditioning $\min_j |\cos(\theta_j/2)|$ below 10^{-2}. The larger EMDB-2 global maxima occur in the IK-refined body pose and the integrated root translation, and reach at most 0.87 of the tolerance.

Finite-precision agreement depends on the arithmetic used for the comparison. Under mixed precision, one-token and full-sequence products round differently. On 3,588 frames from three sequences per dataset, streamed and full mixed-precision outputs differ by 0.26 mm mean camera-frame joint displacement. This is comparable to the 0.28 mm by which the mixed-precision full pass differs from FP32 inference.

On one L40S in FP32, a newest-frame update takes 57.1/57.4 ms median/p95. On the same GPU, recomputing the 120-frame window takes 0.603/0.606 s. Condition encoding takes 16.2 ms, the three field evaluations for both views take 17.6 ms, and decoding, fusion, IK, and EMA take 23.1 ms. The field evaluations replay captured CUDA graphs. IK solves the four limb chains, which share no joints, as one batch; this is bit-identical to solving them in sequence on all 107,747 test frames. Caches and graph buffers add 189 MiB of GPU memory; recomputation needs 14 MiB of transient memory.

C Text Generation and Inference-Path Analysis

C.1 HumanML3D Text-to-Motion

The text-generation comparison controls the captions, requested lengths, motion representation, and evaluation batches across models. We draw 1,000 eligible test caption/length conditions once with subset seed 7 and retain indices 0–991, yielding 992 conditions in 31 complete batches. Lengths are 40–196 frames, rounded down to multiples of four, with mean 141.44. Generation uses seeds 42–46. GENMO uses its native DDIM-5 sampler and PrefixMotion uses NFE-3 Euler; both apply classifier-free guidance (Ho & Salimans, 2022) with scale 2.5. The reported standard deviations reflect variation across these five trials; diversity also includes evaluator subsampling variation. Generated SMPL joints are converted to the HumanML3D representation before applying its pretrained text–motion evaluator.

Under this protocol, PrefixMotion improves five-trial mean R@3 (0.475 to 0.492), FID (10.169 to 8.110), and matching distance (5.444 to 5.291). Table 2 reports means and standard deviations.

Extension to music conditioning. Same-length future-music perturbations leave pre-cutoff outputs bit-exact on eight AIST++ sequences; on four synthetic 2,400-frame repeats, overlap generation is 14.3× faster than GENMO (Appendix D).

C.2 Training Choices and Inference Paths

The spatial–temporal balance depends on both the learned predictor and its numerical integration. At fixed weights, DDIM-5 lowers acceleration error, RTE, jitter, and foot sliding but raises local spatial errors and WA2 (Figure 3). Changing the training objective at a fixed sampler has smaller, mixed effects; the exchanged samplers expose each model to inference states absent from its training. FM/Euler is lowest on all six local spatial metrics (Table 24).

At 50,000 steps, the per-frame-time and clean-prefix design gives the best 3DPW test pose, scalar time (one shared flow time per sequence) gives the lowest held-out validation errors, and mixed-history generation changes the ranking (Appendix F). Clean-state regression lowers acceleration error relative to direct velocity. At 200,000 steps, the full objective improves held-out spatial error, EMDB-1 spatial error, and foot sliding relative to omitting auxiliary regression; the cheaper no-regression variant has lower error on 8 of 13 test metrics (Appendix G).

D Music-Conditioned Generation and Temporal Continuity

Evaluation protocol. We use the 40 conditions in the official cross-modal validation/test splits, generate 600 frames (20 seconds) at 30 FPS, and evaluate against all 411 public SMPL reference motions. Music conditioning uses the public 35-D EDGE feature extractor (Tseng et al., 2023). Motion features use the public FACT (Li et al., 2021) and Bailando (Siyao et al., 2022) kinetic and geometric extractors after 30-to-60 FPS interpolation, first-root centering, and normalization by real-set statistics. FID-k and Div-k use kinetic features and FID-m and Div-m geometric features (Li et al., 2021); PFC is the physical foot-contact score (Tseng et al., 2023) and BAS the beat-alignment score (Li et al., 2021). Chunked generation uses 120-frame windows, 30-frame overlap, coupled full-sequence noise, and joint-space crossfading. The reference archive contains 411 of the 1,408 sequences listed in the public camera mapping. Both models are evaluated using this same reference set and feature-processing procedure. GENMO's original AIST++ evaluator, generated samples, and inference seed are unavailable, so the comparison uses its official pretrained weights under the public protocol specified here.

Table 6: Matched raw-output AIST++ quality on 40 conditions. Both models use identical 120/30 chunking, coupled noise, joint crossfading, and no contact postprocessing. GENMO uses DDIM-50 and PrefixMotion uses Euler-3; both apply classifier-free guidance with scale 2.5, as in all AIST++ experiments. Lower FID and PFC and higher BAS are better; diversity should approach the reference (→). Bold compares GENMO and PrefixMotion.

Method	FID-k ↓	FID-m ↓	Div-k →	Div-m →	PFC ↓	BAS ↑	Sec./clip ↓
Public reference	–	–	9.559	7.264	1.399	0.273	–
GENMO	78.903	**24.343**	1.498	**3.218**	**0.632**	0.213	18.843
PrefixMotion	**68.957**	45.970	**3.791**	3.168	9.978	**0.248**	**1.339**

Under identical windowing and raw decoding, PrefixMotion lowers kinetic FID and increases beat alignment while using three Euler evaluations, compared with GENMO's DDIM-50 sampler (Table 6). Geometric FID, diversity, and contact measures yield different rankings. The following interventions separate sensitivity to music from the effects of contact postprocessing.

Dependence on future conditions. To test whether later music can alter earlier motion, we perturb the conditioning while keeping the sequence length and random tensors fixed. For eight sequences, we hold a full 600-frame noise tensor fixed and replace only the music embedding after cutoffs 120,

240, 360, and 480 with a strong deterministic perturbation. The model still executes at full length, so this tests same-shape future-condition leakage rather than prefix truncation.

Table 7: Dependence on future music. Each field is compared before four cutoffs on eight AIST++ sequences. Exact equality is bitwise. Bold zero deviations attain the minimum possible error in each scope.

Scope	Fields	Comparisons	Exact?	Max abs. ↓
Raw	motion state, camera head, contact logits	32 each	Yes	**0**
Global	pose, orientation, translation, shape	32 each	Yes	**0**

Music interventions. We compare aligned, lagged, and zero music embeddings while holding the sequences, inference noise, generation windows, postprocessing, and scoring beats fixed. Lagging by 60 frames is past-only and pads the first two seconds with the initial feature. Zero conditioning collapses geometric quality and kinetic diversity while producing an artificially low PFC. The lag also increases BAS. Together, these interventions demonstrate that the model uses music conditioning, while showing that PFC and BAS alone cannot establish correct temporal alignment with the supplied music.

Table 8: AIST++ music-conditioning interventions for PrefixMotion. Evaluation beats remain aligned in every row. Diversity targets are the public reference values in Table 6; bold compares the three conditioning variants for each metric.

Condition	FID-k ↓	FID-m ↓	Div-k →	Div-m →	PFC ↓	BAS ↑
Aligned	68.957	45.970	3.791	3.168	9.978	0.248
Lag 60 frames	**61.463**	**45.797**	**4.576**	3.314	10.488	**0.277**
Zero	97.115	478.992	0.604	**6.156**	**0.144**	0.241

Postprocessing and the GENMO comparison. Contact postprocessing materially changes AIST++ metrics and latency. Table 6 therefore compares raw outputs from both models: static-translation correction and IK are disabled for GENMO and PrefixMotion. This common decoding setting separates model and sampler differences from contact correction. Table 9 additionally compares GENMO's default postprocessing with raw output and ablates both contact corrections for PrefixMotion.

Table 9: Effect of contact postprocessing on AIST++. "Configured" uses GENMO's default inference procedure; "raw" disables static translation and IK. Lower FID/PFC and higher BAS are conventionally better. Bold marks the best value per metric across the listed settings.

Model	Postprocessing	FID-k ↓	FID-m ↓	PFC ↓	BAS ↑	Sec./clip ↓
GENMO	Configured	72.774	**22.041**	**0.219**	0.218	20.269
GENMO	Raw	78.903	24.343	0.632	0.213	18.843
PrefixMotion	Raw	**68.957**	45.970	9.978	**0.248**	**1.339**
PrefixMotion	IK	70.465	44.829	4.117	0.245	2.555
PrefixMotion	Static translation	190.668	41.906	2.325	0.239	1.393
PrefixMotion	Static translation + IK	192.014	40.661	0.794	0.240	2.598

The effectiveness of contact correction depends on how predicted confidence corresponds to motion stationarity. We examine this correspondence using a post-hoc 0.25 m/s stationarity threshold on generated motion; the head was trained with a threshold of 0.15 m/s. Across 124,800 foot-channel predictions, kinematic stationarity has 2.47% prevalence while mean predicted confidence is 32.90%. At a 0.5 threshold, precision is 3.03%, recall is 21.66%, F1 is 0.053, Brier score is 0.144, expected calibration error is 0.304, and 96.97% of predicted static contacts are moving. This is a generated-motion self-consistency diagnostic, not calibration against annotated real contacts. The discrepancy helps explain why static translation can lower PFC while distorting root motion and increasing kinetic FID.

Long-horizon stress test. We periodically repeat each 600-frame music condition to 1,200 and 2,400 frames and evaluate four matched sequences. This synthetic extension is evaluated for runtime, overlap disagreement, and temporal continuity. Distributional quality and music alignment on naturally long sequences are outside this repeated-signal experiment; FID, PFC, and BAS are therefore omitted. Both models use raw outputs. Transition ratios divide derivative magnitude at chunk transitions by the sequence-interior mean, so one is ideal. Overlap disagreement is measured between adjacent generated windows before blending.

Table 10: Long-horizon chunking. "Ov." is overlap in frames. Root velocity/acceleration/jerk and body-local jerk are transition ratios (one is ideal). The final columns are pre-blend local/root disagreement in millimeters. Bold compares the two models at the same sequence length and overlap; $\rightarrow 1$ denotes closeness to one.

Model	Frames	Ov.	Sec. ↓	Root vel. $\rightarrow 1$	Root accel. $\rightarrow 1$	Root jerk $\rightarrow 1$	Local jerk $\rightarrow 1$	Local mm ↓	Root mm ↓
PrefixMotion	600	0	**1.014**	60.95	94.33	**81.52**	**2.90**	–	–
PrefixMotion	600	30	**1.332**	1.50	1.88	**1.63**	**1.04**	**135.4**	**104.1**
PrefixMotion	1,200	0	**1.902**	58.93	89.76	**77.91**	**3.46**	–	–
PrefixMotion	1,200	30	**2.470**	1.52	1.89	**1.63**	**1.04**	**113.9**	**102.0**
PrefixMotion	2,400	0	**3.809**	57.08	**85.81**	**74.22**	**3.61**	–	–
PrefixMotion	2,400	30	**5.137**	1.49	**1.84**	**1.60**	**1.01**	**110.0**	**99.1**
GENMO	600	0	13.676	**23.62**	**64.43**	101.75	20.06	–	–
GENMO	600	30	19.065	**1.39**	**1.79**	2.35	1.29	185.8	125.2
GENMO	1,200	0	27.171	**28.79**	**78.43**	122.53	20.37	–	–
GENMO	1,200	30	35.292	**1.35**	**1.72**	2.34	1.29	186.3	118.1
GENMO	2,400	0	54.331	**32.44**	89.67	137.66	20.69	–	–
GENMO	2,400	30	73.215	**1.39**	1.91	2.57	1.38	198.6	128.6

E Solver and Inference Robustness

We examine numerical integration and temporal postprocessing with the generator weights held fixed. Solver selection uses held-out 3DPW, while the test sets quantify the resulting spatial and temporal differences. Both Euler and Heun use three field evaluations, allowing their update rules to be compared at the same evaluation budget.

Table 11: Solver selection on held-out 3DPW validation. Latency is timed on one 120-frame 3DPW test window and covers the complete parallel inference and causal postprocessing after model and data loading on one L40S. Lower is better except window frames/s, which is not the newest-frame online update rate.

Solver	PA ↓	MPJPE ↓	PVE ↓	Accel ↓	Latency (s) ↓	Window frames/s ↑
NFE-3 Heun	**33.239**	**49.739**	61.562	9.244	**0.600**	**199.9**
NFE-3 Euler	33.261	49.743	**61.553**	**9.003**	0.602	199.4

We assess variation across both generated noise and test sequences by comparing NFE-3 Euler and Heun under matched inference seeds 40–44. The first two numerical columns in Table 13 are means and sample standard deviations of the official frame-weighted aggregate metric over seeds. For uncertainty, we first average each sequence over the five seeds and then resample sequences 100,000 times. The paired delta and confidence interval therefore need not equal the difference between the two frame-weighted aggregate means.

Table 12: Solver ablation with fixed model weights on the full benchmark after selection on 3DPW validation. EMDB values are diagnostic test-set evidence, not additional model-selection criteria. Lower is better.

3DPW				
Solver	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Heun	34.910	**55.053**	**66.318**	6.584
Euler	**34.886**	55.075	66.367	**6.361**

EMDB-1				
Solver	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Heun	**44.649**	76.102	87.737	5.968
Euler	44.756	**75.899**	**87.475**	**5.683**

EMDB-2					
Solver	W-MPJPE $_{100}$ ↓	WA-MPJPE $_{100}$ ↓	RTE (%)↓	Jitter↓	Foot slide↓
Heun	266.261	93.328	1.843	49.492	11.600
Euler	**265.948**	**93.015**	**1.838**	**46.279**	**11.425**

Table 13: Solver robustness over five inference seeds with fixed model weights. Δ is Euler minus Heun. Bold marks the lower solver mean for each metric; standard deviations are not ranked. † marks paired confidence intervals that exclude zero.

Dataset	Metric	Euler mean $\pm$ std	Heun mean $\pm$ std	Paired $\Delta \downarrow$ [95% CI]
3DPW	PA-MPJPE↓	**34.910** $\pm$ 0.022	34.935 $\pm$ 0.019	-0.021 [-0.067, $+0.025$]
	MPJPE↓	55.072 $\pm$ 0.070	**55.044** $\pm$ 0.063	$+0.048$ [-0.099, $+0.204$]
	PVE↓	66.371 $\pm$ 0.087	**66.317** $\pm$ 0.078	$+0.071$ [-0.072, $+0.221$]
	Accel↓	**6.361** $\pm$ 0.004	6.583 $\pm$ 0.004	-0.223 [-0.244, -0.204]†
EMDB-1	PA-MPJPE↓	44.745 $\pm$ 0.007	**44.636** $\pm$ 0.008	$+0.110$ [$+0.042$, $+0.183$]†
	MPJPE↓	**75.854** $\pm$ 0.099	76.055 $\pm$ 0.099	-0.184 [-0.351, -0.012]†
	PVE↓	**87.437** $\pm$ 0.096	87.694 $\pm$ 0.099	-0.240 [-0.429, -0.049]†
	Accel↓	**5.681** $\pm$ 0.003	5.968 $\pm$ 0.003	-0.285 [-0.316, -0.255]†
EMDB-2	W-MPJPE$_{100}$↓	**266.151** $\pm$ 0.215	266.413 $\pm$ 0.157	-0.280 [-0.735, $+0.169$]
	WA-MPJPE$_{100}$↓	**92.901** $\pm$ 0.071	93.209 $\pm$ 0.079	-0.293 [-0.414, -0.171]†
	RTE↓	**1.837** $\pm$ 0.001	1.842 $\pm$ 0.001	-0.0067 [-0.0119, -0.0028]†
	Jitter↓	**46.300** $\pm$ 0.089	49.495 $\pm$ 0.090	-3.297 [-3.614, -3.012]†
	Foot slide↓	**11.419** $\pm$ 0.019	11.593 $\pm$ 0.021	-0.185 [-0.220, -0.151]†

Euler's most consistent gains occur in temporal and world-motion quality. It improves acceleration on all 37 3DPW sequences, EMDB-1 acceleration on all 17 sequences, EMDB-2 jitter on all 25 sequences, and foot sliding on 24/25 sequences. Its 3DPW spatial intervals include zero, while EMDB-1 PA-MPJPE increases by 0.110 mm with a confidence interval above zero. This distinction between temporal gains and dataset-dependent spatial changes is consistent with the held-out validation comparison.

EMA, IK, context, and camera. EMA and IK address different sources of temporal error: root motion and body contact, respectively. The EMDB-2 factorial in Table 14 quantifies their separate and combined effects at fixed model weights. The main evaluation uses EMA 0.25 and global IK, specified before test evaluation; static-translation correction is disabled based on 3DPW validation.

Table 14: Euler EMA×IK factorial on EMDB-2 with fixed model weights. Lower is better. EMA primarily stabilizes the root; IK primarily improves contact. The local 3DPW and EMDB-1 outputs are unchanged.

EMA	IK	W-MPJPE$_{100}$↓	WA-MPJPE$_{100}$↓	RTE↓	Jitter↓	Foot slide↓
0.00	Off	266.571	92.944	1.840	68.270	16.308
0.00	On	266.622	93.275	1.840	64.228	**11.252**
0.25	Off	**265.783**	**92.611**	**1.838**	48.804	14.888
0.25	On	265.948	93.015	**1.838**	**46.279**	11.425

Table 15: Held-out 3DPW context selection and window-recomputation update rate for PrefixMotion on one L40S with fixed model weights. Every update recomputes the retained window; no key-value cache is used. Lower is better except updates/s.

Context	PA↓	MPJPE↓	PVE↓	Accel↓	Updates/s↑
30	33.313	49.997	61.789	9.080	**1.82**
60	33.269	49.838	61.639	9.037	1.76
120	**33.261**	**49.743**	**61.553**	9.003	1.66
240	33.443	49.968	61.828	**9.001**	1.49
Unlimited	35.023	54.498	67.170	9.192	1.49†

†Measured on a 240-frame sequence.

The 120-frame context balances spatial accuracy and computation, giving the lowest held-out PA-MPJPE, MPJPE, and PVE. A 240-frame context reduces acceleration error by 0.002 but increases all three spatial errors; unlimited history also increases these errors. With 120-frame recomputation, latency is 0.603 s median and 0.606 s p95 per newest-frame update. This update recomputes only the trailing 120-frame window. Because the band applies in every attention layer and running shape accumulates from the first frame, the benchmark pass gives later frames a longer effective history, so

the trailing-window update reproduces it exactly only within the first 120 frames. The key–value-cached streamer in Appendix B.1 reproduces the benchmark pass at every frame, apart from one ill-conditioned flip-fusion frame.

Camera causality. Holding the supplied trajectory fixed, future-looking camera-feature alignment improves EMDB-1 acceleration from 5.683 to 5.486, EMDB-2 jitter from 46.279 to 45.611, and foot sliding from 11.425 to 11.280; all spatial changes are below 0.06 mm. Because the weights were trained with forward $t \to t+1$ alignment, these differences combine one-frame lookahead with the change from training-time to causal alignment. They therefore measure the combined effect of these changes, rather than future access in isolation.

To measure the effect of camera-trajectory revision, we compare frozen online poses with retrospectively optimized poses from the same MASt3R-SLAM (Murai et al., 2025) reconstruction for all 39 unique EMDB protocol sequences (62,707 frames). MASt3R-SLAM optimizes keyframes rather than every tracked frame. For each keyframe, we compute the left-multiplied SE(3) correction from its frozen online pose to its final pose, interpolate these corrections in SE(3), and apply them to the online poses of intervening frames. Both variants use regenerated, calibrated camera conditions. Their paired differences isolate retrospective optimization under this common protocol; absolute scores use different intrinsics from the main reconstruction comparison.

We exclude failed reconstruction states using tracking and geometric consistency, without ground-truth camera or body metrics. An optimized keyframe estimate is accepted only after the following frame tracks successfully; an estimate still pending at the end of a sequence must also pass loose dimensionless path-length, spatial-extent, and correction-ratio checks. Final estimates are tracker-confirmed for 38 of 39 sequences; the remaining one is validated by the transform, path-length, extent, and correction-ratio checks. The rule rejects two estimates that immediately preceded tracker failure; the latest accepted estimate is used instead.

Table 16: Frozen online and retrospectively optimized MASt3R trajectories using PrefixMotion with fixed weights and Euler sampling. Delta is final minus online; lower is better. Bold compares the online and validated-final estimates within each metric.

Protocol	Metric	Online	Validated final	$\Delta \downarrow$
EMDB-1	PA-MPJPE ↓	45.0138	**45.0136**	-0.0003
	MPJPE ↓	77.9016	**77.9008**	-0.0008
	PVE ↓	88.9509	**88.9499**	-0.0010
	Accel ↓	5.6933	**5.6931**	-0.0002
EMDB-2	W-MPJPE$_{100}$ ↓	363.275	**362.382**	-0.894
	WA-MPJPE$_{100}$ ↓	**128.215**	128.898	$+0.683$
	RTE (%) ↓	**3.6389**	3.6544	$+0.0156$
	Jitter ↓	41.9734	**41.9058**	-0.0676
	Foot slide ↓	**10.6522**	10.6861	$+0.0339$

Sequence-level bootstrap intervals (100,000 resamples) include zero for EMDB-2 W-MPJPE$_{100}$, WA-MPJPE$_{100}$, RTE, and foot sliding. Jitter improves on all 25 sequences, with mean delta -0.0707 and 95% interval $[-0.0954, -0.0485]$. Thus retrospective finalization provides a small smoothness benefit, with no consistent change across the other trajectory and contact measures.

F Causal-FM Training-Design Ablation

We isolate two training choices: per-frame flow times and clean-prefix conditioning. Each variant initializes from the pretrained bidirectional GENMO weights and trains for 50,000 steps; the seven-dataset mixture, 5% 3DPW sampling, global batch 48, seed 42, optimizer schedule, and architecture are shared. The scalar arm uses one shared flow time per sequence and no forced clean prefix. The per-frame arm samples independent valid-frame times without a forced clean prefix. The full method combines per-frame times with a randomized clean-history prefix on 50% of training samples.

Test-set evaluations use seed-42 NFE-3 Euler, context 120, root-velocity EMA 0.25, global IK, and disabled static-translation correction. EMDB is diagnostic test evidence and is not used for model selection.

Table 17: Held-out 3DPW validation for the causal-FM training-design study, logged during training with NFE-3 Heun under mixed precision; values therefore differ slightly from the standalone FP32 evaluation in Table 11. Lower is better.

Time/prefix design	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Full method	33.482	50.080	62.000	9.788
Scalar time	**33.477**	**49.631**	**61.532**	**9.576**
Per-frame, no prefix	33.512	49.876	61.759	9.738

Table 18: Complete benchmark evaluation of three training designs after 50,000 steps. All rows use the same inference settings and seed 42; lower is better.

3DPW

Design	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Full method	**34.704**	**54.513**	**65.586**	**6.481**
Scalar time	35.013	55.558	66.557	6.528
Per-frame, no prefix	34.966	55.257	66.373	6.483

EMDB-1

Design	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Full method	44.910	77.401	89.110	5.978
Scalar time	44.609	**74.746**	**86.432**	6.044
Per-frame, no prefix	**44.574**	74.955	86.650	**5.953**

EMDB-2

Design	W-MPJPE$_{100}$↓	WA-MPJPE$_{100}$↓	RTE (%)↓	Jitter↓	Foot slide↓
Full method	267.037	93.556	1.838	47.724	11.526
Scalar time	**263.260**	**92.831**	**1.826**	48.112	**11.409**
Per-frame, no prefix	265.188	93.150	1.835	**47.469**	11.627

Scalar time gives the lowest errors on held-out validation and on four of five EMDB-2 metrics. The full method gives the best 3DPW test pose, while per-frame time without a clean prefix gives the best EMDB-1 PA-MPJPE and acceleration and the best EMDB-2 jitter. These comparisons assess the two training choices at a common budget; the main benchmark uses the predefined full method trained for 200,000 steps.

Held-out mixed-history stress test. To assess robustness when completed motion is reused as history, we evaluate the three models on held-out 3DPW with completed blocks clamped clean. A block size of 16 generates short continuations, while a block size of one generates one frame at a time. Table 19 reports the change relative to standard joint integration (stress minus standard), so negative is better. Targets, masks, camera support, seed, Euler-3 solver, and postprocessing are identical across all nine evaluation settings.

Table 19: Effect of mixed-history inference on three models trained for 50,000 steps. Evaluation uses held-out 3DPW validation. Values are differences from standard joint integration (stress minus standard); negative is better. PA, MPJPE, PVE, and Accel are sequence means of per-sequence differences. Root jerk is the difference in frame-pooled mean world-frame pelvis jerk, in m/s^3 divided by 10 as for jitter. Bold compares training designs at the same rollout block size.

Training design	Block	PA↓	MPJPE↓	PVE↓	Accel↓	Root jerk↓
Full method	16	$+0.021$	-0.069	-0.096	$\mathbf{+0.106}$	$\mathbf{-0.745}$
Scalar time	16	-0.006	-0.118	-0.163	$+0.295$	$+1.377$
Per-frame, no prefix	16	$\mathbf{-0.012}$	$\mathbf{-0.228}$	$\mathbf{-0.375}$	$+0.203$	-0.136
Full method	1	$+0.313$	$+0.622$	$+0.987$	-0.366	$+0.508$
Scalar time	1	$\mathbf{-0.070}$	$\mathbf{+0.149}$	$\mathbf{+0.243}$	$+0.115$	$+0.380$
Per-frame, no prefix	1	$+0.423$	$+0.662$	$+0.594$	$\mathbf{-0.531}$	$\mathbf{-7.247}$

Sequence-balanced 100,000-resample comparisons show that the full method has a lower stress-minus-standard acceleration change than scalar time at block 16 (-0.189, 95% CI $[-0.271, -0.118]$)

and block 1 (-0.480, $[-0.731, -0.271]$). Under strict block 1, however, it has larger PA-MPJPE and PVE degradation than scalar time and smaller acceleration improvement than per-frame time without prefix forcing. The experiment therefore does not support a general robustness advantage from clean-prefix forcing. Independent per-frame times appear more closely associated with acceleration robustness under strict rollout, while spatial and root-jerk robustness remain mixed.

G CAUSAL-FM LOSS-OBJECTIVE ABLATION

The clean-motion formulation raises two empirical questions: how its time weighting affects optimization, and what the auxiliary regression objective contributes. All three variants initialize from the pretrained GENMO weights and use the same 50,000 updates, data mixture, batch 48, seed 42, optimizer schedule, architecture, and dual-mode forward structure. The full method applies clean-state MSE in both the retained all-zero-state regression pass labeled $\tau = 0$ and the sampled rectified-path pass. The direct-velocity arm preserves the regression objective but replaces only the rectified-path clean-state MSE with MSE on the induced velocity. The no-regression-loss arm keeps a detached all-zero-state $\tau = 0$ forward to provide pseudo-clean targets for partially supervised data in the same way, but excludes its losses from backpropagation. Both ablations therefore preserve the partially supervised examples and their target source.

Test-set evaluations use seed-42 NFE-3 Euler, context 120, root-velocity EMA 0.25, global IK, and disabled static-translation correction. EMDB remains diagnostic test evidence and is not used for model selection.

Table 20: Held-out 3DPW validation after 50,000 training steps in the loss-objective ablation, logged during training with NFE-3 Heun under mixed precision. Lower is better.

Objective	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Full method	33.482	50.080	62.000	9.788
Direct velocity	34.310	51.230	62.801	11.582
No regression loss	**33.381**	**49.925**	**61.587**	**9.663**

Table 21: Complete benchmark evaluation of loss objectives after 50,000 training steps. Lower is better.

3DPW

Objective	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Full method	**34.704**	**54.513**	**65.586**	**6.481**
Direct velocity	35.467	56.423	66.986	7.877
No regression loss	34.752	54.794	65.736	6.496

EMDB-1

Objective	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Full method	44.910	77.401	89.110	5.978
Direct velocity	45.099	76.953	88.484	8.042
No regression loss	**44.703**	**75.022**	**86.609**	**5.905**

EMDB-2

Objective	W-MPJPE$_{100}$↓	WA-MPJPE$_{100}$↓	RTE (%)↓	Jitter↓	Foot slide↓
Full method	267.037	93.556	1.838	47.724	**11.526**
Direct velocity	266.072	93.827	1.827	58.802	12.620
No regression loss	**264.057**	**92.414**	**1.826**	**46.307**	11.564

Clean-state regression yields lower errors than direct velocity regression on every 3DPW metric and both acceleration measures at 50,000 steps. Direct velocity is not the best objective on any of the 13 metrics; its larger EMDB-2 jitter and foot-sliding errors are consistent with the strong near-endpoint weighting $(1 - \tau)^{-2}$ in Section 3. The auxiliary term has a different effect: omitting it improves all four EMDB-1 metrics and four of five EMDB-2 metrics, whereas the full objective retains lower 3DPW test errors and lower EMDB-2 foot sliding. The full-budget comparison below examines whether this pattern persists with longer training.

Auxiliary regression at the full training budget. The no-regression variant is trained to the same budget of 200,000 steps as PrefixMotion, preserving optimizer moments and the learning-rate schedule. The regression forward remains detached and continues to provide pseudo-clean targets; only its loss gradients are absent. Table 22 reports both models after this prespecified number of training steps.

Table 22: Auxiliary regression ablation after 200,000 training steps. Held-out 3DPW is logged during training with NFE-3 Heun under mixed precision; test-set results use seed-42 NFE-3 Euler, context 120, EMA 0.25, global IK, and no static translation correction. Lower is better.

Held-out 3DPW validation

Objective	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Full method	**33.232**	**49.722**	**61.551**	9.373
No regression loss	33.307	50.028	61.654	**9.365**

3DPW test

Objective	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Full method	34.886	55.075	66.367	**6.361**
No regression loss	**34.846**	**54.872**	**65.947**	6.384

EMDB-1

Objective	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
Full method	**44.756**	**75.899**	**87.475**	5.683
No regression loss	45.100	76.671	88.367	**5.648**

EMDB-2

Objective	W-MPJPE$_{100}$↓	WA-MPJPE$_{100}$↓	RTE (%)↓	Jitter↓	Foot slide↓
Full method	265.948	93.015	1.838	46.279	**11.425**
No regression loss	**265.309**	**92.553**	**1.832**	**45.152**	11.552

At the full budget, the auxiliary objective supports held-out spatial accuracy and EMDB-1 spatial reconstruction. Omitting it gives lower error on 8 of 13 test metrics, concentrated in 3DPW geometry and EMDB-2 trajectory and jitter, while increasing every held-out spatial error, every EMDB-1 spatial error, 3DPW acceleration error, and foot sliding. On one GPU with batch 48, omitting auxiliary backpropagation lowers peak allocated memory from 31.1 to 19.6 GiB and raises steady throughput from roughly 1.4 to 1.8 updates/s; inference cost is unchanged. This variant offers a training-efficiency tradeoff with dataset-dependent accuracy. The main comparison retains the predefined dual-mode objective for PrefixMotion.

H Comparison with Staged Adaptation

A staged adaptation baseline sequentially applies causal diffusion training, flow matching, low-rate 3DPW sampling, and a final stage with higher 3DPW sampling, pseudo-global 3DPW targets, and T5-3B (Raffel et al., 2020) text embeddings. This comparison evaluates the complete adaptation strategy, including its inference sampler. The PrefixMotion configuration and its Euler solver are selected independently of this comparison.

Relative to staged adaptation, PrefixMotion lowers 3DPW PA-MPJPE by 0.297 mm, 3DPW/EMDB-1 acceleration by 0.847/1.426, EMDB-2 jitter by 6.477, and foot sliding by 0.003. These gains characterize the combined training-and-sampling strategy. Figure 4 reports all 13 metrics alongside GENMO. Staged adaptation has lower error on the remaining eight: 3DPW MPJPE and PVE, all three EMDB-1 spatial metrics, and EMDB-2 W-MPJPE$_{100}$, WA-MPJPE$_{100}$, and RTE.

I Model Initialization and Training

PrefixMotion initializes from the pretrained bidirectional GENMO SMPL model and is fine-tuned for 200,000 optimization steps using the seven-dataset mixture described in Appendix J, with T5-3B

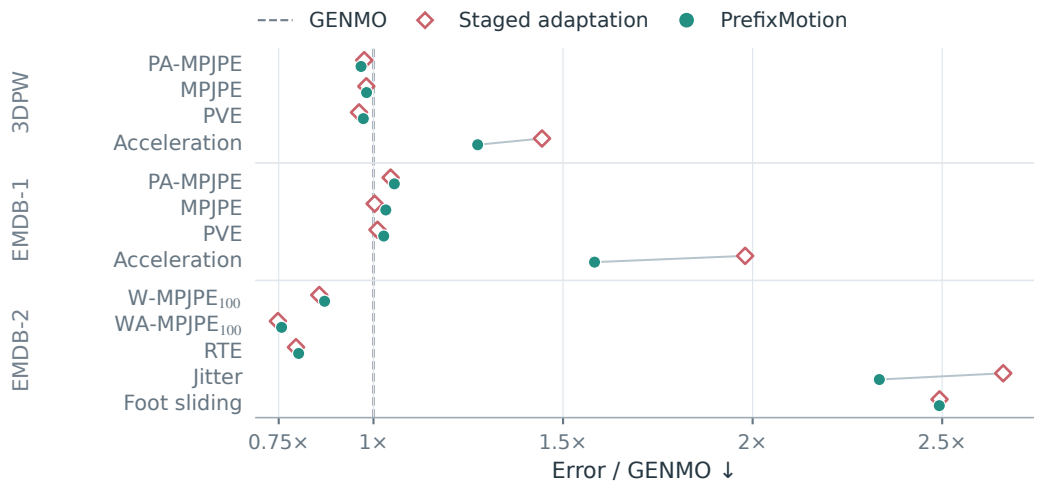

Figure 4: GENMO, staged adaptation, and PrefixMotion. Each error is divided by the corresponding GENMO value; the dashed line marks a ratio of one. Points are offset vertically for visibility. Lower is better. Both adapted models use inference seed 42; staged adaptation uses NFE-3 Heun and PrefixMotion uses the NFE-3 Euler solver. PrefixMotion has lower 3DPW PA-MPJPE, acceleration error, jitter, and foot sliding; staged adaptation has lower error on the other eight metrics. Training schedule and solver both differ between the adapted models. GENMO uses the same baseline values as Table 1.

text embeddings, Motion-X++ text and 2D supervision, 35-D AIST++ music features, and exactly 5% 3DPW sampling.

The staged adaptation baseline in Figure 4 uses a different schedule: pretrained GENMO is adapted through causal diffusion training, 50,000 flow-matching updates, 5,000 updates with 1% 3DPW sampling, and 15,000 updates with 5% 3DPW sampling, pseudo-global 3DPW targets, and T5-3B HumanML3D text embeddings, then evaluated with NFE-3 Heun. The 120-frame context, Euler sampler, camera lift, EMA, IK, and disabled static-translation correction are inference choices applied to fixed PrefixMotion weights.

J Training Objectives and Sampler Interactions

To compare training objectives at a common adaptation budget, all variants use the same seven-dataset mixture: AMASS, BEDLAM, Human3.6M, 3DPW, HumanML3D, Motion-X++, and AIST++. 3DPW has exactly 5% sampling probability; the remaining probability is distributed in proportion to dataset-loader lengths. BEAT2 is excluded. All configurations use seed 42, random 16–120-frame causal windows, per-frame times, 50% random clean prefixes, acceleration weight 400, and mixed precision. In Equation 7, $\mathcal{L}_{\text{transl/contact}}$ is the unit-weight sum of camera-parameter MSE, an L1 loss on world root translation integrated from the predicted local root velocity, and static-contact binary cross-entropy. The FM and diffusion-objective models use the same AdamW optimizer, global batch 48, data mixture, seed, update count, and initial learning rate 2×10^{-5}, with step decays at 40,000, 80,000, 120,000, and 160,000 steps. The two models differ in their training objective. A third variant, Diffusion→FM, uses the diffusion objective for 150,000 steps and flow matching for the final 50,000 steps. The transition preserves model parameters, optimizer moments, and learning-rate schedule. The diffusion-objective model is evaluated with its native DDIM-5 sampler, and FM with the validation-selected NFE-3 Euler sampler. We subsequently exchange these samplers to distinguish changes in the learned model from changes in numerical integration.

FM achieves lower error than the diffusion-objective model on 8 of 13 metrics, including all six 3DPW/EMDB-1 spatial metrics; the diffusion-objective model achieves lower error on both acceleration metrics and EMDB-2 RTE, jitter, and foot sliding. The diffusion-to-FM model improves 7 of 13 metrics relative to FM. In 100,000-resample paired sequence bootstraps, the switch significantly im-

Table 23: Training objectives compared after 200,000 optimization steps with matched data, optimizer, and batch size. Evaluation uses DDIM-5 for the diffusion-objective model and NFE-3 Euler for FM, comparing each objective with its corresponding sampler. Lower is better.

3DPW

Training objective	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
FM	34.886	55.075	66.367	6.361
Diffusion	35.151	55.618	67.092	**5.958**
Diffusion → FM	**34.848**	**54.908**	**66.238**	6.329

EMDB-1

Training objective	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
FM	**44.756**	**75.899**	**87.475**	5.683
Diffusion	46.112	77.849	89.673	**5.051**
Diffusion → FM	44.792	76.537	88.079	5.640

EMDB-2

Training objective	W-MPJPE$_{100}$↓	WA-MPJPE$_{100}$↓	RTE↓	Jitter↓	Foot slide↓
FM	265.948	**93.015**	1.838	46.279	11.425
Diffusion	267.761	93.057	**1.824**	**41.403**	**11.119**
Diffusion → FM	**265.739**	93.089	1.839	45.755	11.435

proves 3DPW acceleration (-0.0389, 95% CI $[-0.0853, -0.0063]$), EMDB-1 acceleration (-0.0441, $[-0.0615, -0.0268]$), and EMDB-2 jitter (-0.5433, $[-0.7156, -0.3797]$). It significantly regresses EMDB-1 MPJPE ($+0.5889$, $[+0.2147, +1.0220]$) and PVE ($+0.5471$, $[+0.1240, +1.0218]$); the remaining intervals include zero. The principal distinction is between spatial accuracy and temporal quality: FM/Euler has lower spatial errors, while diffusion/DDIM-5 has lower acceleration error, RTE, jitter, and foot sliding. Late objective switching preserves this distinction. Since the samplers also differ, these comparisons describe the combined effect of learning and numerical integration.

Objective–sampler interaction. We evaluate both FM and diffusion-objective weights with both samplers while holding seed, context, EMA, IK, camera support, and evaluator settings fixed. The off-diagonal combinations are diagnostic: diffusion-objective weights under Euler see rectified-flow states absent during training, while FM weights under DDIM-5 see diffusion-path states absent during training.

At fixed weights, changing Euler-3 to DDIM-5 causes the largest and most consistent shifts: local spatial errors and WA2 generally worsen, whereas acceleration, RTE, jitter, and foot sliding improve. At a fixed sampler, changing training weights produces smaller, mixed differences. Paired 100,000-resample sequence bootstraps confirm, for both weight sets, the sampler effects on both acceleration metrics, 3DPW PA-MPJPE and PVE, EMDB-1 spatial errors, EMDB-2 WA2, RTE, jitter, and foot sliding.

We further examine whether diffusion training with Euler-3 offers a different spatial–temporal balance. Across inference seeds 40–44, diffusion-objective weights under Euler-3 improve both acceleration metrics, WA2, RTE, and jitter on every seed relative to FM/Euler, but worsen all six local spatial metrics on every seed and worsen foot sliding on four seeds. The combination satisfies all 28 real-sequence prefix checks (seven 3DPW test windows at four prefix lengths) and has the same NFE-3 cost, but its held-out 3DPW paired intervals all include zero and its paired acceleration mean does not improve. We therefore retain the validation-selected FM/Euler configuration for PrefixMotion.

K Evaluation Protocols and Numerical Verification

We verify consistency of the metric implementations before comparing models. PA-MPJPE, MPJPE, PVE, Accel, WA2, WAA, RTE, Jitter, and foot-sliding computations are elementwise identical to the GENMO evaluator on masked synthetic tensors. Evaluation runs in FP32 with TF32 disabled; TF32 amplifies small vertex differences when the SMPL (Loper et al., 2015) joint regressor operates on

Table 24: Joint comparison of training objective and sampler at seed 42. Every row uses context 120, EMA 0.25, global IK, causal camera alignment, and identical evaluation support. Lower is better; bold is best across all four combinations. The two exchanged-sampler settings evaluate inference states that are outside the corresponding training distributions.

3DPW

Weights / sampler	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
FM / Euler-3	**34.886**	**55.075**	**66.367**	6.361
Diffusion / Euler-3	34.931	55.222	66.457	6.262
FM / DDIM-5	35.203	55.724	67.188	5.998
Diffusion / DDIM-5	35.151	55.618	67.092	**5.958**

EMDB-1

Weights / sampler	PA-MPJPE↓	MPJPE↓	PVE↓	Accel↓
FM / Euler-3	**44.756**	**75.899**	**87.475**	5.683
Diffusion / Euler-3	44.972	76.343	87.735	5.524
FM / DDIM-5	46.005	77.750	89.867	5.052
Diffusion / DDIM-5	46.112	77.849	89.673	**5.051**

EMDB-2

Weights / sampler	W-MPJPE$_{100}$↓	WA-MPJPE$_{100}$↓	RTE↓	Jitter↓	Foot slide↓
FM / Euler-3	265.948	93.015	1.838	46.279	11.425
Diffusion / Euler-3	**265.028**	**92.949**	1.835	45.134	11.430
FM / DDIM-5	269.069	93.033	1.826	**41.370**	**11.080**
Diffusion / DDIM-5	267.761	93.057	**1.824**	41.403	11.119

world translations of hundreds of meters. Evaluation targets generated independently by our pipeline and by the upstream pipeline agree exactly on all 37, 17, and 25 sequences of 3DPW, EMDB-1, and EMDB-2, and an independent evaluation of GENMO with seed 42, DDIM-5, upstream preprocessing, and released DROID support reproduces all 13 values in Table 1 exactly.

OnlineHMR protocol. We evaluate OnlineHMR with the official `Tsukasane/Video-OnlineHMR` implementation (`beta_test_w_tracking` branch) and pretrained weights on the original RGB frames, with the supplied benchmark tracks and calibrated benchmark intrinsics. Its current-plus-two-frame memory resets at each track boundary, and shape is estimated directly per frame. We apply neither flip testing nor test-set tuning. EMDB-2 lifts the decoded body through the same released DROID trajectory as PrefixMotion, so causality remains conditional on supplied track and camera prefixes. Full-sequence and prefix executions give identical decoded outputs (maximum difference zero) on representative 3DPW-validation and EMDB sequences. Throughput is 14.03 frames/s excluding detector, tracker, and camera estimation, which are shared protocol inputs. Aggregate results appear in Table 1. EMDB intrinsics remain model-specific: OnlineHMR uses benchmark calibration, while PrefixMotion uses the released upstream preprocessing intrinsics. This protocol compares body estimators with supplied tracks and trajectories. Its masks, preprocessing, and per-frame shape treatment differ from OnlineHMR's published evaluation, and DROID replaces its MASt3R-SLAM trajectory estimator.

Table 25: Sequence-balanced paired bootstrap for PrefixMotion versus OnlineHMR under shared finalized tracks and DROID support but model-specific EMDB intrinsics. Delta is PrefixMotion minus OnlineHMR; negative favors PrefixMotion. "Flip" is the main PrefixMotion setting and "no flip" matches OnlineHMR's released no-flip inference setting. Intervals use 100,000 sequence resamples. Bold marks the smaller mean delta between the two inference settings for each metric; confidence bounds are not ranked.

Dataset	Metric	Flip $\Delta \downarrow$ [95% CI]	No-flip $\Delta \downarrow$ [95% CI]
3DPW	PA-MPJPE ↓	$\mathbf{-4.152}$ $[-5.097, -3.287]$	-3.380 $[-4.447, -2.387]$
	MPJPE ↓	$\mathbf{-11.956}$ $[-14.092, -9.677]$	-11.349 $[-13.581, -9.004]$
	PVE ↓	$\mathbf{-12.020}$ $[-14.588, -9.409]$	-10.779 $[-13.420, -8.090]$
	Accel ↓	$\mathbf{-1.965}$ $[-2.258, -1.700]$	-1.414 $[-1.710, -1.145]$
EMDB-1	PA-MPJPE ↓	$\mathbf{-5.397}$ $[-7.411, -3.666]$	-4.140 $[-6.025, -2.568]$
	MPJPE ↓	$\mathbf{-5.350}$ $[-8.976, -1.656]$	-3.762 $[-7.226, -0.204]$
	PVE ↓	$\mathbf{-7.351}$ $[-11.728, -2.999]$	-5.407 $[-9.566, -1.156]$
	Accel ↓	$\mathbf{-3.409}$ $[-3.972, -2.906]$	-2.484 $[-2.973, -2.047]$
EMDB-2	W-MPJPE$_{100}$ ↓	$\mathbf{+26.597}$ $[+5.549, +48.883]$	$+28.518$ $[+7.230, +50.324]$
	WA-MPJPE$_{100}$ ↓	$\mathbf{+5.577}$ $[+1.710, +9.620]$	$+6.122$ $[+2.132, +10.279]$
	RTE (%) ↓	$\mathbf{+0.540}$ $[+0.295, +0.845]$	$+0.541$ $[+0.294, +0.850]$
	Jitter ↓	$\mathbf{-38.695}$ $[-42.489, -34.972]$	-36.942 $[-40.710, -33.227]$
	Foot slide ↓	$\mathbf{-10.726}$ $[-12.035, -9.478]$	-10.361 $[-11.631, -9.140]$

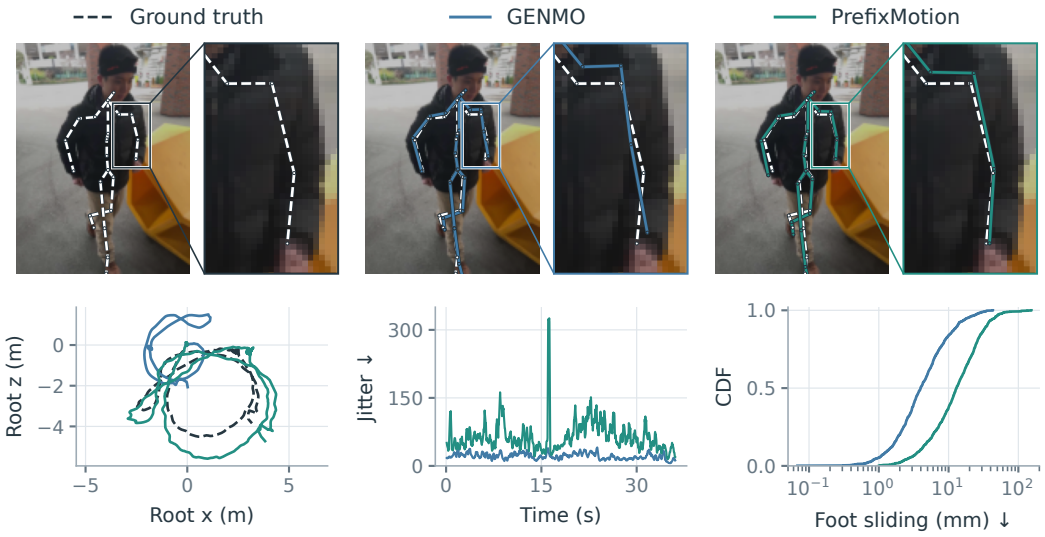

Figure 5: Spatial and temporal error analysis for GENMO and PrefixMotion. Above are camera-frame overlays for ground truth, GENMO, and PrefixMotion; each has an arm detail at the same $3.57\times$ magnification and image-coordinate crop, linked to the boxed region in the full view. In this detail, PrefixMotion's shoulder, elbow, and wrist projections lie closer to ground truth. Below, from left to right, are first-camera-anchored root trajectories, joint jitter, and the evaluator foot-sliding distribution. The fixed cases use the selection criteria below. Ground truth is projected with calibrated EMDB intrinsics and predictions with their model's estimated intrinsics; these overlays are unaligned. Jitter uses a nine-frame rolling mean, and foot sliding is shown as an empirical CDF on a logarithmic horizontal axis.

L Scope and Error Analysis

The diagnostic cases in Figure 5 use inference seed 42 and fixed selection criteria for PrefixMotion: lowest EMDB-1 PA-MPJPE for local pose, highest EMDB-2 W-MPJPE$_{100}$ for root drift, and highest normalized jitter-plus-foot-sliding score for temporal error.

Causal model versus online system. The prefix guarantee concerns generation from compatible supplied person-track and camera-trajectory prefixes with coupled noise. The video benchmarks provide finalized tracks and DROID-SLAM trajectories, so a live system also requires prefix-compatible upstream estimates. The 39-sequence MASt3R comparison finds negligible local changes under trajectory revision and small, mixed world-space changes; only EMDB-2 jitter improves consistently. Rejecting two reconstruction states immediately preceding tracking failure also highlights the need to manage online camera state. The OnlineHMR comparison fixes the same tracks and DROID support and therefore evaluates body estimation separately from detection, tracking, and SLAM.

Streaming inference. The key–value-cached streamer reproduces the parallel pass frame by frame and takes 0.057 s per update on one L40S, versus 0.603 s for recomputing the 120-frame window (Appendix B.1). Its equivalence holds to FP32 rounding; under mixed precision, agreement is limited by half-precision rounding. Per-frame condition encoding and geometry account for about 70% of the remaining latency. A live system would still need online tracking and camera estimation.

Geometry, dynamics, and world state. Relative to our evaluation of pretrained GENMO, PrefixMotion improves the 3DPW spatial and EMDB-2 trajectory metrics; in the stated OnlineHMR comparison, it improves local accuracy, acceleration error, jitter, and foot sliding. The full metric comparison appears in Table 1. A benchmark-wide error decomposition of the staged-adaptation model (Appendix H) further localizes opportunities for improvement: distal limbs dominate camera-space acceleration and root translation dominates world jitter. These findings motivate learned causal dynamics and online world-state estimation as extensions to the current generator.

Attribution and generation scope. The causal diffusion/FM comparison matches initialization, data, optimizer, schedule, batch, seed, and updates, while DDIM-5 and NFE-3 Euler differ in numerical integration. The complete objective–sampler comparison evaluates both off-diagonal combinations, but each exposes weights to inference states absent during its training. No bidirectional-FM control has been trained, so the effect of removing future context is not isolated. HumanML3D covers fixed caption/length pairs under the stated sampling protocol. Mixed video-plus-text conditioning and Motion-X++ generation remain untested. The AIST++ study covers 40 conditions from the public split under a composite public evaluator, not exact reproduction of GENMO's unpublished evaluation; the long-horizon study repeats a 20-second music signal and therefore supports continuity/runtime conclusions only. Appendix K documents the shared evaluation support and the GENMO baseline verification.

M EXPLORATORY TEMPORAL MODELING

Recurrent adaptation and post-hoc calibration. We investigate whether a lightweight recurrent state can refine generated motion while preserving causal computation. A unidirectional GRU is added to transformer blocks 12–15 of the trained PrefixMotion model. It has 3,981,870 trainable parameters, while the base generator remains frozen. The adapter is trained through the deterministic NFE-3 Euler sampler, with losses on its final output that are constrained relative to a frozen, matched supervised control. At inference, the adapter residual is scaled by $\alpha = 0.50$.

Table 26 reports differences between the adapted model and PrefixMotion over inference seeds 40–44. At this residual scale, all 13 aggregate errors decrease on every seed, with changes generally below 1%.

Table 26: Exploratory recurrent-state adaptation. Values are mean differences between the adapted model and PrefixMotion over five matched inference seeds; negative is better. "SD" is the population standard deviation across the five inference seeds. Bold negative deltas indicate improvement over the unadapted model ($\Delta = 0$); SD is not ranked. The main model is fixed independently of this exploratory comparison.

Dataset	Metric	Mean Δ ↓	SD	Improved seeds ↑
3DPW	PA-MPJPE ↓	**−0.1798**	0.0014	5/5
	MPJPE ↓	**−0.3801**	0.0041	5/5
	PVE ↓	**−0.3886**	0.0063	5/5
	Accel ↓	**−0.0474**	0.0011	5/5
EMDB-1	PA-MPJPE ↓	**−0.0353**	0.0022	5/5
	MPJPE ↓	**−0.0234**	0.0056	5/5
	PVE ↓	**−0.0855**	0.0069	5/5
	Accel ↓	**−0.0560**	0.0003	5/5
EMDB-2	W-MPJPE$_{100}$ ↓	**−1.9520**	0.0446	5/5
	WA-MPJPE$_{100}$ ↓	**−0.2633**	0.0054	5/5
	RTE ↓	**−0.002411**	0.000024	5/5
	Jitter ↓	**−0.2429**	0.0047	5/5
	Foot slide ↓	**−0.003746**	0.000944	5/5

The residual scale was selected after examining the test results: held-out 3DPW favored $\alpha = 1.10$, while $\alpha = 0.50$ gave more uniform test-set improvements. At the default scale ($\alpha = 1$), the adapter also significantly increases EMDB-2 WA-MPJPE$_{100}$, RTE, and foot sliding relative to the matched supervised control. The measurements therefore characterize the combined adapter and post-hoc calibration, rather than isolating the terminal objective. They motivate further study of recurrent refinement with independent model selection and evaluation; the prespecified PrefixMotion model remains the basis of the main comparisons.

Additional preliminary studies examine history representations and external cues using an intermediate adapted model or frozen-feature probes, and camera supervision using short continuations of PrefixMotion against a matched control. These are separate from the controlled ablations of the main model.

History conditioning. The history-conditioning comparisons examine how the representation of completed motion affects its continuation. Finalized-history block sampling with one-frame blocks reduces acceleration error but increases PA-MPJPE, MPJPE, and PVE. Independent rectified history noise increases acceleration error, while variance-preserving history noise increases both acceleration error and root jerk. History modifications that smooth motion also cost spatial accuracy.

External priors and visual cues. Frozen-feature and small-adapter studies examine whether stronger external cues improve dynamics. Frozen DuoMo world-prior transfer improves individual trajectory or jitter measures, but not trajectory, jitter, and foot sliding jointly. PromptHMR (Wang et al., 2025) and frozen OnlineHMR token features improve velocity measures while increasing acceleration error. Multi-HMR 2 cues improve frame-local body, hand, foot, and anchored-root position estimates, with higher acceleration error. Part-local DINOv3 (Siméoni et al., 2025) features improve velocity but not acceleration; causal Video Swin (Liu et al., 2022) features increase both

errors. In these preliminary comparisons, more accurate frame-local cues alone do not ensure better temporal dynamics.

Camera supervision. Changing the allocation of camera losses across datasets has limited benefit in the tested configurations. Combined dataset-routed supervision increases 9 of 13 errors; orientation-only routing is neutral. Body-local routing increases EMDB-2 W-MPJPE$_{100}$ by 0.532 mm, and the largest improvement from camera-loss routing is 0.231 mm in W-MPJPE$_{100}$, accompanied by higher acceleration error. Together with the cue experiments, these results motivate joint learning of temporal state and camera-grounded dynamics, extending beyond frozen features and small adapters.

N Reproducibility Checklist

The code release upon publication will include:

- exact training and evaluation configurations, random seeds, and numerical precision settings;
- documented dataset preprocessing;
- evaluation scripts that report complete per-sequence 3DPW, EMDB-1, and EMDB-2 results;
- scripts for the inference ablations over EMA, IK, static translation, NFE, solver, context, camera lift, and smoothing;
- the key–value-cached streamer with its frame-by-frame equivalence and latency scripts; and
- scripts for the AIST++ music interventions, future-condition leakage, contact calibration, and 600–2,400-frame chunking experiments.