FLOODDIFFUSION 2: EFFICIENT AND PATH CONTROLLABLE STREAMING MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present FloodDiffusion 2 (FD2), an efficient and controllable framework that builds upon FloodDiffusion (FD1), a state-of-the-art streaming motion generation model. While FD1 produces plausible motion, it suffers from low efficiency and limited controllability, as its attention design requires repeated computation over the entire history, and it lacks precise trajectory control for real-world applications. To address these limitations and improve generation quality, FD2 introduces three advances. First, *Partial Attention* makes finalized history representations independent of the active window, enabling KV-cached inference and shared-history packing for efficient training. Second, we establish a necessary-and-sufficient Bregman criterion for regression losses to preserve diffusion's conditional-mean velocity field. This criterion guides an FK-induced quadratic loss that incorporates motion geometry without online FK evaluation. Third, FD2 introduces precise path conditioning to control the character's root trajectory while preserving natural body motion. Experiments show that FD2 reduces training computation by 4.6× and accelerates denoising by 11.29×, reaching 2.303 ms per update on long sequences. Alongside these efficiency gains, FD2 improves motion quality over FD1 and achieves state-of-the-art FID scores among streaming methods, with 0.048 on SEED and 0.053 on HumanML3D.

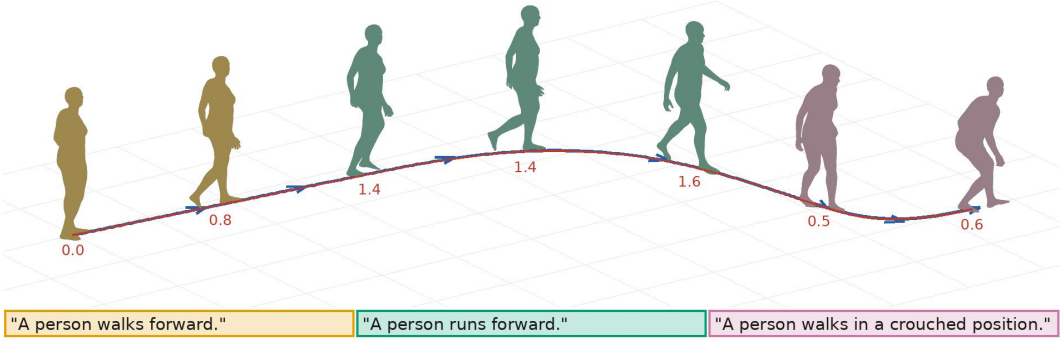

Figure 1: **FloodDiffusion 2** generates continuous human motion from time-varying text prompts while following a specified path. The character transitions smoothly from walking to running and then crouched walking. Blue and red curves denote the target and generated center-of-mass trajectories, respectively; numbers indicate the generated planar speed in m/s.

1 INTRODUCTION

Streaming text-to-motion generation produces motion incrementally while accepting changing conditions (Zhao et al., 2025; Xiao et al., 2025; Cai et al., 2026; Zhao et al., 2026). Interactive virtual characters require efficient generation and precise trajectory control alongside natural, text-aligned motion. FloodDiffusion (FD1), a state-of-the-art streaming motion generation model, realizes this

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.

process through a frame-staggered diffusion schedule: frames enter and complete denoising successively as the active region advances (Cai et al., 2026).

FD1 has two practical limitations. First, its bidirectional attention allows finalized history to depend on the changing active window. History representations therefore change during denoising, preventing standard KV caching and requiring repeated computation during training and inference. Second, FD1 provides only text-based control, which cannot precisely specify even a character's walking direction. Real-world applications require the character to follow a specified route while performing different actions. We therefore introduce FloodDiffusion 2 (FD2) to improve efficiency and enable precise path control while preserving natural motion.

To improve efficiency, FD2 retains FD1's schedule and sampler and employs Partial Attention. History queries read their causal prefixes; active queries read the complete history and interact bidirectionally within the active window. This preserves access to the inputs of the streaming velocity field while making history representations independent of subsequent frames. FD2 uses this structure to support standard KV-cached inference and shared-history training. During inference, finalized history representations can be cached and reused as the active window advances. During training, one clean reference encoding is shared across multiple isolated denoising windows, reducing repeated history computation.

To improve generation quality, we align supervision with human perception of motion. Errors in motion features, such as local joint rotations, do not directly reflect their effects on body geometry: rotation errors can accumulate along kinematic chains and produce noticeable positional errors at the hands and feet. A natural approach is to introduce a forward kinematics (FK) loss that penalizes errors in the reconstructed body geometry. However, directly applying a nonlinear FK loss can shift the regression optimum away from the conditional-mean velocity field required by diffusion.

We therefore ask which losses preserve this velocity field while incorporating geometric supervision. Sampled velocities are regression targets, but the field to be learned is their conditional mean. Applying the classical mean-elicitation result (Banerjee et al., 2005; Gneiting, 2011), we establish a necessary-and-sufficient criterion under the stated regularity and integrability conditions: the compatible losses are exactly target-first Bregman divergences generated by strictly convex potentials, up to a target-only term. Guided by this criterion, we construct an FK-induced quadratic from a dataset-averaged kinematic response, incorporating geometric sensitivity while preserving the conditional mean and avoiding online FK evaluation.

FD2 incorporates root-path conditioning to enable precise trajectory control. The model receives a frame-aligned root trajectory together with time-varying text prompts, allowing spatial movement and action semantics to be specified jointly. In our SEED representation, the root is a mass-weighted center-of-mass ground projection, with explicit root-relative pelvis translation and rotation. This separates the commanded trajectory from the body's articulation, reducing gait-cycle sway while preserving natural body motion. Figure 1 illustrates changing actions along a specified path.

Experiments on SEED and HumanML3D demonstrate improvements in both efficiency and motion quality. Shared-history packing reduces training denoiser computation by $4.6\times$. With 4,500 history frames, KV caching accelerates denoising by $11.29\times$ to 2.303 ms per update on an RTX 4090. Alongside these efficiency gains, FD2 improves HumanML3D FID from FD1's 0.057 to 0.053 and achieves an FID of 0.048 on SEED, yielding state-of-the-art FID scores among the compared streaming text-to-motion methods on both benchmarks.

Our contributions are:

- **Partial Attention for efficient streaming diffusion** enables standard KV-cached inference and reduces training computation through shared-history multi-window supervision.
- **A necessary-and-sufficient characterization of diffusion-compatible losses** connects conditional-mean velocity regression to Bregman divergences and guides an FK-induced quadratic objective for geometric supervision.
- **Root-path conditioning** enables precise trajectory control while preserving natural root-relative body motion.

2 Related Work

Motion generation and streaming. Motion generation spans continuous diffusion (Tevet et al., 2023; Zhang et al., 2024), autoregression (Zhang et al., 2023a), retrieval augmentation (Zhang et al., 2023b), and masked token modeling (Guo et al., 2024). MotionStreamer and ARDY generate incrementally (Xiao et al., 2025; Zhao et al., 2026). FD1 uses frame-staggered diffusion (Cai et al., 2026), building on Diffusion Forcing's per-element noise levels (Chen et al., 2024). FD2 retains FD1's schedule and sampler.

Efficient streaming diffusion. Transformer-XL reuses history states for autoregressive generation (Dai et al., 2019). Rolling Forcing combines causal history with a bidirectional active window for KV caching (Liu et al., 2026). It uses distribution-matching distillation from a pretrained teacher and, like Self Forcing (Huang et al., 2025), trains on self-generated histories. FD2 employs Partial Attention for direct velocity supervision on motion data, sharing a clean reference encoding across isolated denoising windows.

Geometry-aware supervision. MDM supplements feature regression with position, velocity, and foot-contact losses (Tevet et al., 2023). Such geometric supervision motivates our objective, but nonlinear FK losses can shift the conditional-mean velocity target. We apply classical mean-elicitation theory (Banerjee et al., 2005; Gneiting, 2011) to identify compatible regression losses and propose an FK-induced quadratic objective that preserves this target without online FK evaluation.

Spatially controlled motion generation. GMD and OmniControl support spatial joint and trajectory constraints alongside text (Karunratanakul et al., 2023; Xie et al., 2024). Kimodo overwrites constrained features with clean values and supplies a control mask; its denoiser predicts root before body, using a smoothed root distinct from the pelvis (Rempe et al., 2026). ARDY overwrites constrained global root channels while generating latent body motion and accepts future path constraints beyond the active window (Zhao et al., 2026). FD2 similarly holds three relative root channels clean and diffuses explicit body features within FD1's frame-staggered schedule. Its SEED root is a mass-weighted center-of-mass ground projection, with separate pelvis motion preserving root-relative articulation.

3 Methods

FD2 retains FD1's frame-staggered diffusion schedule and sampler (Cai et al., 2026). We employ Partial Attention for efficient inference and training, characterize diffusion-compatible losses to design an FK-induced geometric objective, and incorporate root-path conditioning for precise trajectory control.

3.1 Preliminaries

We use a 138-D rotation-direct state for SEED and native 263-D features for HumanML3D. For a clean standardized motion sequence $\mathbf{z}$, independent noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, and frame-aligned conditions $\mathbf{c}$, frame k follows

$$\alpha_t^k = \mathrm{clamp}(t - k/n_s, 0, 1), \qquad \beta_t^k = 1 - \alpha_t^k, \tag{1}$$

where n_s controls the active-window width. The noisy state and sampled velocity are

$$\mathbf{x}_t^k = \alpha_t^k \mathbf{z}^k + \beta_t^k \boldsymbol{\epsilon}^k, \qquad \mathbf{v}_t^k = \dot{\alpha}_t^k \mathbf{z}^k + \dot{\beta}_t^k \boldsymbol{\epsilon}^k. \tag{2}$$

Away from schedule knots, finalized history $H = \{k : \alpha_t^k = 1\}$ precedes the active window $A = \{k : 0 < \alpha_t^k < 1\}$, where $\mathbf{v}_t^k = \mathbf{z}^k - \boldsymbol{\epsilon}^k$. Future frames remain pure noise; history and future have zero sampled velocity. The prescribed marginal velocity is

$$u_t^A(\mathbf{x}_t^{H\cup A}, \mathbf{c}^{H\cup A}) = \mathbb{E}[\mathbf{v}_t^A \mid \mathbf{x}_t^{H\cup A}, \mathbf{c}^{H\cup A}]. \tag{3}$$

FD1's streaming-locality result makes this processed prefix sufficient under its causal-conditioning assumption. Appendix A.1 states the assumption and exact interval indices.

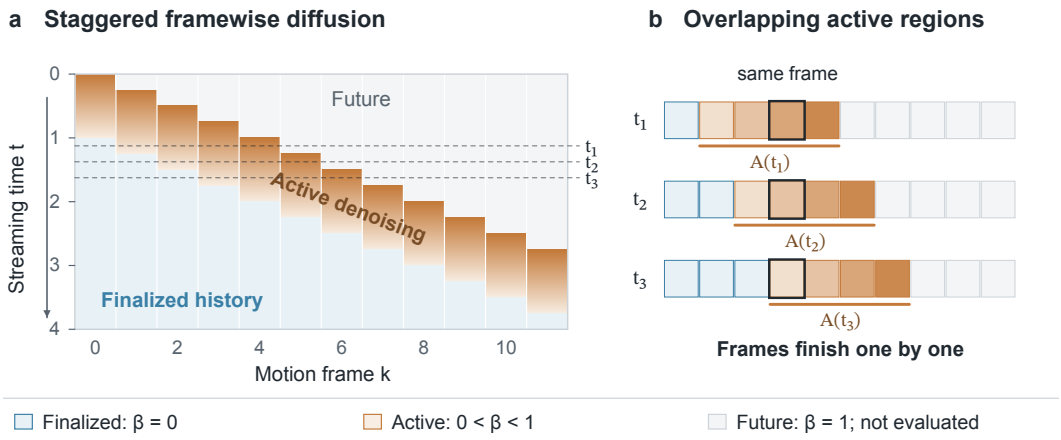

Figure 2: Frame-staggered diffusion schedule retained from FD1. (a) Time-shifted denoising ramps separate finalized history, the active window, and future noise. (b) Overlapping active windows advance as frames finish denoising one by one. The illustration uses $n_s = 4$; β denotes the noise amplitude.

3.2 Partial Attention

We employ Partial Attention to enable KV-cached inference and shared-history training. History queries are causal, while active queries read the complete processed prefix:

$$\mathbf{A}^{\mathrm{p}}_{j,k} = \begin{cases} 1 & j \in H,\ k \le j, \\ 1 & j \in A,\ k \in H \cup A, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Thus history can be encoded independently of later frames, while active outputs retain access to the information in Equation (3). Appendix A.1 establishes this decomposition and gives a data-law counterexample to fully causal active outputs.

KV-cache inference. With row-wise normalization and feed-forward operations, causal history states remain unchanged when their inputs, positions, noise metadata, and frame-aligned conditions are fixed. FD2 caches their layer-wise keys and values, re-encoding each newly finalized frame once under the causal mask before appending it. When prompts change, the conditions of finalized frames remain fixed, preserving their cached representations. Appendix A.2 proves prefix invariance.

Shared-history training. For each sequence, we sample one diffusion time from each of K strata and pack the resulting noisy windows with one clean reference sequence C. A window A_w starting at s_w reads only $C_{<s_w} \cup A_w$; reference rows are causal. Noisy copies remain isolated even when their source intervals overlap. Under the conditions in Appendix A.3, their predictions match isolated calls, and the loss is averaged over valid target tokens. This shares history computation across K training windows (Figure 3).

3.3 Diffusion-Compatible Losses

We seek geometric supervision that reflects the effects of motion-feature errors while preserving diffusion's conditional-mean velocity field (Equation (3)). Although training supervises sampled velocities, the loss must uniquely recover their conditional mean for every conditional target distribution.

Applying the classical mean-elicitation result (Banerjee et al., 2005; Gneiting, 2011), we obtain a necessary-and-sufficient criterion: under the regularity and integrability conditions in Appendix B, these losses are exactly target-first Bregman divergences, up to a target-only term:

$$\ell(y, a) = D_\varphi(y, a) + c(y), \qquad D_\varphi(y, a) = \varphi(y) - \varphi(a) - \nabla\varphi(a)^\top (y - a), \tag{5}$$

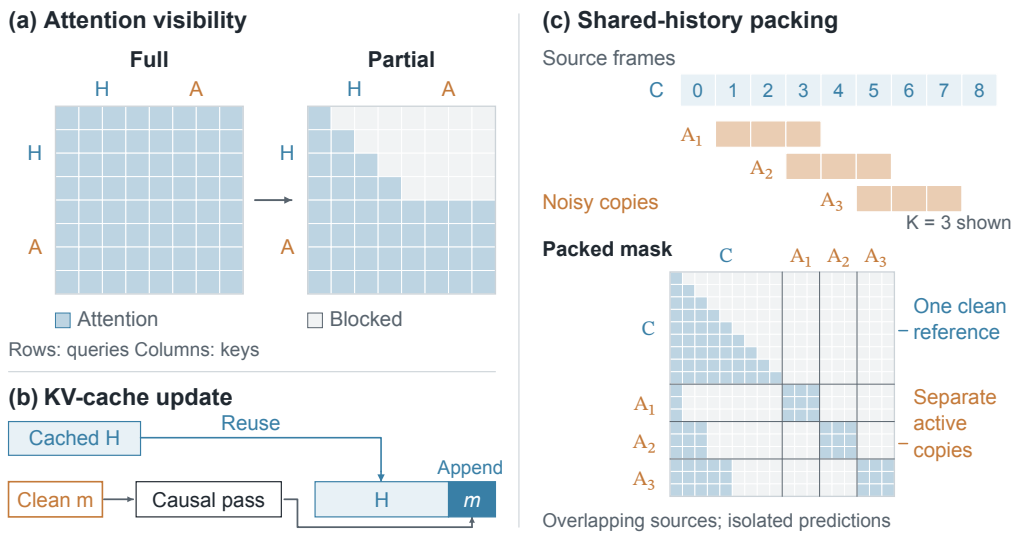

Figure 3: Partial Attention enables KV caching and shared-history training. (a) History queries read their causal prefixes, while active queries read the complete history and interact bidirectionally within the active window. (b) Newly finalized frames are re-encoded causally before their keys and values are appended to the cache. (c) Isolated noisy windows share one causal clean reference, reducing repeated history computation.

with strictly convex φ, sampled target y, and prediction a. Let X denote the predictor inputs, Y the sampled velocity, and $\mu = \mathbb{E}[Y \mid X]$. The conditional risk identity is

$$\mathbb{E}[\ell(Y, a) - \ell(Y, \mu) \mid X] = D_\varphi(\mu, a). \tag{6}$$

Strict convexity makes μ the unique optimum. Thus geometric weighting can change without shifting the required conditional mean. Figure 4 summarizes this criterion and our construction.

Direct nonlinear FK supervision. Let $F_{\boldsymbol{\eta}}$ decode a clean standardized state to geometric output, with $\boldsymbol{\eta}$ supplying the relative representation's root context. At an active frame, the predicted velocity gives $\widehat{\mathbf{z}} = \mathbf{x}_t + \beta_t \widehat{\mathbf{v}}$. Direct supervision with a fixed geometric metric $M \succeq 0$ would minimize

$$\ell_{\text{direct}}(\mathbf{v}, \widehat{\mathbf{v}}) = \|F_{\boldsymbol{\eta}}(\mathbf{z}) - F_{\boldsymbol{\eta}}(\mathbf{x}_t + \beta_t \widehat{\mathbf{v}})\|_M^2 . \tag{7}$$

For nonlinear $F_{\boldsymbol{\eta}}$, averaging geometric outputs generally differs from decoding the mean state. Consequently, this loss need not recover the conditional-mean velocity; Appendix B provides a counterexample. The controlled experiment in Appendix B.7 further shows that nonlinear FK supervision can bias sampling toward high-density regions, reducing diversity.

An FK-induced quadratic. We retain geometric sensitivity through a linear response in standardized coordinates. For the 138-D state, $J_{\boldsymbol{\eta},\mathbf{z}}$ is the neutral-SMPL-H surface-decoder Jacobian: a small state perturbation $\boldsymbol{\delta}$ produces local geometric error $\boldsymbol{\delta}^\top J_{\boldsymbol{\eta},\mathbf{z}}^\top M J_{\boldsymbol{\eta},\mathbf{z}} \boldsymbol{\delta}$, where $M \succeq 0$ weights vertices by surface area. For the 263-D features, $J_{\boldsymbol{\eta},\mathbf{z}}$ is a designed kinematic joint-response operator. We average and normalize the response Gram matrices using training data:

$$\overline{G} = \mathbb{E}_{(\boldsymbol{\eta},\mathbf{z})\sim p_{\text{train}}}[J_{\boldsymbol{\eta},\mathbf{z}}^\top M J_{\boldsymbol{\eta},\mathbf{z}}], \qquad W_{\text{FK}} = \frac{d\,\overline{G}}{\operatorname{tr}(\overline{G})}. \tag{8}$$

Here d is the number of predicted channels. The averaged matrix is independent of the current regression target; target-dependent per-example weights generally elicit a weighted mean. The clean-state residual is $\boldsymbol{\delta} = \beta_t \mathbf{e}$ for velocity residual $\mathbf{e} = \widehat{\mathbf{v}} - \mathbf{v}$. Omitting β_t^2 reweights time/frame contributions while preserving their conditional-mean optima. We use the fixed channel metric

$$W_\gamma = \frac{I_d + \gamma W_{\text{FK}}}{1 + \gamma}, \qquad \mathcal{L}_\gamma = \frac{1}{d}\mathbf{e}^\top W_\gamma \mathbf{e}, \qquad \gamma \geq 0. \tag{9}$$

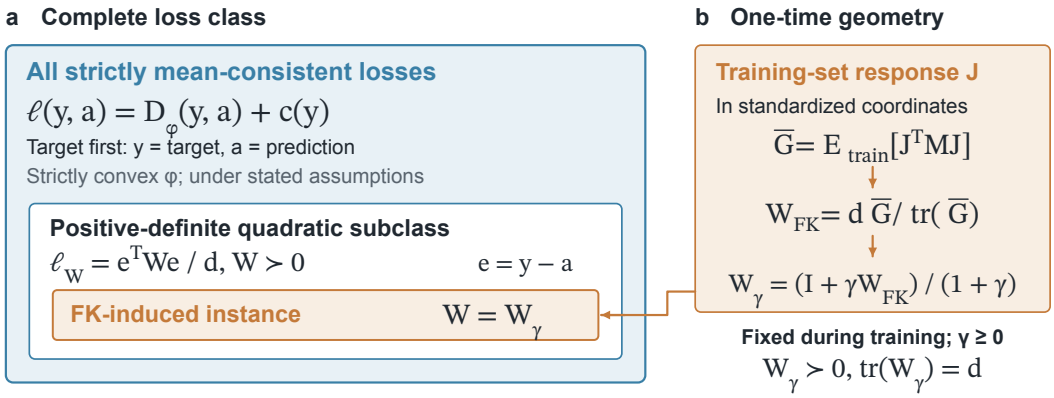

Figure 4: Diffusion-compatible losses and the FK-induced quadratic objective. (a) Under the stated regularity and integrability conditions, target-first Bregman divergences with strictly convex potentials, up to a target-only term, are necessary and sufficient for uniquely recovering the conditional mean. (b) Kinematic response Gram matrices estimated from training data are averaged and normalized, then combined with an identity term to obtain a fixed positive-definite metric, avoiding online FK evaluations.

The identity term makes W_γ positive definite, and trace normalization keeps its trace at d. The potential $\varphi_\gamma(\mathbf{u}) = \mathbf{u}^\top W_\gamma \mathbf{u}/d$ generates this quadratic as a target-first Bregman divergence, so Equation (6) guarantees the same conditional-mean velocity optimum. The matrix is estimated once and held fixed; training requires no online FK or Jacobians. FD2-Path restricts the response to its predicted channels before averaging and normalization. Appendices B.4 and C give the complete construction and decoder.

3.4 Root-Path Conditioning

FD2-Path, the path-conditioned variant of FD2, receives a frame-aligned planar root trajectory (x_k, z_k, ψ_k) together with text. We encode the trajectory into three relative root channels $\mathbf{c}_k$: heading increments and heading-local displacement. These channels remain clean and are concatenated with the noisy remaining state $\mathbf{x}_t^k$:

$$\widehat{\mathbf{v}}_t^k = f_\theta\big([\mathcal{N}_c(\mathbf{c}_k) \,\|\, \mathbf{x}_t^k], t, E_{\text{text}}(\mathbf{y}); \mathbf{A}^{\text{p}}\big)\,. \tag{10}$$

Here $\mathcal{N}_c$ standardizes the root conditions per channel and $\mathbf{y}$ denotes the text prompt. Only the remaining $d = 135$ (138-D) or 260 (263-D) channels are diffused and supervised. Sampling combines the prescribed root channels with the generated body state.

Center-of-mass root and relative body motion. For the 138-D SEED state, we separate the planar root from the articulated pelvis. Given the 22 body joints $\mathbf{J}_{k,i}$, the root is the ground projection of their fixed mass-weighted average:

$$\mathbf{q}_k = \Pi_{\text{ground}}\left(\sum_{i=0}^{21} w_i \mathbf{J}_{k,i}\right), \qquad \sum_i w_i = 1. \tag{11}$$

The weights follow biomechanical segment parameters (de Leva, 1996). To compare root definitions, we evaluate their smoothness on the HumanML3D training split: this root has 20.3% lower planar jerk and 9.6% lower curvature than a pelvis projection (Appendix Table 8). The pelvis retains its translation and 3D rotation relative to the root, allowing natural body motion around the commanded trajectory. Together with the 21 parent-local body rotations, these channels decode directly to neutral SMPL-H. The supplied root channels remain fixed during sampling. Tracking error is measured from the generated body's center-of-mass projection in Equation (11), which depends jointly on the predicted pelvis offset and articulation. Appendix C gives the weights, heading definition, and encoding conventions.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and training. HumanML3D uses the standard split and evaluator (Guo et al., 2022a), with FD1's preprocessing and duration handling. SEED (Bones Studio, 2026) uses the public Kimodo training partition and category-held-out Content test split with the v1.1 evaluator (NVIDIA, 2026). SEED R@3/FID use TMR-SOMA-RP-v1 embeddings; Skate measures contact-time foot velocity (cm/s). Appendix D.6 details conversion and evaluation.

Our 114M-parameter transformer uses frozen UMT5-XXL (Chung et al., 2023), $n_s = 30$, $K = 8$, and $\gamma = 1$ (0 without FK). On HumanML3D, FD2 uses the 60K checkpoint with CFG 4, and FD2-Path uses the 55K checkpoint with CFG 3. Both SEED models use 300K checkpoints with CFG 2. Appendix D.1 details the configuration.

4.2 ATTENTION AND EFFICIENCY

Attention masks. Partial Attention retains comparable FID to Full attention (0.058 vs. 0.057) and improves R@3 and MM-Dist; fully causal attention gives FID 0.096 (Table 1).

Table 1: Attention-mask ablation on HumanML3D with the same 263-D configuration, checkpoint budget, and $K = 1$. Diversity closer to real motion (9.503) is better.

Attention	R@3↑	FID↓	MM-Dist↓	Div.→
Full	0.810	**0.057**	2.887	9.579
Causal	0.793	0.096	3.005	9.369
Partial	**0.814**	0.058	**2.839**	**9.438**

Multi-window training. At 512 windows per step, $K = 8$ packing reduces training denoiser FLOPs by $4.6\times$ and GPU count from four to one (Table 2). One GPU also supports 1,024 windows. FD2 retains motion quality, which improves with FK supervision (Table 4).

Table 2: Multi-window training with 300 frames and 16 text tokens (rounded mean caption length). TFLOPs cover denoiser forward/backward; memory sums peak usage across H200 GPUs.

K	GPUs	Batch/GPU	Windows/step	TFLOPs↓	Total peak GPU GiB↓
1	4	128	512	90.666	164.27
4	1	128	512	29.827	54.27
8	1	64	512	19.846	36.61
16	1	32	512	**14.903**	**27.97**
8	1	128	1,024	39.691	72.37

Cached inference. At 4,500 history frames, caching reduces backbone latency from 26.002 to 2.303 ms per update ($11.29\times$) and computation by $137\times$ (Table 3).

Table 3: KV-cache inference on RTX 4090: batch 1, 30 active frames, 16 cached text tokens, CUDA Graph, PyTorch SDPA, no CFG. FLOPs and latency include the complete backbone update, causal re-encoding, and K/V writes for the newly finalized frame. Cache counts bf16 history K/V.

	Full, uncached		Partial, KV cached		Cache	Speedup
History	GFLOPs↓	ms↓	GFLOPs↓	ms↓	MiB	×
0	6.26	**1.810**	6.26	1.831	0	0.99
90	23.26	2.107	**6.53**	**1.953**	2.8	1.08
450	96.61	3.066	**6.90**	**1.962**	14.1	1.56
4500	1506.95	26.002	**11.01**	**2.303**	140.6	**11.29**

4.3 Motion Quality and Path Control

HumanML3D. FK lowers FD2's FID from 0.058 (FD1: 0.057) to 0.053, the best among compared streaming text-only methods (Table 4). FD2-Path leads the compared path-conditioned methods in R@3, FID, MM-Dist, and 3D pelvis error. Root channels are fixed by conditioning.

Table 4: HumanML3D text-only and path-conditioned generation: means with 95% confidence intervals; bold/underline denote best/second-best per condition. Text-only baselines and adapted PRIMAL follow FD1; path results are our evaluations. Root measures mean planar control-point error; 3D, mean pelvis error against ground truth (both cm). Skate: fraction of frames with foot sliding. Control targets: Appendix D.5.

Method	Stream	R@3↑	FID↓	MM-Dist↓	Div.→	Root↓	3D↓	Skate↓
Text-only								
Real motion	–	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	–	–	–
TM2T (Guo et al., 2022b)	✗	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$8.589^{\pm.076}$	–	–	–
T2M (Guo et al., 2022a)	✗	$0.736^{\pm.002}$	$1.087^{\pm.021}$	$3.347^{\pm.008}$	$9.175^{\pm.083}$	–	–	–
MDM (Tevet et al., 2023)	✗	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\underline{9.559}^{\pm.086}$	–	–	–
MLD (Chen et al., 2023)	✗	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	–	–	–
MotionDiffuse (Zhang et al., 2024)	✗	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	–	–	–
T2M-GPT (Zhang et al., 2023a)	✗	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$9.722^{\pm.082}$	–	–	–
ReMoDiffuse (Zhang et al., 2023b)	✗	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$2.974^{\pm.016}$	$9.018^{\pm.075}$	–	–	–
MoMask (Guo et al., 2024)	✗	$0.807^{\pm.002}$	$\mathbf{0.045}^{\pm.002}$	$2.958^{\pm.008}$	$9.677^{\pm.032}$	–	–	–
PRIMAL (Zhang et al., 2025)	✓	$0.780^{\pm.002}$	$0.511^{\pm.006}$	$3.120^{\pm.019}$	$\mathbf{9.520}^{\pm.068}$	–	–	–
MotionStreamer (Xiao et al., 2025)	✓	$0.802^{\pm.003}$	$0.092^{\pm.003}$	$2.909^{\pm.015}$	$9.722^{\pm.037}$	–	–	–
FloodDiffusion (Cai et al., 2026)	✓	$\underline{0.810}^{\pm.003}$	$0.057^{\pm.002}$	$\underline{2.887}^{\pm.007}$	$9.579^{\pm.062}$	–	–	–
FD2 w/o FK (ours)	✓	$0.807^{\pm.003}$	$0.058^{\pm.004}$	$2.903^{\pm.011}$	$9.405^{\pm.043}$	–	–	–
FD2 (ours)	✓	$\mathbf{0.815}^{\pm.003}$	$\underline{0.053}^{\pm.004}$	$\mathbf{2.832}^{\pm.012}$	$9.425^{\pm.065}$	–	–	–
Text + Path								
GMD (Karunratanakul et al., 2023)	✗	$0.777^{\pm.003}$	$0.133^{\pm.012}$	$3.129^{\pm.020}$	$\underline{9.445}^{\pm.078}$	$\underline{0.003}^{\pm.000}$	$7.19^{\pm.13}$	$0.144^{\pm.001}$
OmniControl (Xie et al., 2024)	✗	$\underline{0.804}^{\pm.001}$	$\underline{0.100}^{\pm.004}$	$\underline{2.928}^{\pm.012}$	$9.688^{\pm.007}$	$5.48^{\pm.09}$	$\underline{6.47}^{\pm.07}$	$\mathbf{0.071}^{\pm.000}$
DART (Zhao et al., 2025)	✓	$0.722^{\pm.003}$	$0.854^{\pm.008}$	$3.513^{\pm.019}$	$\mathbf{9.498}^{\pm.014}$	$7.26^{\pm.05}$	$12.40^{\pm.15}$	$0.130^{\pm.001}$
FD2-Path (ours)	✓	$\mathbf{0.816}^{\pm.003}$	$\mathbf{0.050}^{\pm.004}$	$\mathbf{2.829}^{\pm.010}$	$9.312^{\pm.069}$	$\mathbf{0.000}^{\pm.000}$	$\mathbf{5.23}^{\pm.08}$	$\underline{0.082}^{\pm.000}$

SEED. FD2 has the lowest streaming text-only FID and lower foot skate than ARDY/Kimodo, although its R@3 remains below ARDY (Table 5). FK supervision improves all three text-only metrics and reduces FD2-Path's tracking error from 0.18 to 0.11 cm and foot skate from 4.07 to 3.24 cm/s.

Table 5: SEED Content Overview with the v1.1 evaluator. Ground truth and Kimodo text-only R@3/FID are official (NVIDIA, 2026). Generated-motion Skate uses shared contact thresholds (parentheses: official predicted contacts). All path models receive the same 917 dense paths. Position pools conditioned frames: generated-body CoM for FD2-Path, pelvis for Kimodo/ARDY. Bold: best streaming result per condition. †ARDY uses RP01 training; FD2/Kimodo use the public SEED partition. Motion Data by Bones Studio.

					Position (cm)↓	
Method	Stream	R@3↑	FID↓	Skate↓	Mean	P95
Text-only						
Ground truth	–	89.09	0.000	1.85	–	–
Kimodo-SOMA-SEED-v1.1 (Rempe et al., 2026)	✗	81.13	0.035	3.34 (3.70)	–	–
ARDY (Zhao et al., 2026)†	✓	**76.34**	0.055	3.17 (4.43)	–	–
FloodDiffusion 2 w/o FK loss (ours)	✓	70.88	0.053	3.84	–	–
FloodDiffusion 2 (ours)	✓	73.12	**0.048**	**3.08**	–	–
Text + Path						
Kimodo-SOMA-SEED-v1.1 (Rempe et al., 2026)	✗	78.84	0.031	4.14 (5.14)	2.37	7.44
ARDY (Zhao et al., 2026)†	✓	75.03	0.052	3.63 (5.11)	1.60	3.91
FloodDiffusion 2-Path w/o FK loss (ours)	✓	75.07	0.036	4.07	0.18	0.55
FloodDiffusion 2-Path (ours)	✓	**76.01**	**0.034**	**3.24**	**0.11**	**0.32**

4.4 Qualitative Control

Figures 5 and 6 illustrate action and gait changes under text and path conditioning.

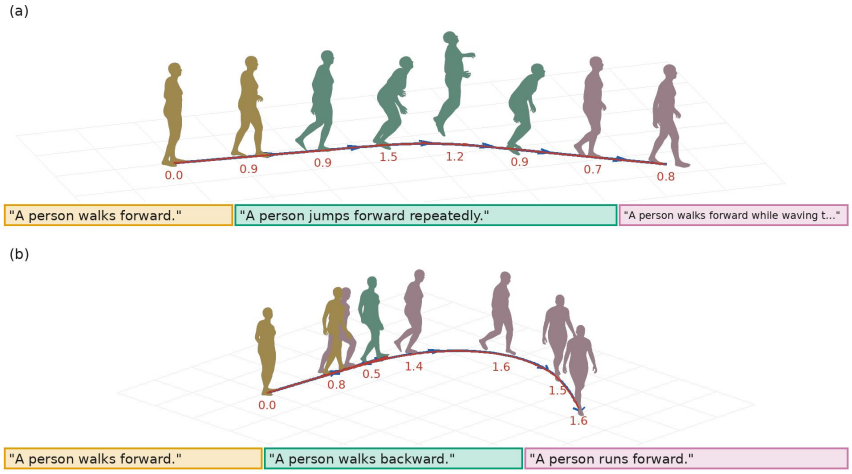

Figure 5: Prompt switching along a commanded path. Body colors indicate the active prompt. Blue is the commanded center-of-mass path, red the generated center-of-mass track; numbers show planar speed in m/s. The examples include repeated jumping and backward walking.

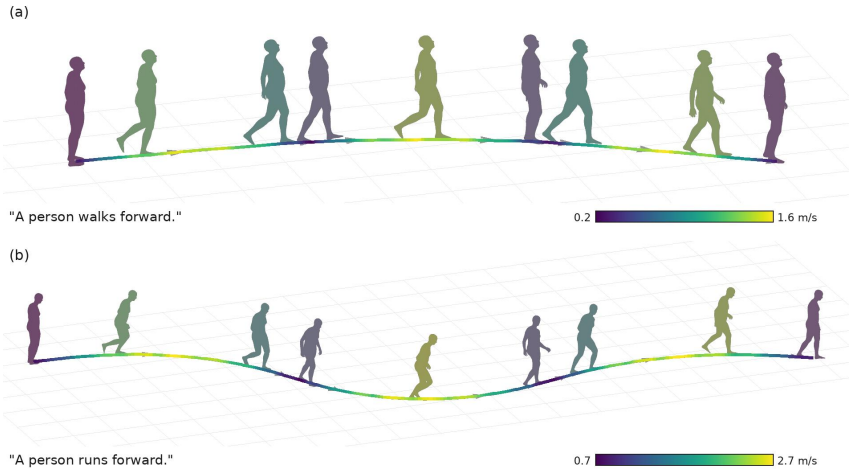

Figure 6: Speed control under fixed text prompts. The walking (a) and running (b) sequences follow varying commanded speed profiles. Body and trajectory colors indicate generated planar speed, illustrating gait adaptation along the commanded path while the action prompt remains unchanged.

5 Limitations and Conclusions

Conclusion. FD2 improves FD1 through Partial Attention for efficient training and inference, an FK-induced quadratic guided by the necessary-and-sufficient Bregman criterion, and root-path conditioning. Experiments on HumanML3D and SEED demonstrate improved motion quality, efficiency, and controllability.

Limitations. Finalized errors persist in history, and generation beyond the training horizon requires extrapolation. Cache memory and per-update attention cost grow linearly with history length. Our fixed, dataset-averaged FK quadratic leaves room for more adaptive mean-consistent geometric weighting. Motion and transition diversity are limited by training-data coverage, particularly for rare actions and unfamiliar combinations.

AI USE STATEMENT

Generative AI tools assisted with manuscript and citation checks, LaTeX editing, schematic artwork, and benchmarking code. The authors reviewed the final manuscript and take responsibility for its content.

REFERENCES

Arindam Banerjee, Xin Guo, and Hui Wang. On the Optimality of Conditional Expectation as a Bregman Predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005. doi: 10.1109/TIT.2005.850145.

Bones Studio. BONES-SEED: Skeletal Everyday Embodiment Dataset. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/bones-studio/seed. Motion Data by Bones Studio; accessed 2026-09-05.

Yiyi Cai, Yuhan Wu, Kunhang Li, You Zhou, Bo Zheng, and Haiyang Liu. FloodDiffusion: Tailored Diffusion Forcing for Streaming Motion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2295–2304, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cai_FloodDiffusion_Tailored_Diffusion_Forcing_for_Streaming_Motion_Generation_CVPR_2026_paper.html.

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*, volume 37, pp. 24081–24125. Curran Associates, Inc., 2024. doi: 10.52202/079017-0759. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2aee1c4159e48407d68fe16ae8e6e49e-Abstract-Conference.html.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your Commands via Motion Diffusion in Latent Space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010, 2023. doi: 10.1109/CVPR52729.2023.01726.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. UniMax: Fairer and More Effective Language Sampling for Large-Scale Multilingual Pretraining. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kXwdL1cWOAi.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285/.

Paolo de Leva. Adjustments to Zatsiorsky–Seluyanov's Segment Inertia Parameters. *Journal of Biomechanics*, 29(9):1223–1230, 1996. doi: 10.1016/0021-9290(95)00178-6.

Tilmann Gneiting. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138.

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*, pp. 5142–5151, 2022a. doi: 10.1109/CVPR52688.2022.00509.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *Computer Vision – ECCV 2022*, volume 13695 of *Lecture Notes in Computer Science*, pp. 580–597, Cham, 2022b. Springer Nature Switzerland. doi: 10.1007/978-3-031-19833-5_34. URL https://link.springer.com/chapter/10.1007/978-3-031-19833-5_34.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative Masked Modeling of 3D Human Motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024. doi: 10.1109/CVPR52733.2024.00186.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308. Curran Associates, Inc., 2025. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html. Main Conference Track.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided Motion Diffusion for Controllable Human Motion Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2151–2162, 2023. doi: 10.1109/ICCV51070.2023.00205.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling Forcing: Autoregressive Long Video Diffusion in Real Time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/935151cc6cb5d8b6816133b75233775a-Abstract-Conference.html.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of Motion Capture As Surface Shapes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5441–5450, 2019. doi: 10.1109/ICCV.2019.00554.

NVIDIA. Kimodo Results. Kimodo documentation, 2026. URL https://research.nvidia.com/labs/sil/projects/kimodo/docs/benchmark/results.html. Results for Kimodo-SOMA-SEED-v1.1 and Kimodo-SOMA-RP-v1.1; accessed 2026-09-05.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling Controllable Human Motion Generation. *arXiv preprint arXiv:2603.15546*, 2026. doi: 10.48550/arXiv.2603.15546. URL https://arxiv.org/abs/2603.15546.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. doi: 10.1145/3130800.3130883.

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. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. MotionStreamer: Streaming Motion Generation via Diffusion-Based Autoregressive Model in Causal Latent Space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10086–10096, 2025. doi: 10.1109/ICCV51701.2025.00940.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *International Conference on Learning Representations*, volume 2024, pp. 28176–28194, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7827290f07f63485b849b66cfa3e5dd0-Abstract-Conference.html.

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*, pp. 14730–14740, 2023a. doi: 10.1109/CVPR52729.2023.01415.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 364–373, 2023b. doi: 10.1109/ICCV51070.2023.00040.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-Driven Human Motion Generation With Diffusion Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024. doi: 10.1109/TPAMI.2024.3355414.

Yan Zhang, Yao Feng, Alpár Cseke, Nitin Saini, Nathan Bajandas, Nicolas Heron, and Michael J. Black. PRIMAL: Physically Reactive and Interactive Motor Model for Avatar Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12725–12736, 2025. doi: 10.1109/ICCV51701.2025.01182.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A Diffusion-Based Autoregressive Motion Model for Real-Time Text-Driven Motion Control. In *International Conference on Learning Representations*, volume 2025, pp. 23569–23592, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3b057de5a2e38bd8fa10201866c20dbf-Abstract-Conference.html.

Kaifeng Zhao, Mathis Petrovich, Haotian Zhang, Tingwu Wang, Siyu Tang, and Davis Rempe. ARDY: Autoregressive Diffusion with Hybrid Representation for Interactive Human Motion Generation. *ACM Transactions on Graphics*, 45(4):86:1–86:14, 2026. doi: 10.1145/3811284.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746. IEEE, 2019. doi: 10.1109/CVPR.2019.00589.

A Partial Attention

This section supports the Partial Attention construction in Section 3.2. We first recall FD1's streaming-locality result and examine the information available to active predictions. We then establish conditions for reusing finalized-history keys and values, followed by the equivalence between shared-history packing and isolated training windows.

A.1 Complete Active Information and Probability Flow

Notation follows FloodDiffusion (Cai et al., 2026). Let $\mathbf{c}^{0:T}$ collect the frame-aligned clean conditions of a sampling trajectory, and let $\mathbf{z} \in \mathbb{R}^{T\times d}$ be the clean diffused sequence, $\mathbf{z} \sim p_{\mathrm{data}}(\cdot \mid \mathbf{c}^{0:T})$, with d diffused channels per frame. With $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ independent of $(\mathbf{z}, \mathbf{c}^{0:T})$, the streaming interpolation for frame k is

$$\mathbf{x}_t^k = \alpha_t^k \mathbf{z}^k + \beta_t^k \boldsymbol{\epsilon}^k, \qquad \dot{\mathbf{x}}_t^k = \dot{\alpha}_t^k \mathbf{z}^k + \dot{\beta}_t^k \boldsymbol{\epsilon}^k. \tag{12}$$

The frame-dependent coefficients follow the lower-triangular schedule

$$\alpha_t^k = \mathrm{clamp}\big(t - \tfrac{k}{n_s},\, 0,\, 1\big), \qquad \beta_t^k = 1 - \alpha_t^k. \tag{13}$$

They induce marginals $p_t = \mathrm{Law}(\mathbf{x}_t \mid \mathbf{c}^{0:T})$. Away from the schedule knots, finalized history, the active window, and future are $H = [0, m(t))$, $A = [m(t), n(t))$, and $F = [n(t), T)$, where

$$\begin{aligned} m(t) &= \min(T, \max(0, \lfloor (t-1)n_s \rfloor + 1))\,, \\ n(t) &= \min(T, \lfloor t n_s \rfloor + 1)\,. \end{aligned} \tag{14}$$

Let $u_t(\mathbf{x}, \mathbf{c}^{0:T})$ be the marginal velocity field of this path.

Assumption A.1 (Causal Dependency; Assumption 3.7 of Cai et al., 2026)**.** *As in the streaming formulation of FloodDiffusion, a state prefix depends only on the corresponding condition prefix:*

$$p_{\mathrm{data}}(\mathbf{z}^{0:l} \mid \mathbf{c}^{0:T}) = p_{\mathrm{data}}(\mathbf{z}^{0:l} \mid \mathbf{c}^{0:l}), \qquad l < T. \tag{15}$$

Theorem A.2 (Streaming Locality; Theorem 3.8 of Cai et al., 2026)**.** *Under Assumption A.1 and the schedule of Equation* (13)*, away from the finitely many knots,*

$$u_t(\mathbf{x}_t, \mathbf{c}^{0:T}) = \big[\mathbf{0}_H,\ u_t^{m(t):n(t)}\big(\mathbf{x}_t^{0:n(t)}, \mathbf{c}^{0:n(t)}\big),\ \mathbf{0}_F\big]. \tag{16}$$

The field vanishes outside the active window, and its active block depends only on the processed prefix.

We abbreviate the function $u_t^{m(t):n(t)}$ as u_t^A.

The following propositions justify causal history encoding while retaining the complete active window as input.

Proposition A.3 (Causal history encoding can preserve complete prefix information)**.** *Fix t. There exist a causal history encoder $\phi = (\phi_0, \ldots, \phi_{m(t)-1})$, whose component ϕ_k reads only the rows $0{:}k{+}1$, and a measurable readout g with*

$$u_t^A\big(\mathbf{x}_t^{0:n(t)}, \mathbf{c}^{0:n(t)}\big) = g\big(\phi\big(\mathbf{x}_t^{0:m(t)}, \mathbf{c}^{0:m(t)}\big),\ \mathbf{x}_t^A, \mathbf{c}^A\big). \tag{17}$$

Since history rows are saturated, $\mathbf{x}_t^{0:m(t)} = \mathbf{z}^{0:m(t)}$: the encoder runs on the finalized clean prefix.

Proof. Take the causal projections $\phi_k(\mathbf{x}_t^{0:k+1}, \mathbf{c}^{0:k+1}) = (\mathbf{x}_t^k, \mathbf{c}^k)$. Their outputs and the active rows retain the entire processed prefix. If ρ restores its original arrangement, $g = u_t^A \circ \rho$. □

Proposition A.4 (Fully causal active predictions can be suboptimal)**.** *Fix a time t away from the schedule knots with two active rows $j < k$. There exists a data law satisfying Assumption A.1 under which every predictor whose output at row j reads only the rows $0{:}j{+}1$ has expected squared error exceeding that of the full-information conditional mean u_t^j by a positive constant independent of the predictor.*

Proof. Take all conditions to be deterministic. Let (z^j, z^k) be jointly Gaussian with zero means, positive variances σ_j^2, σ_k^2, and correlation $0 < |\rho| < 1$. All other clean coordinates are independent of this pair and of each other; all Gaussian noise coordinates are independent of each other and of the clean data. This law satisfies Assumption A.1. Both selected rows lie on the open ramp, and the other processed rows provide no additional information about their velocities. For this fixed t, write $\alpha_i = \alpha_t^i$ and $\beta_i = \beta_t^i$, with the same convention for their derivatives. By Gaussian regression, $u_t^j = \mathbb{E}[\dot{x}_t^j \mid x_t^j, x_t^k] = a\, x_t^j + b\, x_t^k$, with

$$b = \frac{\mathrm{Cov}(\dot{x}_t^j, x_t^k \mid x_t^j)}{\mathrm{Var}(x_t^k \mid x_t^j)}, \qquad \mathrm{Cov}(\dot{x}_t^j, x_t^k \mid x_t^j) = \frac{\alpha_k \rho \sigma_j \sigma_k\, \beta_j(\dot{\alpha}_j \beta_j - \dot{\beta}_j \alpha_j)}{\alpha_j^2 \sigma_j^2 + \beta_j^2}. \tag{18}$$

On the open ramp of Equation (13), $\dot{\alpha}_j \beta_j - \dot{\beta}_j \alpha_j = 1$, so $b \neq 0$. For a predictor f causal at row j, the additional rows in its prefix are independent of $(x_t^j, x_t^k, \dot{x}_t^j)$. Projection onto functions of that prefix therefore gives

$$\mathbb{E}\big[(f(\mathbf{x}_t^{0:j+1}) - \dot{x}_t^j)^2\big] - \mathbb{E}\big[(u_t^j - \dot{x}_t^j)^2\big] \geq b^2\, \mathrm{Var}(x_t^k \mid x_t^j) > 0. \tag{19}$$

For vector-valued frames, embed this construction in one coordinate and take all remaining coordinates to be independent. □

A.2 History-Prefix Invariance

Consider a layered computation with Partial Attention and row-wise normalization, residual, and feed-forward operations.

Lemma A.5 (Layer-wise suffix independence)**.** *Assume that the input embedding of each history row i depends only on rows $0{:}i{+}1$, and that conditioning operations, including cross-attention, read only the conditions attached to that row. Fix a history position j and two inputs that agree on every row $i \leq j$: state tokens, noise metadata, position IDs, and per-row conditions. Under Partial Attention, the hidden state and the layer-wise keys and values at j are identical at every layer in deterministic evaluation, whatever the later suffix. They also equal those of an inclusive causal execution with the same prefix inputs.*

Proof. Induction over layers. A history query $i \leq j$ reads only keys $0{:}i{+}1 \subseteq 0{:}j{+}1$, its cross-attention reads only the condition attached to row i, and normalization, residual, and feed-forward act row-wise; so equality of rows $0{:}j{+}1$ before a layer implies equality after it, and the input layer is equal by hypothesis. Inclusive causal attention has the same visible sets on these history rows. □

Thus finalized-history K/V can be reused while new rows are encoded; a prompt switch affecting only later rows preserves the cache.

A.3 Multi-Window Packing

For K copied windows $A_1, \ldots, A_K$ with start rows $s_1, \ldots, s_K$ over one reference sequence C, the packed mask is

$$\mathcal{V}(q) = \begin{cases} C_{\leq q}, & q \in C, \\ C_{<s_w} \cup A_w, & q \in A_w, \end{cases} \tag{20}$$

so no window reads another, even when their source ranges overlap.

Proposition A.6 (Packed/isolated equivalence)**.** *With untruncated history, deterministic evaluation (or coupled stochastic masks), and identical reference content, target content, metadata, and noise draw between each packed window and its isolated execution, every prediction on A_w equals its isolated prediction in exact arithmetic.*

Proof. Reference rows compute their states causally, unaffected by every packed target (Lemma A.5). A query in A_w therefore sees exactly the computation graph of its isolated call (Equation (20)) and has no path to another window; a layer-wise induction equates the hidden states on A_w. □

Let $\mathcal{I}_w \subseteq A_w$ contain the valid prediction rows of copy w, and let $N_w = |\mathcal{I}_w|$. For $N_w > 0$, define its mean per-row quadratic loss by

$$L_w = \frac{1}{N_w} \sum_{i \in \mathcal{I}_w} \frac{1}{d} \mathbf{e}_{w,i}^\top W_\gamma \mathbf{e}_{w,i},$$

where $\mathbf{e}_{w,i}$ is the prediction residual at row i of copy w, with W_γ from Equation (9). Windows with no valid prediction rows are omitted from the loss. Provided $\sum_{w=1}^{K} N_w > 0$, the packed objective is

$$L_{\text{packed}} = \frac{\sum_{w:N_w>0} N_w L_w}{\sum_{w=1}^{K} N_w}. \tag{21}$$

Thus packing preserves the isolated predictions and their token-count-weighted objective. The equivalence is conditional on the realized reference inputs. If reference perturbations are used during training, their states and noise metadata must also be shared by the packed and isolated executions.

B Diffusion-Compatible Regression Losses

This section expands Section 3.3. We first state the conditional-mean target of velocity regression and the necessary-and-sufficient Bregman criterion for preserving it. We then show why direct nonlinear geometric losses need not preserve this target, construct the fixed FK-induced quadratic, and give its two motion-state implementations. Matrix diagnostics and a controlled rotation experiment illustrate the construction and its effect on sampling.

B.1 Conditional Targets Train the Posterior Average

We write $\dot{\mathbf{x}}_t$ for the sampled velocity denoted by $\mathbf{v}_t$ in the main text. The active velocity is the conditional mean of the sampled interpolation target (Definition 3.3 of Cai et al., 2026):

$$u_t^A\big(\mathbf{x}_t^{0:n(t)}, \mathbf{c}^{0:n(t)}\big) = \mathbb{E}\big[\dot{\mathbf{x}}_t^A \,\big|\, \mathbf{x}_t^{0:n(t)}, \mathbf{c}^{0:n(t)}\big] \tag{22}$$

Fix t away from the schedule knots with $A \neq \varnothing$, and write $X_t = (\mathbf{x}_t^{0:n(t)}, \mathbf{c}^{0:n(t)})$. We apply the loss to each active row: $Y_k = \dot{\mathbf{x}}_t^k$ and its prediction $f_t^k(X_t)$ take values in an open convex set $\Omega \subseteq \mathbb{R}^d$, with d predicted channels per frame. The per-time risk is $\mathcal{R}_t(f) = \mathbb{E}[|A|^{-1} \sum_{k \in A} \ell(Y_k, f_t^k(X_t))]$.

Definition B.1 (Mean consistency)**.** *A loss $\ell(y, a)$ is* mean-consistent *on Ω if, for every finitely supported distribution P on Ω, its mean $\mu_P = \mathbb{E}_{y\sim P}[y] \in \Omega$ minimizes $R_P(a) = \mathbb{E}_{y\sim P}[\ell(y, a)]$. It is* strictly *mean-consistent if this minimizer is unique.*

Proposition B.2 (Strictly mean-consistent training recovers the posterior average)**.** *Let the risks be finite, and suppose that for each active row k, the conditional mean $\mathbb{E}[Y_k \mid X_t]$ uniquely minimizes its conditional expected loss under ℓ. Then u_t^A uniquely minimizes $\mathcal{R}_t$ up to null sets. Applying this result at almost every diffusion time recovers the streaming velocity field, whose zero extension reproduces the conditional data law under the sampling formulation of Cai et al. (2026) (Theorems 3.4 and 3.8).*

Proof. The tower property expresses the excess risk as the expected average of the per-row conditional excess risks. Each is nonnegative and vanishes only at $f_t^k(X_t) = u_t^k(X_t)$ by hypothesis and Equation (22). The sampling claim follows from Theorem A.2 and the conditional-generation theorem of Cai et al. (2026). □

Definition B.1 quantifies over finitely supported laws because they suffice for the necessity direction below. For a Bregman loss $\ell(Y, a) = D_\varphi(Y, a) + c(Y)$, the identity in Equation (29) extends the conclusion to laws whose mean lies in Ω and for which $\mathbb{E}\|Y\|$, $\mathbb{E}|\varphi(Y)|$, and $\mathbb{E}|c(Y)|$ are finite. For the quadratic losses below, a finite second moment suffices.

B.2 A Necessary and Sufficient Characterization

Lemma B.3 (Identification of a mean-consistent gradient). *Let $g : \Omega \times \Omega \to \mathbb{R}^d$ be continuous. Suppose that, for every $a \in \Omega$ and every finite collection $\{(\lambda_i, y_i)\}_{i=1}^n$ with $\lambda_i > 0$, $\sum_i \lambda_i = 1$, and $\sum_i \lambda_i y_i = a$, one has $\sum_i \lambda_i g(y_i, a) = 0$. Then there is a matrix-valued function $H(a)$ such that*

$$g(y, a) = H(a)(a - y), \qquad y, a \in \Omega. \tag{23}$$

Proof. Fix a and write $h(v) = g(a + v, a)$ on the open convex set $V = \Omega - a$, which contains the origin. The one-point distribution gives $h(0) = 0$. On a sufficiently small ball $B(0, r) \subset V$, a two-point zero-mean distribution supported on v and $-v$ gives $h(-v) = -h(v)$. Whenever $v, w, v + w$ and their negatives lie in that ball, the equal-weight distribution on $v, w, -(v + w)$ gives $h(v) + h(w) + h(-(v + w)) = 0$, hence $h(v + w) = h(v) + h(w)$. Repeated addition gives rational homogeneity; continuity gives real homogeneity. Therefore $h(v) = A(a)v$ near zero for a linear map $A(a)$.

For arbitrary $v \in V$, choose $\rho > 0$ small enough that $-\rho v \in B(0, r)$. The distribution assigning masses $\rho/(1+\rho)$ and $1/(1+\rho)$ to v and $-\rho v$ has mean zero, so $\rho h(v)+h(-\rho v) = 0$. Local linearity makes the second term $-\rho A(a)v$, hence $h(v) = A(a)v$ throughout V. Setting $H(a) = -A(a)$ proves Equation (23). □

The theorem below states the smooth multivariate form of the classical mean-consistency criterion (Banerjee et al., 2005; Gneiting, 2011).

Theorem B.4 (Necessary and sufficient loss characterization). *Let $\ell : \Omega \times \Omega \to \mathbb{R}$ be three times continuously differentiable. It is mean-consistent for every finitely supported law whose support and mean lie in Ω if and only if*

$$\ell(y, a) = D_\varphi(y, a) + c(y), \qquad D_\varphi(y, a) = \varphi(y) - \varphi(a) - \nabla\varphi(a)^\top (y - a), \tag{24}$$

for a convex potential $\varphi \in C^4(\Omega)$ and a target-only function $c \in C^3(\Omega)$. The loss is strictly mean-consistent if and only if φ is strictly convex.

Proof. Necessity. Let $g(y, a) = \nabla_a \ell(y, a)$. Because Ω is open, mean consistency for every finitely supported law P with mean a gives $\mathbb{E}_{y\sim P}[g(y, a)] = 0$. Lemma B.3 therefore yields $g(y, a) = H(a)(a-y)$. For fixed $y_0 \in \Omega$ and sufficiently small $h > 0$, $H(a)e_j = [g(y_0, a)-g(y_0+he_j, a)]/h$ for each coordinate basis vector e_j; hence $H \in C^2$ since $g \in C^2$. For the law degenerate at a, the action a minimizes $\ell(a, \cdot)$, so

$$H(a) = \nabla_a^2 \ell(y, a)\big|_{y=a} \succeq 0;$$

Thus H is symmetric. Differentiating g and using symmetry of the action Hessian gives

$$\sum_k \big(\partial_j H_{ik}(a) - \partial_i H_{jk}(a)\big)(a_k - y_k) = 0.$$

This holds for every y in the open set Ω, so each coefficient vanishes: for every i, j, k,

$$\partial_j H_{ik}(a) = \partial_i H_{jk}(a). \tag{25}$$

Together with $H_{jk} = H_{kj}$, this implies $\partial_k H_{ij} = \partial_j H_{ik}$, the row-wise integrability condition.

Fix $a_0 \in \Omega$, put $r = a - a_0$, and define

$$p_i(a) = \int_0^1 \sum_j H_{ij}(a_0 + tr) r_j \, dt, \tag{26}$$

$$\varphi(a) = \int_0^1 p(a_0 + tr)^\top r \, dt. \tag{27}$$

Differentiating the first integral and using $\partial_k H_{ij} = \partial_j H_{ik}$ gives

$$\partial_k p_i(a) = \int_0^1 \left[t \frac{d}{dt} H_{ik}(a_0 + tr) + H_{ik}(a_0 + tr) \right] dt = H_{ik}(a).$$

Since H is symmetric, the same calculation for the second integral gives $\partial_k \varphi(a) = p_k(a)$ and hence $\nabla^2 \varphi(a) = H(a) \succeq 0$. Thus φ is convex and lies in C^4 because $H \in C^2$. Direct differentiation gives

$$\nabla_a D_\varphi(y, a) = \nabla^2 \varphi(a)(a - y) = g(y, a).$$

Consequently $\nabla_a[\ell(y, a) - D_\varphi(y, a)] = 0$. Since Ω is connected, the bracket equals $c(y) = \ell(y, y) \in C^3$, proving Equation (24).

Sufficiency. For any law P with mean μ_P for which the expectations exist,

$$\mathbb{E}_{y\sim P}[D_\varphi(y, a)] = \mathbb{E}_{y\sim P}[\varphi(y)] - \varphi(a) - \nabla\varphi(a)^\top(\mu_P - a), \tag{28}$$

$$\mathbb{E}_{y\sim P}[D_\varphi(y, a)] - \mathbb{E}_{y\sim P}[D_\varphi(y, \mu_P)] = D_\varphi(\mu_P, a) \geq 0. \tag{29}$$

The target-only term $c(y)$ cancels. Thus the mean is a minimizer, and strict convexity of φ makes it unique. Conversely, strict mean consistency for every degenerate law requires $D_\varphi(y, a) > 0$ whenever $y \neq a$, which is the strict supporting-plane characterization of differentiable strict convexity. □

The argument order in Equation (24) is essential: the random diffusion target is the first argument and the network prediction is the second.

B.3 Direct Nonlinear Losses Need Not Be Mean-Consistent

The model predicts velocity rather than a clean state, so a loss through a nonlinear map of the clean state must first reconstruct that state. For one active token, write

$$\mathbf{x}_t = \alpha_t \mathbf{z} + \beta_t \boldsymbol{\epsilon}, \qquad \dot{\mathbf{x}}_t = \dot{\alpha}_t \mathbf{z} + \dot{\beta}_t \boldsymbol{\epsilon}, \qquad \Delta_t = \alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t. \tag{30}$$

At an active interior diffusion time with $\Delta_t \neq 0$ and $\beta_t \neq 0$, a proposed velocity a implies the clean-state estimate

$$R_t(a) = \frac{\dot{\beta}_t \mathbf{x}_t - \beta_t a}{\Delta_t}, \qquad R_t(\dot{\mathbf{x}}_t) = \mathbf{z}. \tag{31}$$

For our linear active ramp, $\dot{\alpha}_t = 1$ and $\dot{\beta}_t = -1$, so $\Delta_t = -1$ and $R_t(a) = \mathbf{x}_t + \beta_t a$, as in the main text. Let $\boldsymbol{\eta}$ be a clean decoder context from the processed prefix, and let $\mathcal{S}_{\boldsymbol{\eta}} : R_t(\Omega) \to \mathbb{R}^m$ be any differentiable nonlinear map from the clean state to a geometric output. Define the nonlinear composite $\mathcal{K}_t = \mathcal{S}_{\boldsymbol{\eta}} \circ R_t$. The geometric decoder is denoted by $F_{\boldsymbol{\eta}}$ in the main text.

Proposition B.5 (A direct nonlinear loss is not generally mean-consistent)**.** *Let $\mathcal{K}_t : \Omega \to \mathbb{R}^m$ be the differentiable nonlinear map above and let $M \succeq 0$ be a fixed output metric. Define $\ell_{\mathrm{dir}}(y, a) = \|\mathcal{K}_t(y) - \mathcal{K}_t(a)\|_M^2$, where $\|v\|_M^2 = v^\top M v$. For a target law P on Ω with finite mean $\mu_P \in \Omega$ and $\mathbb{E}_{y\sim P}\|\mathcal{K}_t(y)\|_2^2 < \infty$,*

$$\nabla_a \, \mathbb{E}_{y\sim P}[\ell_{\mathrm{dir}}(y, a)] = 2 J_{\mathcal{K}_t}(a)^\top M \left(\mathcal{K}_t(a) - \mathbb{E}_{y\sim P}[\mathcal{K}_t(y)]\right), \tag{32}$$

which need not vanish at $a = \mu_P$.

Proof. The moment assumption makes the risk finite for every $a \in \Omega$. Expanding the square gives

$$\begin{aligned} R_P(a) = \mathbb{E}_{y\sim P}[\mathcal{K}_t(y)^\top M \mathcal{K}_t(y)] \\ - 2\mathcal{K}_t(a)^\top M \mathbb{E}_{y\sim P}[\mathcal{K}_t(y)] + \mathcal{K}_t(a)^\top M \mathcal{K}_t(a). \end{aligned}$$

Differentiating this finite expression with respect to a gives Equation (32). A single point rigidly rotated in a plane already makes the failure explicit. The nonzero affine scale in Equation (31) can be absorbed into the scalar angle coordinate. Let

$$K(\theta) = (\cos\theta, \sin\theta), \qquad M = I_2,$$

and let P put mass $\frac{3}{4}$ at 0 and $\frac{1}{4}$ at $\frac{\pi}{2}$, so $\mu_P = \pi/8$ and $\mathbb{E}_{y\sim P}[K(y)] = (3/4, 1/4)$. Since $J_K(\mu_P) = (-\sin\mu_P, \cos\mu_P)^\top$, the derivative of the direct expected loss at the mean equals

$$2J_K(\mu_P)^\top (K(\mu_P) - \mathbb{E}_{y\sim P}[K(y)]) = \tfrac{1}{2}\left(3\sin\tfrac{\pi}{8} - \cos\tfrac{\pi}{8}\right) \neq 0.$$

Adding this loss with any positive coefficient to a mean-consistent quadratic leaves a nonzero gradient at μ_P, so the sum is not generally mean-consistent. □

B.4 FK-Induced Quadratic Objective

The dataset-averaged response Gram matrix. Here $\boldsymbol{\xi}$ denotes the standardized clean state written as $\mathbf{z}$ in the main text. Let $J_{\boldsymbol{\eta},\boldsymbol{\xi}} \in \mathbb{R}^{m\times d}$ assign a linear response into an output space with fixed metric $M \succeq 0$ to each clean context $\boldsymbol{\eta}$ and state $\boldsymbol{\xi}$, expressed in the standardized training coordinate. The 263-D features use a designed kinematic joint response. For the 138-D state, J is the Jacobian of the surface decoder $F_{\boldsymbol{\eta}}$, $\partial F_{\boldsymbol{\eta}}(\boldsymbol{\xi})/\partial\boldsymbol{\xi}$. Assume $0 < \mathbb{E}\|M^{1/2}J_{\boldsymbol{\eta},\boldsymbol{\xi}}\|_F^2 < \infty$. Define the dataset-averaged Gram matrix and its normalization

$$\overline{G} = \mathbb{E}_{(\boldsymbol{\eta},\boldsymbol{\xi})}\big[J_{\boldsymbol{\eta},\boldsymbol{\xi}}^\top M J_{\boldsymbol{\eta},\boldsymbol{\xi}}\big], \qquad W_{\mathrm{FK}} = \frac{d\,\overline{G}}{\mathrm{tr}(\overline{G})}, \qquad W_\gamma = \frac{I_d + \gamma W_{\mathrm{FK}}}{1+\gamma}, \qquad \gamma \geq 0, \tag{33}$$

where the expectation runs over the training entries used for matrix estimation. Every summand is PSD, and $\mathrm{tr}(\overline{G}) = \mathbb{E}\|M^{1/2}J_{\boldsymbol{\eta},\boldsymbol{\xi}}\|_F^2 \in (0,\infty)$, so $W_\gamma \succeq (1+\gamma)^{-1}I_d \succ 0$ and $\mathrm{tr}(W_{\mathrm{FK}}) = \mathrm{tr}(W_\gamma) = d$. Because $\boldsymbol{\xi}$ is the standardized training coordinate, the response composes with $\mathbf{q} = \boldsymbol{\mu}_{\mathrm{ch}} + \boldsymbol{\sigma}_{\mathrm{ch}} \odot \boldsymbol{\xi}$, so every column already contains the corresponding channel standard deviation.

Theorem B.6 (The FK-induced quadratic preserves the diffusion field)**.** *Let W_γ be the deterministic matrix in Equation* (33)*. For every $\gamma \geq 0$, the loss*

$$\ell_\gamma(y,a) = \frac{1}{d}(y-a)^\top W_\gamma (y-a) \tag{34}$$

is a strictly mean-consistent target-first Bregman divergence: under every target law with finite second moment its unique optimum is that law's mean. Together with Proposition B.2, its population solution yields the original probability path and terminal distribution.

Proof. The positive-definite matrix W_γ defines the strictly convex potential

$$\varphi_\gamma(u) = \frac{1}{d}u^\top W_\gamma u, \qquad \nabla\varphi_\gamma(u) = \frac{2}{d}W_\gamma u, \qquad \nabla^2\varphi_\gamma(u) = \frac{2}{d}W_\gamma \succ 0. \tag{35}$$

Substitution into Equation (24) gives $D_{\varphi_\gamma}(y,a) = \ell_\gamma(y,a)$, and Equation (29) yields

$$\mathbb{E}_{y\sim P}[\ell_\gamma(y,a)] = \mathbb{E}_{y\sim P}[\ell_\gamma(y,\mu_P)] + \frac{1}{d}(a-\mu_P)^\top W_\gamma (a-\mu_P). \tag{36}$$

The final term vanishes only at $a = \mu_P$. Proposition B.2 then gives the sampling statement. □

When J is the Jacobian of a decoder $F_{\boldsymbol{\eta}}$, the local Taylor expansion

$$\|F_{\boldsymbol{\eta}}(\boldsymbol{\xi}+\delta) - F_{\boldsymbol{\eta}}(\boldsymbol{\xi})\|_M^2 = \delta^\top J_{\boldsymbol{\eta},\boldsymbol{\xi}}^\top M J_{\boldsymbol{\eta},\boldsymbol{\xi}}\delta + o(\|\delta\|_2^2) \tag{37}$$

shows the local output geometry retained by W_{FK}. Since $D_a R_t = -\beta_t I_d/\Delta_t$, mapping a velocity residual back to clean coordinates scales this metric by β_t^2/Δ_t^2. We omit this positive, target-independent factor, reweighting time/frame contributions while preserving their conditional-mean optima.

A target-dependent per-sample matrix is not interchangeable with Equation (33). For $\ell(y,a) = (y-a)^\top W(y)(y-a)$ and $Y \sim P$ with finite mean, let W be measurable with $W(y) \succeq 0$, $\mathbb{E}[\|W(Y)\|_F(1+\|Y\|_2^2)] < \infty$, and $\mathbb{E}[W(Y)] \succ 0$. Minimizing over $a \in \mathbb{R}^d$ gives

$$\mathbb{E}_{y\sim P}[W(y)]\, a = \mathbb{E}_{y\sim P}[W(y)\, y], \tag{38}$$

Thus the unique minimizer is $a^* = (\mathbb{E}[W(Y)])^{-1}\mathbb{E}[W(Y)Y]$, which generally differs from μ_P. Matrices $W(X) \succ 0$ depending only on predictor inputs preserve $\mathbb{E}[Y \mid X]$. The offline average in Equation (33) avoids FK and Jacobian evaluations during training.

B.5 Response Operators for the Two Motion States

The common construction uses different geometric responses for the two motion states: a surface-decoder Jacobian for 138-D SEED motion and a designed joint-response operator for native 263-D HumanML3D features. Both are expressed in standardized predicted coordinates and averaged offline before training. The path-conditioned branches restrict the raw Gram matrix to predicted channels before trace normalization.

Surface instantiation: the 138-D state. Here $F_{\boldsymbol{\eta}}$ is neutral-SMPL-H mesh FK with N_v template vertices. Let $\mathcal{T}$ be the triangle set of the neutral-shape template, and let $A_f^{(0)}$ be the rest-template area of triangle f. The lumped and normalized area assigned to vertex v is

$$\widetilde{a}_v = \frac{1}{3}\sum_{f\in\mathcal{T}:\, v\in f} A_f^{(0)}, \qquad a_v = \frac{\widetilde{a}_v}{\sum_{u=1}^{N_v}\widetilde{a}_u}, \qquad \sum_{v=1}^{N_v} a_v = 1, \tag{39}$$

and the area-weighted squared vertex-displacement metric is

$$M = \mathrm{diag}(a_1,\ldots,a_{N_v})\otimes I_3, \qquad \|\delta\mathbf{v}\|_M^2 = \sum_{v=1}^{N_v} a_v\|\delta\mathbf{v}_v\|_2^2. \tag{40}$$

The metric uses fixed rest-template areas and vertex correspondences. Area weighting approximates mean-squared surface displacement, and unit-sum normalization fixes the objective scale with $\mathrm{tr}(M) = 3$.

Let $\mathcal{N}_m$ be the fixed channel-wise standardization of the physical state $\mathbf{m} = [\mathbf{c}\|\mathbf{s}]$. Its restrictions to the root and articulated channels are $\mathcal{N}_c$ and $\mathcal{N}_s$, respectively: $\mathcal{N}_m(\mathbf{m}) = [\mathcal{N}_c(\mathbf{c})\|\mathcal{N}_s(\mathbf{s})]$. Let $b \in \{\mathrm{FD2}, \mathrm{Path}\}$ index the text-only and path-conditioned branches, with $d_{\mathrm{FD2}} = 138$ and $d_{\mathrm{Path}} = 135$. For valid frame i of training motion c, let $h_{c,i}$ contain the teacher-forced predecessor root position and heading (or the initial pose at the first frame), and define

$$(\boldsymbol{\xi}_{c,i}^{(b)}, \boldsymbol{\eta}_{c,i}^{(b)}, d_b) = \begin{cases} (\mathcal{N}_m(\mathbf{m}_{c,i}), h_{c,i}, 138), & b = \mathrm{FD2}, \\ (\mathcal{N}_s(\mathbf{s}_{c,i}), (h_{c,i}, \mathbf{c}_{c,i}), 135), & b = \mathrm{Path}. \end{cases} \tag{41}$$

The context $\boldsymbol{\eta}_{c,i}^{(b)}$ remains in physical coordinates. The FK function $F_{\boldsymbol{\eta}_{c,i}^{(b)}}(\boldsymbol{\xi}_{c,i}^{(b)}) \in \mathbb{R}^{3N_v}$ includes inverse channel normalization, root–pelvis reconstruction, the neutral-shape SMPL-H pose correctives, and linear blend skinning (LBS). Differentiating its final world-space vertices only with respect to the predicted block gives

$$J_{c,i}^{(b)} = \frac{\partial F_{\boldsymbol{\eta}_{c,i}^{(b)}}(\boldsymbol{\xi}_{c,i}^{(b)})}{\partial \boldsymbol{\xi}_{c,i}^{(b)}} \in \mathbb{R}^{3N_v\times d_b}. \tag{42}$$

Let $\mathcal{C}_{\mathrm{train}}$ be the calibration set and let $\mathcal{I}_c$ contain its selected valid non-initial frames for entry c. For the full-training-set HumanML3D calibration reported in Table 6, these are all valid non-initial frames from the 21,466 listed training entries. The average is

$$\overline{G}_{\mathrm{FK}}^{(b)} = \frac{1}{\sum_{c\in\mathcal{C}_{\mathrm{train}}}|\mathcal{I}_c|}\sum_{c\in\mathcal{C}_{\mathrm{train}}}\sum_{i\in\mathcal{I}_c}(J_{c,i}^{(b)})^\top M J_{c,i}^{(b)} \succeq 0. \tag{43}$$

Each branch is normalized independently through Equation (33) with $d = d_b$, yielding $W_{\mathrm{FK}}^{(b)}$ and $W_\gamma^{(b)}$.

For text-only FD2, the Jacobian in Equation (42) contains all 138 columns. In particular, the planar root displacement moves every surface vertex, while the heading increment changes both the displacement direction and the orientation of the complete surface. For FD2-Path, the predecessor context and current root channels are fixed, and the Jacobian contains only the predicted 135-D articulated block. With the same calibration poses, this is equivalently the articulated principal block of the *raw* text-only pullback,

$$\overline{G}_{\mathrm{FK}}^{(\mathrm{Path})} = \left(\overline{G}_{\mathrm{FK}}^{(\mathrm{FD2})}\right)_{\mathrm{art,art}}. \tag{44}$$

It is then trace-normalized with $d_{\mathrm{Path}} = 135$, independently of the 138-D branch.

Joint instantiation: the 263-D features. Here the prediction is the native 263-D feature frame, and the response is a designed per-frame operator $J_{\mathrm{kin}} \in \mathbb{R}^{66\times 263}$ into the 22 joint centers, under the mean-squared joint-displacement metric $M = \frac{1}{22}I_{66}$. It maps perturbations of the root channels and the 21 stored root-relative joint positions $\mathbf{p}_j$ to articulated joint displacements. The root yaw channel rotates the complete root-relative skeleton about the vertical axis; the planar displacement

and height channels translate every joint rigidly. A perturbation $\delta\mathbf{p}_j$ of stored joint j induces the minimum swing rotation of its incoming bone $\mathbf{b}_j = \mathbf{p}_j - \mathbf{p}_{\mathrm{pa}(j)}$,

$$\delta\boldsymbol{\omega}_j = \frac{\mathbf{b}_j \times \delta\mathbf{p}_j}{\|\mathbf{b}_j\|_2^2}, \qquad \delta\mathbf{p}_k^{\mathrm{out}} = \delta\boldsymbol{\omega}_j \times \big(\mathbf{p}_k - \mathbf{p}_{\mathrm{pa}(j)}\big) \quad \text{for every } k \text{ in the subtree of } j, \tag{45}$$

where $\mathrm{pa}(j)$ is the parent of j. The output response retains the component of $\delta\mathbf{p}_j$ perpendicular to $\mathbf{b}_j$ and propagates the induced swing through the subtree. A degenerate bone contributes zero response for that joint. Velocity, rotation, and contact channels receive no response, so their columns vanish and the dataset-averaged Gram matrix has rank at most 67; the identity term in W_γ restores positive definiteness. The average assigns equal weight to every frame across 23,384 training entries, including the additional HumanAct data, and Equation (33) normalizes with $d = 263$. The path-conditioned variant removes the three conditioned root channels—yaw and planar displacement, while the root height stays diffused—as the principal block $[3{:}263)$ of the raw Gram matrix, renormalized with $d = 260$.

B.6 Spectral Structure of the FK Matrices

The diagnostics in this subsection use the full-training-set HumanML3D calibrations of the 138-D and 263-D representations. They describe the geometric response matrices for these calibration data.

For a normalized FK matrix W_{FK}, define the effective rank $r_{\mathrm{eff}} = \exp(-\sum_j p_j \log p_j)$, where $p_j = \lambda_j(W_{\mathrm{FK}})/\operatorname{tr}(W_{\mathrm{FK}})$ over the nonzero PSD spectrum, and the off-diagonal ratio $r_{\mathrm{off}} = \|W_{\mathrm{FK}} - \operatorname{diag}(W_{\mathrm{FK}})\|_F / \|W_{\mathrm{FK}}\|_F$.

Table 6: Spectral properties of the HumanML3D-calibrated FK matrices. Values describe W_{FK} before forming W_γ; by construction $\operatorname{tr}(W_{\mathrm{FK}}) = d$. Numerical rank counts eigenvalues above $10^{-8}\lambda_{\max}$. Root share is the fraction of trace in the three path-conditioned root channels.

Branch	d	$\operatorname{tr}(\overline{G}_{\mathrm{FK}})$	Numerical rank	$\lambda_{\min}(W_{\mathrm{FK}})$	$\lambda_{\max}(W_{\mathrm{FK}})$	r_{eff}	r_{off}	Root trace share
Kinematic response, FD2	263	0.59209720	67/263	0	103.9970	8.2594	0.6302	0.098%
Kinematic response, FD2-Path	260	0.59151672	64/260	0	102.9012	8.2405	0.6300	–
Surface pullback, FD2	138	0.03065992	136/138	0	22.8292	12.9788	0.5543	0.842%
Surface pullback, FD2-Path	135	0.03040161	135/135	2.280×10^{-6}	22.0458	13.0258	0.5405	–

The 138-D surface matrix has off-diagonal ratio 55.43% (Figure 7); 99.11% of the off-diagonal squared energy couples different semantic channel groups rather than coordinates within one 6D rotation block. Its ten largest eigenvalues carry 90% of the trace. The two FD2 zero modes arise because, within one frame, planar root displacement can cancel the oppositely scaled planar pelvis offset. The identity term in Equation (33) removes this gauge. Fixing the root path gives the 135-D matrix full numerical rank. At $\gamma = 1$, $W_\gamma = (I + W_{\mathrm{FK}})/2$ has condition numbers 23.8292 and 23.0458 for FD2 and FD2-Path, respectively, and trace d_b.

B.7 Effect of FK Supervision on the Sampling Distribution

We compare pure nonlinear FK, pure FK-induced quadratic, and MSE supervision in a controlled 6-D rotation experiment. The dataset contains 8,192 equally weighted rotations distributed across eight regions. Their centers are the identity, $+90^\circ$ rotations about the x, y, and z axes, -90° rotations about the same axes, and a $100\sqrt{2}^\circ$ rotation about $(1,1,0)^\top/\sqrt{2}$, in that order. The region sizes are $(2458, 1720, 1311, 983, 655, 491, 328, 246)$. Each rotation is independently perturbed by left-multiplying its center by an isotropic Gaussian rotation-vector perturbation, with per-component standard deviations $(4,5,6,7,9,11,13,16)^\circ$, respectively. Thus region 1 has the highest density, while all entries have the same sampling probability. Rotations are encoded by their first two columns. The FK map $F : \mathbb{R}^6 \to \mathbb{R}^9$ orthonormalizes these columns and returns the positions of three orthogonal unit-axis endpoints.

For $\mathbf{x}_t = t\mathbf{z} + (1-t)\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I_6)$, let a be the predicted velocity, $y = \mathbf{z} - \boldsymbol{\epsilon}$, and $\widehat{\mathbf{z}} = \mathbf{x}_t + (1-t)a$. The three standalone losses are

$$\begin{aligned} \ell_{\mathrm{FK}}(y,a) &= \tfrac{1}{9}\|F(\widehat{\mathbf{z}}) - F(\mathbf{z})\|_2^2, \\ \ell_{\mathrm{quad}}(y,a) &= \tfrac{1}{6}(a-y)^\top W_{\mathrm{FK}}(a-y), \qquad \ell_{\mathrm{MSE}}(y,a) = \tfrac{1}{6}\|a-y\|_2^2. \end{aligned} \tag{46}$$

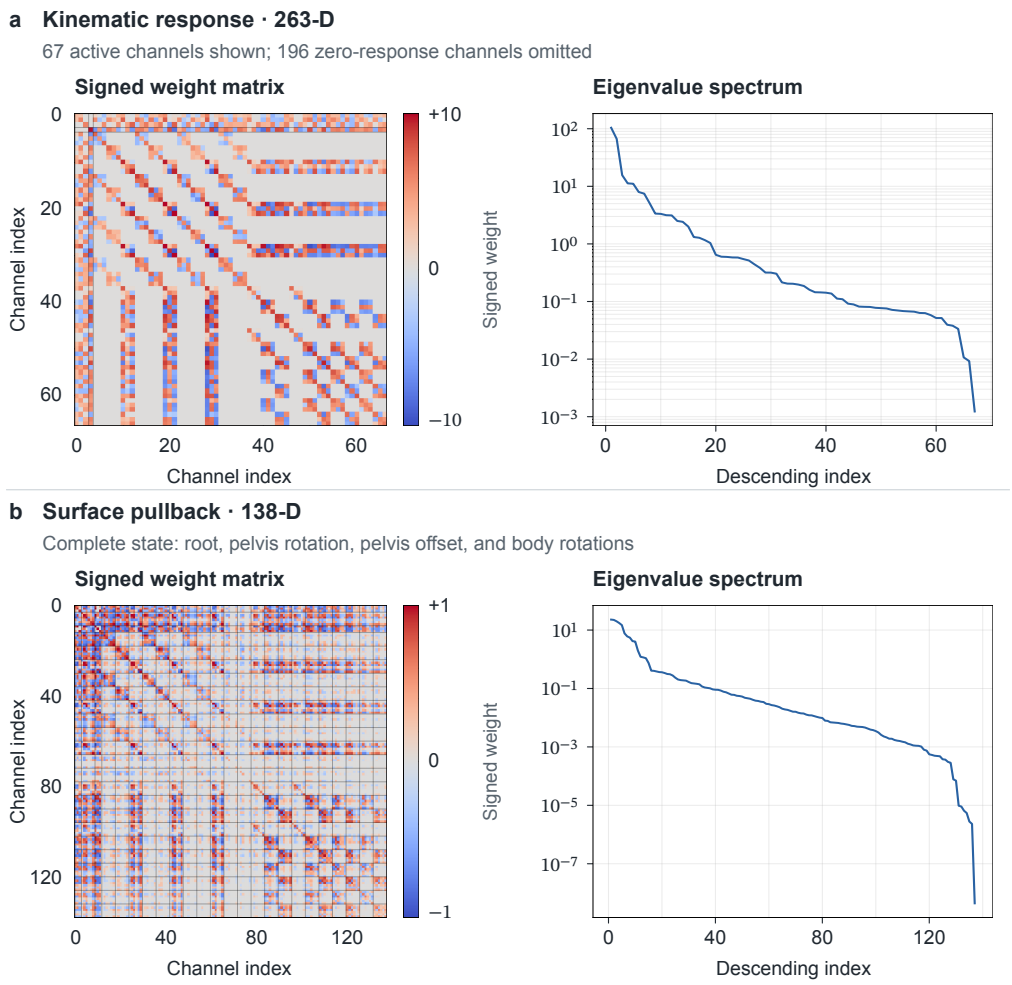

Figure 7: HumanML3D-calibrated FK matrices. Trace-normalized kinematic-response (263-D) and surface-pullback (138-D) matrices with descending eigenvalue spectra. The 263-D heatmap shows the active 67-channel block; omitted channels have zero weight. Each heatmap uses its own signed color scale.

Here W_{FK} is the trace-normalized average of $J_F(\mathbf{z})^\top J_F(\mathbf{z})$ over the dataset, as in Equation (33); it is positive definite in this experiment.

For the stated empirical data law, the exact conditional expectations can be written using posterior weights over the $N = 8{,}192$ training rotations:

$$w_i(\mathbf{x}_t, t) = \frac{\exp\!\left(-\|\mathbf{x}_t - t\mathbf{z}_i\|_2^2/[2(1-t)^2]\right)}{\sum_{j=1}^{N} \exp(-\|\mathbf{x}_t - t\mathbf{z}_j\|_2^2/[2(1-t)^2])}. \tag{47}$$

At fixed $(\mathbf{x}_t, t)$, the corresponding velocity target is $y_i = (\mathbf{z}_i - \mathbf{x}_t)/(1-t)$. Thus the conditional targets are $\overline{y} = \sum_i w_i y_i$ for MSE and the fixed quadratic, and $\overline{F} = \sum_i w_i F(\mathbf{z}_i)$ for the direct geometric loss. Removing the parameter-independent conditional variance gives losses $\|a - \overline{y}\|_2^2/6$, $(a-\overline{y})^\top W_{\mathrm{FK}}(a-\overline{y})/6$, and $\|F(\mathbf{x}_t + (1-t)a) - \overline{F}\|_2^2/9$, respectively.

All three velocity networks have three hidden layers of 256 SiLU units and share the same initialization and training batches within each seed. We use 50,000 AdamW updates with batch size 1,024, a cosine learning-rate schedule from 2×10^{-4} to 2×10^{-5}, and $t \sim \mathcal{U}(0.001, 0.999)$. Expected training losses are evaluated using exact conditional expectations over the empirical dataset, with parameter-independent variance terms omitted. The predicted velocity is integrated directly in the

six-dimensional coordinate. For each of two training seeds, we generate 8,192 samples from the same Gaussian initial states using 512 Euler steps and decode the final rotations with F.

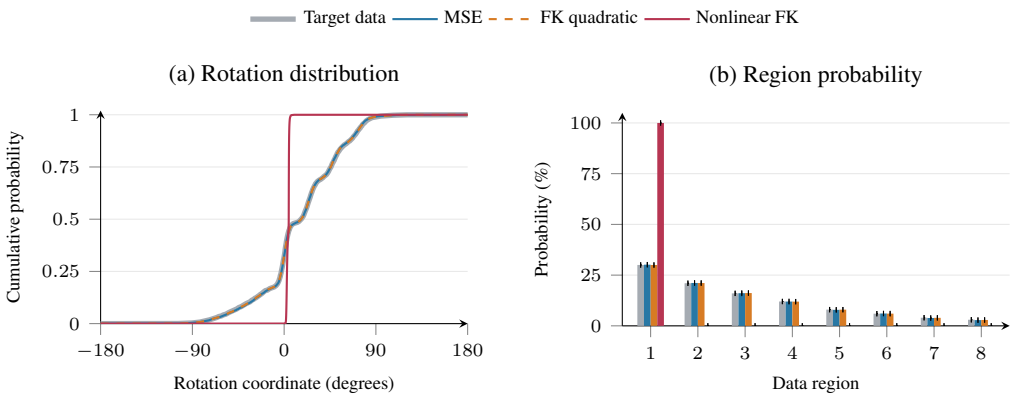

Figure 8: Sampling under three standalone regression losses. (a) Cumulative distributions of the decoded rotation vector projected onto $(1, 2, 3)^\top/\sqrt{14}$. (b) Probability assigned to each data region, determined by the nearest training rotation in FK output space. Curves and bars show means over two training seeds; whiskers show their ranges. MSE and the FK-induced quadratic retain all eight regions, whereas nonlinear FK concentrates samples in the highest-density region.

As shown in Figure 8, MSE and the FK-induced quadratic assign 30.15% and 29.98% of samples to the highest-density region, respectively, close to its 30.00% target probability, and retain the remaining regions. Pure nonlinear FK instead assigns 99.99% to this region in both runs. This experiment illustrates how nonlinear FK supervision can bias sampling toward high-density regions and reduce diversity, while the fixed quadratic preserves the conditional-mean target.

C Root-Path Conditioning and Motion Representation

This section expands the root-path conditioning in Section 3.4. FD2-Path receives a frame-aligned planar root position and heading, encodes them as three relative root channels, and keeps these channels clean while diffusing the remaining body state. The root defines the commanded path; the pelvis retains its relative translation and rotation so that body articulation can vary around this path.

SMPL-H reconstruction requires parent-local joint rotations and a global translation. HumanML3D-263 includes RIC joint positions and an IK-derived 21×6 rotation block; its joint positions do not uniquely determine bone twist, and its root stream retains heading rather than full pelvis orientation. Conversion to SMPL therefore requires approximate retargeting.

For SEED generation, our 138-D state stores the heading-relative pelvis rotation and the 21 source parent-local body rotations directly. With its translation channels, it reconstructs the 22 body joints and SMPL-H surface up to global planar placement and numerical precision. All encoding and decoding use neutral SMPL-H (Romero et al., 2017; Mahmood et al., 2019) with $\boldsymbol{\beta} = \mathbf{0}$ and the fixed model-mean hand pose, excluding actor-specific shape and variable hand articulation.

The representation diagnostics below use HumanML3D source motions to evaluate root smoothness and rotation recovery. The HumanML3D generation benchmark in the main text instead uses the native 263-D features.

C.1 Root–Pelvis Separation

We separate a planar root-motion node from the articulated pelvis. The root carries ground position and heading; the pelvis stores its vertical motion, local displacement, and residual 3D orientation relative to that node. This separates the user-facing trajectory from gait-cycle pelvis sway.

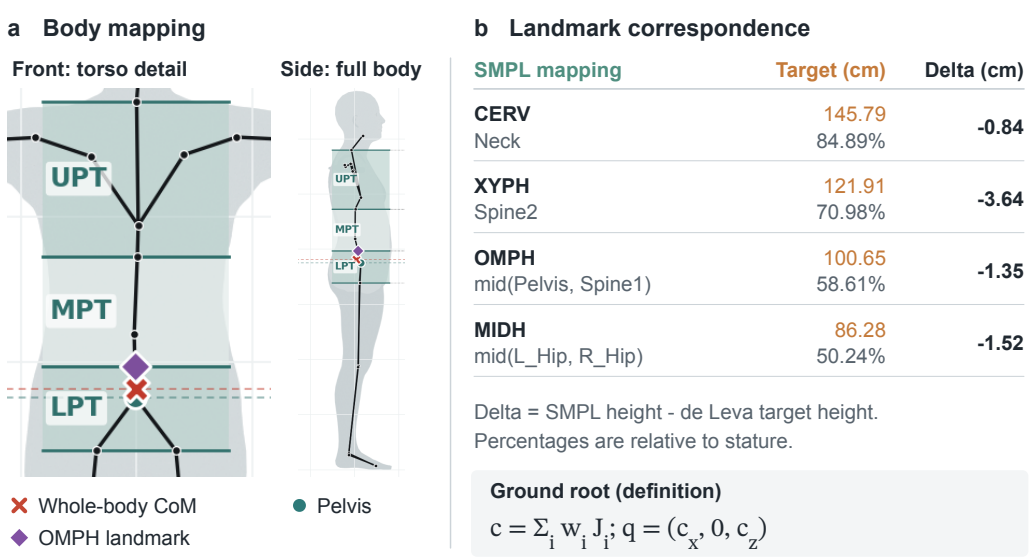

Figure 9: Biomechanical center-of-mass mapping to the neutral SMPL-H body. A front torso detail and a full-body side view locate the anatomical landmarks and the mass-weighted whole-body center of mass. The table compares the landmark heights from de Leva (1996) with their SMPL counterparts. The ground root is the planar projection of the resulting center of mass.

C.2 Mass-Weighted Ground-Projected Root Pivot

Let $\mathbf{J}_{k,i} \in \mathbb{R}^3$, $i = 0, \ldots, 21$, denote the 22 SMPL-H body-joint positions. We define an approximate whole-body pivot as

$$\mathbf{q}_k = \Pi_{\text{ground}}\left(\sum_{i=0}^{21} w_i \mathbf{J}_{k,i}\right), \qquad \sum_i w_i = 1, \tag{48}$$

where $\Pi_{\text{ground}}(x, y, z) = (x, 0, z)$. The weights are derived from the male-column adjusted Zatsiorsky–Seluyanov body-segment parameters of de Leva (1996) and are fixed across sequences. For a segment of mass m whose center is at fraction r from proximal landmark P to distal landmark D, the lever rule assigns $m(1-r)$ to P and mr to D.

We map the landmarks of de Leva (1996) to SMPL as follows: CERV≈Neck, XYPH≈Spine2, OMPH≈ mid(Pelvis, Spine1), and MIDH= mid(L_Hip, R_Hip). The OMPH and MIDH shares are each split equally between the two joints of their respective midpoint mappings. Because the 22-joint skeleton has no separate head-vertex (crown) landmark or hand joint, head mass is assigned to Head and hand mass to Wrist; heel is approximated by Ankle. Table 7 lists every resulting node assignment, and Figure 9 visualizes the anatomical-to-SMPL mapping and resulting ground projection.

A pelvis projection can oscillate laterally during walking. The mass-weighted construction is designed to let opposing limb contributions cancel before ground projection, while the difference between pelvis and pivot remains explicitly represented. We verify this effect on the complete HumanML3D training split in Table 8. Of 21,466 listed training entries, 21,462 are valid for this computation. Relative to the pelvis projection, the mass-weighted center-of-mass projection reduces planar jerk by 20.3% and curvature by 9.6%.

The heading-local root displacement and pelvis offset are

$$\mathbf{d}_k = \Pi_{xz}\big(R_y(\psi_k)^\top(\mathbf{q}_k - \mathbf{q}_{k-1})\big) \in \mathbb{R}^2, \tag{49}$$

$$\mathbf{o}_k = R_y(\psi_k)^\top(\mathbf{J}_{k,\text{Pelvis}} - \mathbf{q}_k) \in \mathbb{R}^3. \tag{50}$$

Table 7: SMPL-H node weights for the ground-projected root pivot. Weights follow the segment masses and center locations in de Leva (1996), using the stated landmark mapping and normalization to unit sum.

SMPL joint	Weight	SMPL joint	Weight
Pelvis	0.05846	Neck	0.07875
L_Hip	0.11777	L_Collar	0.00000
R_Hip	0.11777	R_Collar	0.00000
Spine1	0.05846	Head	0.06940
L_Knee	0.08225	L_Shoulder	0.01146
R_Knee	0.08225	R_Shoulder	0.01146
Spine2	0.17064	L_Elbow	0.02443
L_Ankle	0.02668	R_Elbow	0.02443
R_Ankle	0.02668	L_Wrist	0.01351
Spine3	0.00000	R_Wrist	0.01351
L_Foot	0.00605	R_Foot	0.00605

Table 8: Smoothness of the global root definition. Statistics are computed over the complete HumanML3D training split at 30 fps, with each motion weighted equally. Jerk is the mean norm of the third temporal derivative of planar position; curvature is evaluated at frames whose planar speed exceeds $0.05\,\mathrm{m/s}$.

Root definition	Jerk ↓ ($\mathrm{m/s^3}$)	Curvature ↓ ($\mathrm{m^{-1}}$)
Pelvis projection (263-style)	33.450	32.632
Mass-weighted center-of-mass projection (ours)	**26.650**	**29.504**

C.3 Heading from Hips and Shoulders

Pelvis orientation contains local body motion and can fluctuate even when the intended travel direction is stable. We instead estimate a skeletal lateral direction by averaging the hip and shoulder lines. On frames with $\mathbf{e}_y \times \mathbf{a}_k \neq \mathbf{0}$, define

$$\mathbf{a}_k = (\mathbf{J}_{k,\mathrm{RHip}} - \mathbf{J}_{k,\mathrm{LHip}}) + (\mathbf{J}_{k,\mathrm{RShoulder}} - \mathbf{J}_{k,\mathrm{LShoulder}}), \tag{51}$$

$$\mathbf{f}_k = \frac{\mathbf{e}_y \times \mathbf{a}_k}{\max(\|\mathbf{e}_y \times \mathbf{a}_k\|_2, \varepsilon)}, \qquad \psi_k = \mathrm{atan2}(f_{k,x}, f_{k,z}), \tag{52}$$

$$\Delta\psi_k = \mathrm{wrap}(\psi_k - \psi_{k-1}). \tag{53}$$

The heading uses both landmark pairs without temporal smoothing, and the representation stores its wrapped increment. On the nondegenerate domain specified above, the relative motion features are invariant to global yaw. The ε guard stabilizes normalization.

C.4 138-D Layout and Reconstruction

The full per-frame vector is

$$\mathbf{m}_k = [\underbrace{\Delta\psi_k}_{1}, \underbrace{\mathbf{d}_k}_{2}, \underbrace{\mathbf{r}_k}_{6}, \underbrace{\mathbf{o}_k}_{3}, \underbrace{\rho_6(R^{\mathrm{local}}_{k,1}), \ldots, \rho_6(R^{\mathrm{local}}_{k,21})}_{21\times 6}], \tag{54}$$

where $\mathbf{r}_k = \rho_6(R_y(\psi_k)^\top R_{k,\mathrm{Pelvis}})$ and ρ_6 is the continuous two-column rotation representation of Zhou et al. (2019). For a rotation matrix R, $\rho_6(R)$ contains its first two columns, $[R_{00}, R_{10}, R_{20}, R_{01}, R_{11}, R_{21}]$; ρ_6^{-1} applies the standard Gram–Schmidt reconstruction. Root quantities are heading-relative, and the 21 body rotations are parent-local.

Starting from planar position $\mathbf{q}_0$ and heading ψ_0, decoding for $k \geq 1$ uses

$$\psi_k = \psi_{k-1} + \Delta\psi_k, \tag{55}$$

$$\mathbf{q}_k = \mathbf{q}_{k-1} + R_y(\psi_k)[d_{k,x}, 0, d_{k,z}]^\top, \tag{56}$$

$$\mathbf{J}_{k,\mathrm{Pelvis}} = \mathbf{q}_k + R_y(\psi_k)\mathbf{o}_k, \tag{57}$$

$$R_{k,\mathrm{Pelvis}} = R_y(\psi_k)\rho_6^{-1}(\mathbf{r}_k). \tag{58}$$

Table 9: Source-rotation round trip. Frame-weighted mean geodesic error over the HumanML3D source-motion collection (26,846 motions; two SVD failures excluded for 263-D).

Representation	Dim.	Pelvis error (°)↓	Body-local error (°)↓
HumanML3D (Guo et al., 2022a)	263	1.635	30.680
MotionStreamer (Xiao et al., 2025)	272	$< 10^{-5}$	$< 10^{-5}$
FloodDiffusion 2 (ours, 138-D)	138	$< 10^{-5}$	$< 10^{-5}$

At $k = 0$, the stored pelvis residual and body rotations are decoded directly under $(\mathbf{q}_0, \psi_0)$. With valid encoded rotations and the same initial pose, this reconstruction recovers the original positions and rotations; heading is recovered modulo 2π. The parent-local body rotations then recover all remaining joint transforms through forward kinematics. The neutral-shape SMPL-H pose corrective blend shapes and LBS use those transforms to produce the final corresponding surface vertices. Changing $(\mathbf{q}_0, \psi_0)$ places the same relative joints and mesh at a new ground location and orientation.

C.5 Path Encoding

FD2-Path converts a complete per-frame planar pose sequence (x_k, z_k, ψ_k) through the encoder of its motion state. In the 138-D state, $\mathbf{c}_k = [\Delta\psi_k, \mathbf{d}_k]$ uses the wrapped heading change from $k-1$ to k and displacement expressed in heading ψ_k (Equations (49) and (53)). The native 263-D features instead store the heading rate and heading-local displacement leaving frame k.

The three encoded root channels are standardized with the same fixed channel statistics used during training and concatenated with the noisy body channels. Only the remaining 135 or 260 channels are predicted and supervised, as in Equation (10). During sampling, the supplied root channels remain fixed; decoding combines them with the generated pelvis and body state. Tracking is evaluated from the decoded motion using the conventions in Appendix D.5.

C.6 Source-Rotation Fidelity

We encode and decode 26,846 HumanML3D source motions and compare the recovered rotations with the source local rotations. The reported errors are frame-weighted. Two motions are omitted from the 263-D result because its SVD-based reconstruction does not converge. Table 9 reports the mean geodesic error. This round trip measures preservation of the source rotations through encoding and decoding.

D Implementation and Evaluation Details

This section collects the training configuration and measurement protocols for the experiments in Section 4. The timing and FLOP measurements describe denoiser computation; the benchmark-specific evaluation protocols define motion quality, path tracking, and foot skate.

D.1 Training Configuration

Shared architecture and optimization. FD2 uses a 114M-parameter transformer with 8 layers, width 1024, FFN width 2048, and 8 heads. Frozen UMT5-XXL (Chung et al., 2023) provides frame-aligned text features. We use $n_s = 30$, $K = 8$, and bf16 training. AdamW uses learning rate 2×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.99)$, cosine decay, and a batch of 128 source sequences per GPU. The full model uses $\gamma = 1$ in Equation (9); the ablation without FK uses $\gamma = 0$.

FD2 directly diffuses native 263-D features on HumanML3D and the 138-D rotation-direct state on SEED, using a common neutral-body convention. The latter decodes to SMPL-H as described in Appendix C. The full-training calibration diagnostics use 21,466 HumanML3D entries for the 138-D state and 23,384 entries for the 263-D features, the latter including additional HumanAct data. These are the calibration sets for the matrix diagnostics in Table 6. The common construction and branch-specific normalization are given in Appendix B.5.

Table 10: Dataset-specific checkpoint and inference settings. The number of predicted channels excludes the three clean root channels for FD2-Path.

Dataset	Variant	Predicted channels	Checkpoint step	CFG scale
HumanML3D	FD2	263	60K	4
HumanML3D	FD2-Path	260	55K	3
SEED	FD2	138	300K	2
SEED	FD2-Path	135	300K	2

D.2 Training-Efficiency Measurements

For the training-efficiency measurements in Table 2, the $K = 1$ baselines use four GPUs for 512 windows per step and eight GPUs for 1,024 windows, with 128 source sequences per GPU. Each NVIDIA H200 has 141 GB HBM3e memory; reported memory sums peak usage across GPUs. Packing uses the per-GPU batch sizes listed in the table, which differ from the default training batch.

D.3 KV-Cache Inference Measurements

Timing protocol. Table 3 compares uncached Full attention with KV-cached Partial Attention under the following common measurement protocol. We benchmark the 114M-parameter backbone (8 layers, width 1024, FFN width 2048, 8 heads, and 138-D input/output) on an NVIDIA RTX 4090 using PyTorch 2.9.1 and CUDA 12.8. Both paths use CUDA Graph replay and PyTorch SDPA with automatic backend selection. Each call processes one denoiser branch without CFG at batch 1, with 30 active frames and 0, 90, 450, or 4500 history frames. Text features are supplied as GPU-resident 16×4096 tensors; text projection and cross-attention are included in the measured backbone. Projection and feed-forward GEMMs use bf16 except for the time/output projections. Stored K/V uses bf16; normalization and residual arithmetic use FP32, with TF32 enabled.

After 8 warm-up calls, we measure 9 groups of 20 replays with CUDA events and report the median group-mean latency. Latency measures steady-state backbone execution with cached text features; startup, text encoding, motion decoding, and rendering are excluded.

Cached execution processes the 30 active frames and the newly finalized history frame in one forward pass, including causal re-encoding and K/V writes (31 rows, or 30 with empty history). Q/K/V projections and feed-forward layers process the rows jointly. Self-attention evaluates two query groups: the finalized frame attends to its inclusive causal prefix, and active frames attend to the complete visible prefix. Concatenating their outputs implements the Partial Attention mask. History K/V uses $2 \times 8 \times H \times 1024 \times 2$ bytes at history length H, or 140.625 MiB at $H = 4500$.

FLOPs count two operations per multiply–add in all linear and input-convolution projections and in the attention products $QK^\top$ and AV. Text and time projections, all eight transformer blocks, the output projection, and the newly finalized frame are included. Elementwise operations and memory transfers are excluded from the FLOP count.

D.4 HumanML3D Evaluation

We use the standard HumanML3D split and released text-to-motion evaluator (Guo et al., 2022a), with FD1's preprocessing and duration handling. Table 4 reports the mean and 95% confidence interval for repeated evaluations. The HumanML3D generation models operate on the native 263-D representation; the 138-D round-trip and root-smoothness diagnostics in Appendix C evaluate representation properties separately.

D.5 Path-Control Evaluation Conventions

For FD2-Path, Position in Table 5 is the distance from the mass-weighted center-of-mass projection recomputed from the generated body joints using Equation (48) to the commanded root trajectory. Both are measured in the decoded neutral SMPL-H coordinate system, before the SOMA conversion and its inverse scale adjustment. Distances are pooled over conditioned frames, computed in meters,

and reported in centimeters. For the 917-case evaluation, the mean and P95 use 183,130 conditioned frames.

On HumanML3D, GMD and DART target planar pelvis positions; OmniControl targets 3D pelvis positions. FD2-Path conditions on planar root position and heading. Root and 3D in Table 4 measure the mean xz-plane control-point error and 3D pelvis error relative to ground truth, respectively, pooled over frames. Skate follows the OmniControl convention at 20 fps: a foot is in contact when its height is below 5 cm in both adjacent frames, and skates when both its instantaneous planar speed and its five-frame average exceed 0.5 m/s. A frame counts as skating if either foot skates; we report the per-sequence fraction averaged over sequences.

D.6 SEED Evaluation

We use the public Kimodo training partition and category-held-out Content test split of SEED, with the official v1.1 evaluator (NVIDIA, 2026). Path-conditioned methods receive dense paths for the same 917 cases in Table 5.

We decode the 138-D output into the 22 SMPL-H body-joint rotations and the world-space pelvis trajectory. The rotations are mapped to SOMA77 using the inverse joint correspondence of the SEED training-data conversion; the neck rotation is assigned to Neck1, and unmapped joints retain identity rotations. Pelvis translations are divided by the training-conversion scale factor 0.915609. The official SOMA skeleton then produces 77-joint positions by forward kinematics at 30 fps. These positions are passed to the released TMR-SOMA-RP-v1 motion encoder, which applies its standard joint selection, canonicalization, and normalization. R@3 and FID use the official embedding and aggregation pipeline.

For generated motions, SEED Skate uses the same contact thresholds for all methods: 0.10 m in height and 0.15 m/s in speed, implemented by the official motion conversion routine. We report the mean 3D foot speed over contact frames in cm/s, averaged over sequences. The official metric returns zero for a sequence with no contact frames. Parenthesized Kimodo and ARDY scores use the official models' predicted contact labels on the same generated motions.