QUANTOPO: TOPOLOGY-DIRECTED HIERARCHICAL TEXT-TO-MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-motion generation has advanced in realism and text alignment, yet composite descriptions require global sequence organization alongside diverse, valid realizations. We introduce QuanTopo, a topology-directed hierarchical framework that learns the discrete branch of a hybrid generator as an explicit motion plan rather than only a reconstruction code. QuanTopo addresses a target–use gap: a useful plan needs both an objective that specifies what it should preserve and a structured role in synthesis. To define plan content, persistent homology summarizes multiscale relations along a trajectory of pose, root motion, and local change, emphasizing connectivity, revisitation, and cycles beyond framewise fidelity. To make this content actionable, the plan establishes anatomical context before continuous realization enters joint synthesis, while the continuous pathway models alternative timings, postures, and coordination patterns. A predictor conditioned on text and supplied duration connects the same trajectory target to inference. QuanTopo obtains FID scores of 0.027/0.122 on HumanML3D/KIT-ML and reaches 0.804 composite-action sequence satisfaction on a prespecified subset of composite HumanML3D captions, versus 0.782 for the strongest baseline. In this hybrid setting, ablations and interventions support aligning a plan's information target with its role in synthesis.

1 INTRODUCTION

Text-to-motion generation maps natural-language descriptions to plausible three-dimensional movement. Recent models can produce fluent motions for a wide range of actions, but composite descriptions expose a deeper requirement than local pose quality. The instruction "walk forward, turn around, and then raise both arms" determines a sequence of semantic stages while leaving stride, timing, posture, and inter-limb coordination underdetermined. A successful generator must therefore preserve sequence-level organization without removing the freedom needed to realize the same instruction in different valid ways. Figure 1 highlights the two sequence-level phenomena targeted by QuanTopo: ordered semantic stages in a composite action and loop, crossing, and return structure in the root trajectory.

Recent work approaches this challenge through complementary motion representations. Continuous diffusion, latent-variable, and flow models learn expressive distributions over motion space or continuous latent spaces (Zhang et al., 2022; Tevet et al., 2023; Chen et al., 2023; Li et al., 2026). Discrete generators compress motion into token sequences that provide a compact interface for language conditioning (Zhang et al., 2023a; Guo et al., 2024; Pinyoanuntapong et al., 2024a;b), while hybrid and hierarchical systems distribute responsibility across prototypes and residuals, temporal scales, body regions, or action intervals (Wang et al., 2026; Athanasiou et al., 2023; Petrovich et al., 2024). These directions have substantially improved motion quality, diversity, and controllability, and they increasingly separate coarse organization from fine realization.

Across these families, however, intermediate representations are still defined mainly by reconstruction, quantization, or generative prediction. Such objectives determine how motion can be recovered or sampled, while leaving the intended planning content and its role in synthesis implicit. We call this mismatch the *target–use gap*. A useful motion plan should satisfy four linked requirements: it should capture relations that organize an entire sequence, be inferable from the conditions available at test time, enter the generator through a structured synthesis pathway, and leave sufficient freedom

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

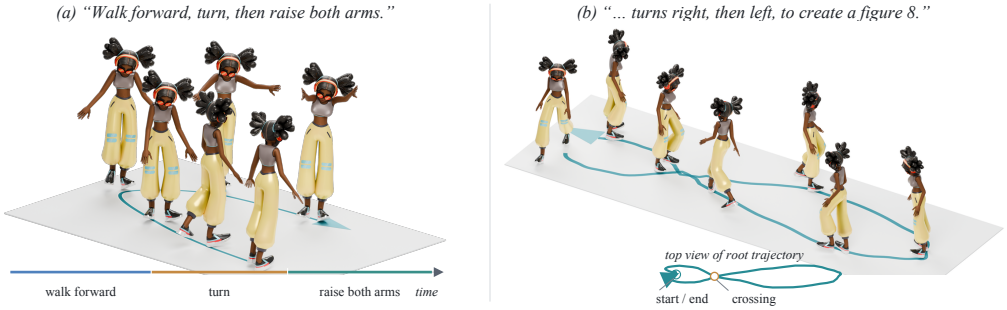

Figure 1: QuanTopo generation examples. (a) The motion follows the ordered stages of a compositional description. (b) A figure-eight motion preserves its loops, crossing, and return structure; the inset shows the top-view root trajectory. Only key frames are displayed.

for multiple valid executions. These requirements motivate jointly designing what the plan preserves and how the generator uses it.

QuanTopo builds on the observation that motion is not only a sequence of poses, but also a trajectory formed by articulation, root travel, and local change. Motion organization is therefore relational: it depends on how states encountered at different times connect, recur, and form larger structures. Persistent homology tracks how these relations emerge as the distance scale expands (Cohen-Steiner et al., 2007; Edelsbrunner et al., 2002), providing a compact description of connectivity, revisitation, and cyclic structure beyond a single scale or individual frame. QuanTopo uses this description to guide the global content of a discrete plan and requires the plan to remain directly usable for motion synthesis. To close the train–test loop, a predictor conditioned on language and supplied duration estimates the same trajectory descriptor used during representation learning. The descriptor supplies geometric organization, while language and the ordered token sequence carry semantic progression.

The second half of the design gives the plan a structured point of use. Anatomical routing provides a middle ground between broadcasting one global code uniformly and assigning unrelated plans to individual joints: coordinated joints share part context while retaining distinct local features. QuanTopo establishes this context from the plan before combining it with a continuous realization during joint synthesis. The continuous generator then models alternative timings, postures, and coordination patterns compatible with the selected plan. Together, topology-directed planning and plan-driven synthesis form a direct path from global trajectory relations to structured body generation while preserving performance-level variation.

Experiments support both the resulting generator and this design choice. QuanTopo obtains FID scores of 0.027 on HumanML3D and 0.122 on KIT-ML, and improves composite-action sequence satisfaction from 0.782 for the strongest baseline to 0.804. Matched trajectory controls, component ablations, and condition interventions further test how the planning target and synthesis pathway contribute to these gains.

Our contributions are threefold:

- We formulate the target–use gap in learned motion planning and show that, in a hybrid text-to-motion generator, jointly specifying plan content and its synthesis role produces a more informative and usable discrete representation. This perspective distinguishes explicit planning from generic compression.
- We introduce QuanTopo, which combines topology-directed discrete planning with anatomically structured, plan-driven synthesis. Persistent trajectory supervision specifies global plan content, while plan-only reconstruction and anatomical routing make that content usable before conditional flow supplies realization-level variation.
- We evaluate QuanTopo on standard text-to-motion benchmarks and a composite-action protocol. Beyond overall generation quality, matched summary controls, component ablations, and condition interventions test how trajectory supervision, anatomical construction, and continuous realization contribute to the resulting motions.

2 Related Work

2.1 Text-to-Motion Generation

Continuous approaches learn text-conditioned distributions in motion space or continuous latent spaces through variational modeling, diffusion, flow, and latent consistency (Petrovich et al., 2022; Tevet et al., 2023; Zhang et al., 2022; Chen et al., 2023; Ma et al., 2022; Li et al., 2026; Dai et al., 2024). Retrieval guidance, efficient sampling, and large-scale pretraining further improve quality or efficiency (Zhang et al., 2023c; Zeng et al., 2025; Zhou et al., 2024; Tencent Hunyuan 3D Digital Human Team, 2025).

Discrete approaches compress motion into token sequences. T2M-GPT predicts codes autoregressively; MMM and MoMask use masked prediction, with MoMask adding residual quantization levels (Zhang et al., 2023a; Pinyoanuntapong et al., 2024a; Guo et al., 2024). BAMM combines autoregressive and bidirectional prediction, while MoGenTS constructs joint-level spatiotemporal tokens (Pinyoanuntapong et al., 2024b; Yuan et al., 2024). DC-Motion retains the continuous residual between encoded features and quantized prototypes (Wang et al., 2026). QuanTopo instead gives the discrete component a trajectory-level objective and an explicit role in synthesis.

2.2 Hierarchical Motion Modeling and Compositional Generation

Hierarchy has been introduced along temporal, semantic, and anatomical axes. TEACH connects successive actions; SINC and STMC compose interval-aligned body-part actions; and B2A-HDM refines low-dimensional semantic motion in a high-dimensional diffusion stage (Athanasiou et al., 2022; 2023; Petrovich et al., 2024; Xie et al., 2024). Compositional systems often receive action sequences, temporal intervals, or part-specific descriptions as generation-time conditions; with only a caption and duration, intermediate sequence organization must be inferred rather than supplied (Athanasiou et al., 2022; 2023; Petrovich et al., 2024). FineMoGen, SALAD, and EnergyMoGen model relations among phases, parts, joints, frames, words, or diffusion energies (Zhang et al., 2023b; Hong et al., 2025; Zhang et al., 2025a). KinMo uses kinematic groups, PB-MLD separates upper- and lower-body latents, and MotionHiFlow constructs coarse-to-fine flows across temporal resolutions and skeletal topology (Zhang et al., 2025b; Liu et al., 2026; Li et al., 2026). QuanTopo instead uses anatomical hierarchy as the synthesis interface for a separately supervised discrete plan.

2.3 Topological Representation Learning

Persistent homology records structures across a filtration and is stable under suitable perturbations (Edelsbrunner et al., 2002; Cohen-Steiner et al., 2007). Persistence Images and PersLay aggregate unordered diagrams into fixed or learned representations (Adams et al., 2017; Carrière et al., 2020). Topological objectives organize learned representations (Moor et al., 2020; Gabrielsson et al., 2020; Balabin et al., 2024), and persistence diagrams have also compared collective motion through velocity fields (Legarda Herranz et al., 2026). QuanTopo instead integrates a persistent summary of an articulated motion-state trajectory into text-conditioned discrete planning.

3 Method

3.1 Overview

QuanTopo realizes the planning requirements identified in Section 1 through a planning–realization hierarchy. A persistent trajectory target specifies the global relations that the plan should retain; plan-only synthesis and anatomical routing make that plan actionable; a text-and-length predictor makes the same target available at inference; and a conditional flow retains multiple realizations compatible with a sampled plan. Figure 2 summarizes representation learning and text-conditioned inference.

Problem formulation. Given text c and an externally supplied length T, the target motion is $x = (X_1, \ldots, X_T)$, where $X_t \in \mathbb{R}^{J\times 3}$. The model uses a discrete plan $q \in \{1, \ldots, K\}^L$ and a continuous realization $r \in \mathbb{R}^{T\times d_r}$. With $\hat{z} = P_{\theta_z}(c, T)$ denoting text-and-length-to-topology

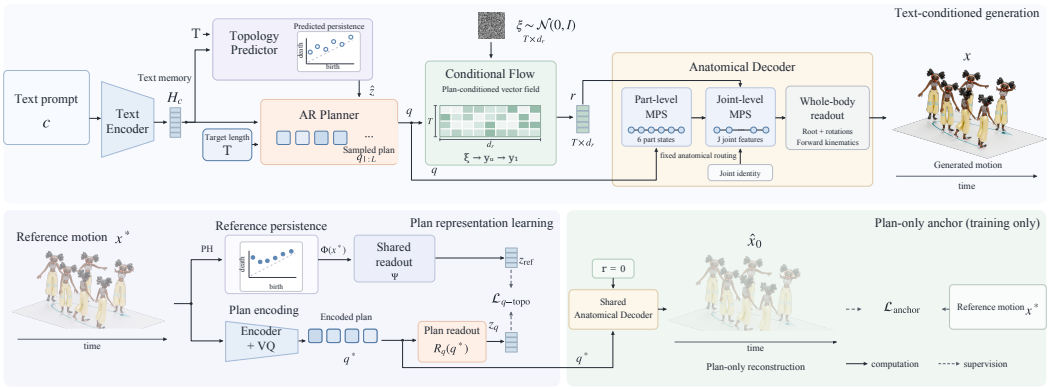

Figure 2: QuanTopo overview. Text and supplied length predict a trajectory descriptor, the AR planner samples q, conditional flow generates r, and the anatomical decoder produces motion. During representation learning, reference persistence supervises the encoded plan and $D(q^*, 0)$ provides the plan-only anchor. Solid and dashed arrows denote computation and supervision.

prediction followed by the shared readout, generation is organized as

$$p(q, r \mid c, T) = p_\theta(q \mid c, \hat{z}, T)\, p_\phi(r \mid c, q, T), \qquad \tilde{x} = D_\psi(q, r). \tag{1}$$

The low-rate sequence q is selected before realization sampling and carries sequence-level organization. The frame-rate variable r is then generated conditionally on q, allowing different executions without requiring the plan to encode every local detail. Equation 1 specifies these computational roles rather than an identifiable semantic decomposition. We write the deterministic decoder D_ψ as D below.

3.2 Topology-Directed Plan Representation

Trajectory relations. We combine pose, spatial movement, and local change in the state

$$s_t = [\mathrm{vec}(X_t - X_{t,\mathrm{root}}),\ X_{t,\mathrm{root}},\ \mathrm{vec}(\Delta X_t)], \qquad \Delta X_t = X_{t+1} - X_t, \quad 1 \le t < T. \tag{2}$$

At the final valid frame we set $\Delta X_T = X_T - X_{T-1}$; a one-frame sequence uses $\Delta X_1 = 0$. This convention is identical for training, validation, and test data. Root-relative pose describes articulation, the root term retains the traveled path, and displacement records local change.

Framewise reconstruction judges states at aligned times, but sequence organization also depends on relations that emerge as neighborhoods expand. We therefore construct a sparse filtered complex from state neighbors and temporal adjacency. Across its filtration, H_0 records component mergers and H_1 records the birth and death of cycles, producing the persistence diagram $\Phi(x)$. A fixed Gaussian readout maps this unordered set to $z = S(x) = \Psi(\Phi(x)) \in \mathbb{R}^{d_z}$. The resulting descriptor provides a direction-agnostic reference for multiscale trajectory organization, including clustering, revisitation, and cyclic structure. Semantic progression and temporal direction are supplied instead by language and the ordered encoder, decoder, and AR objectives. Appendix A.1 gives the exact complex, normalization, and persistence readout.

Paired representation. A temporal encoder produces $h_\ell = E_q(x)_\ell$, which is quantized by the nearest codeword (van den Oord et al., 2017). A continuous encoder reads motion and the upsampled plan:

$$q_\ell = \arg\min_k \|h_\ell - e_k\|_2^2, \quad \bar{q} = \mathrm{Up}(e_{q_1}, \ldots, e_{q_L}), \quad r = E_r(x, \bar{q}). \tag{3}$$

The decoder reconstructs $\tilde{x} = D(q, r)$. Conditioning E_r on the selected plan makes r a plan-specific remainder rather than an independent second description of the motion. The generator learns this same conditional coordinate system, so a plan can support multiple realizations with different local timing and style. For a reference motion x^*, we denote its encoded pair by (q^*, r^*).

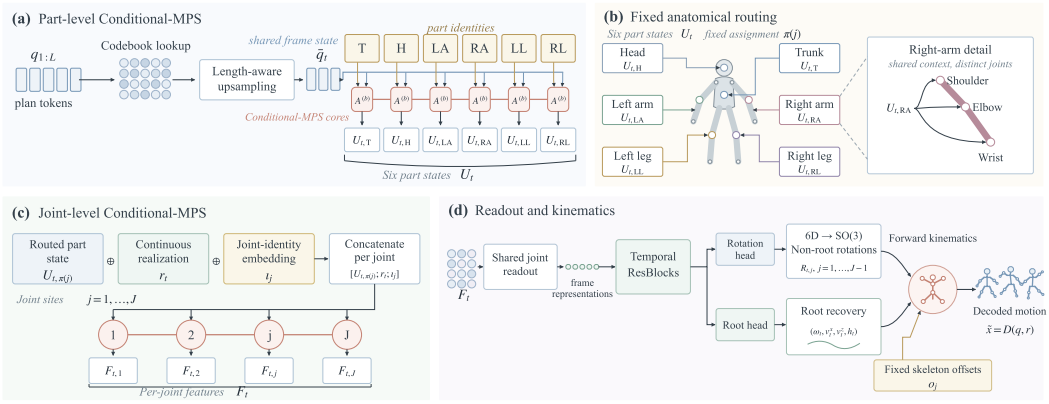

Figure 3: Plan-driven anatomical synthesis. (a) Plan embeddings form six part states. (b) Fixed assignment $\pi(j)$ routes them to joints. (c) Each joint concatenates its routed context, realization r_t, and identity ι_j. (d) Shared readout, temporal modeling, and forward kinematics decode $\tilde{x}$.

An informative, synthesizable plan. To target the discrete branch directly, a plan readout predicts the reference summary and the same decoder reconstructs motion with zero continuous input:

$$b(q) = \frac{\sum_\ell m_\ell e_{q_\ell}}{\sum_\ell m_\ell}, \qquad R_q(q) = f(b(q)), \tag{4}$$

$$\mathcal{L}_{q\text{-topo}} = \|R_q(q) - S(x)\|_2^2, \qquad \mathcal{L}_{\text{anchor}} = \mathcal{L}_{\text{rec}}(D(q,0), x). \tag{5}$$

$z_q = R_q(q^*)$ denotes the encoded-plan descriptor in Figure 2, and $\hat{x}_0 = D(q^*, 0)$ its plan-only reconstruction. m_ℓ masks valid positions. The descriptor loss specifies which global trajectory relations the plan should retain, while plan-only reconstruction requires the ordered plan to support synthesis before continuous detail is introduced. Together, they make q both trajectory-referenced and operational in the decoder. Because $b(q)$ is permutation-invariant (Zaheer et al., 2017), token order is learned through the ordered encoder, decoder, and AR objectives rather than attributed to the topological readout.

3.3 Plan-Driven Anatomical Synthesis

Broadcasting one plan vector identically to every joint leaves the decoder to rediscover body organization, whereas assigning an independent plan to each joint can fragment coordination and greatly enlarge the conditioning interface. Six parts—trunk, head, left/right arm, and left/right leg—form a middle ground: joints that commonly coordinate share context, while bilateral and axial regions remain distinguishable.

Let $\pi(j)$ assign joint j to the trunk, head, left/right arm, or left/right leg. The plan first forms part context and then enters joint construction:

$$U_t = F_{\text{part}}(\bar{q}_t), \qquad F_t = F_{\text{joint}}(\{r_t, U_{t,\pi(j)}, \iota_j\}_{j=1}^J). \tag{6}$$

$U_{t,p}$ is a part state, $F_{t,j}$ a joint feature, and ι_j a joint-identity embedding. The right shoulder, elbow, and wrist, for example, receive the same right-arm context while retaining distinct joint identities. We instantiate both levels with conditional MPS/tensor-train operators (Oseledets, 2011). Their repeated contractions exchange information across sites while the bond width controls parameter growth, providing a compact alternative to a dense jointwise map. Affine conditioning modulates the local cores $A^{(b)}$ (Perez et al., 2018). Fixed routing determines which plan context reaches each joint, whereas the MPS determines how part or joint sites interact; a parameter-matched Conditional-MLP isolates the effect of this operator from routing and model size. Appendix A.3 gives the contractions, site order, and decoder conventions. This ordering is deliberate: q establishes shared anatomical context before r supplies within-plan variation, making their distinct roles explicit; Tables 2 and 4 test these responsibilities separately.

A shared temporal readout predicts non-root six-dimensional rotations (Zhou et al., 2019) and root motion. With local rotation $R_{t,j}$, global rotation $G_{t,j}$, parent $p(j)$, and fixed offset o_j,

$$G_{t,j} = G_{t,p(j)} R_{t,j}, \qquad X_{t,j} = X_{t,p(j)} + G_{t,p(j)} o_j, \qquad j \neq 0. \tag{7}$$

Fixed offsets and valid rotations preserve bone lengths.

3.4 Text-Conditioned Planning and Continuous Generation

A frozen CLIP encoder provides contextual text tokens H_c (Radford et al., 2021). A topology head combines them with a learned embedding of normalized supplied length and predicts a weighted birth–death point set $\widehat{\Phi}(c, T)$; the shared readout gives $\hat{z} = \Psi(\widehat{\Phi}(c, T))$. Predicting the set, rather than only its fixed-dimensional readout, respects the absence of a canonical point ordering and the unequal existence and lifetime of topological features. Reusing Ψ keeps representation learning and text-conditioned inference in the same descriptor space. Set matching, point-existence, and descriptor losses train the head. The planner then generates

$$p_\theta(q \mid c, \hat{z}, T) = \prod_{\ell=1}^{L} p_\theta(q_\ell \mid q_{<\ell}, H_c, \hat{z}, T). \tag{8}$$

Zero-preserving geometric modulation. The AR planner forms state a from the text context, supplied length, and token prefix, and maps it to base logits. Geometry adds a bounded response difference:

$$\ell_\theta(a, c, z) = \ell_0(a, c) + \alpha \,\mathrm{clip}[\rho(a, z \oslash \sigma_z) - \rho(a, 0), -\tau, \tau]. \tag{9}$$

σ_z is a positive descriptor scale. Because both calls share weights and context, the response difference vanishes exactly at $z = 0$, recovers the language-only logits, and therefore does not alter the base language logits. The descriptor biases plan-token selection rather than prescribing a root path; the ordered plan and continuous realization still determine the motion. Scaling balances descriptor coordinates, and clipping bounds the geometric correction when the predicted diagram is inaccurate. Training uses predicted descriptors, matching the condition available at inference. A soft-plan auxiliary also applies the plan readout to expected token embeddings, aligning the predicted plan distribution with $S(x)$.

Continuous realizations. A text and plan can admit several valid timings, postures, and local trajectories. To preserve this within-plan variation, we use Flow Matching (Lipman et al., 2023) to learn the conditional distribution of r. Let y^* be the channelwise-normalized $r^* = E_r(x, q^*)$, let $\xi \sim \mathcal{N}(0, I)$ and $u \sim \mathcal{U}[0, 1]$, and define $y_u = (1 - u)\xi + uy^*$. Then

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\frac{\|v_\phi(y_u, u; H_c, q^*, T) - (y^* - \xi)\|_2^2}{T d_r}\right]. \tag{10}$$

The vector field cross-attends to text and the upsampled plan (Vaswani et al., 2017). At inference, $y_0 = \xi$ is integrated to y_1, which is inverse-normalized to decoder-native r before returning $D(q, r)$. A gated roll-in occasionally trains the flow with accepted frozen-planner sequences to reduce exposure mismatch. Representation learning precedes text planning and continuous generation, and reference diagrams are precomputed.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We use the established HumanML3D (Guo et al., 2022) and KIT-ML (Plappert et al., 2016) splits and the HumanML3D evaluator. Following Guo et al. (2022), we report R-Precision, Matching Distance, Diversity, and MultiModality (MM); FID follows Heusel et al. (2017). Test-record lengths are supplied for QuanTopo and unified reevaluation, while source-reported baselines retain their original protocols. Component metrics add full/plan-only MPJPE, plan information error e_q, and within-plan joint variation V_{joint}. For composite instructions, Seq. measures joint event completion and precedence satisfaction edgewise, Coverage measures event coverage, and Order|Covered measures ordering accuracy conditional on coverage. Standard metrics assess distributional fidelity and text–motion alignment; the composite metrics test joint event realization and prescribed order.

Table 1: Results on HumanML3D (Guo et al., 2022) and KIT-ML (Plappert et al., 2016). $\pm$ indicates a 95% confidence interval; bold and underline mark the best and second-best results for metrics with a directional ranking. Diversity is reported without ranking.

Method	R@1↑	R@2↑	R@3↑	FID↓	Match↓	Div.↔	MM↑
HumanML3D							
MDM (Tevet et al., 2023)	—	—	$0.611^{\pm 0.007}$	$0.544^{\pm 0.044}$	$5.566^{\pm 0.027}$	—	$\mathbf{2.799}^{\pm 0.072}$
MLD (Chen et al., 2023)	$0.481^{\pm 0.003}$	$0.673^{\pm 0.003}$	$0.772^{\pm 0.002}$	$0.473^{\pm 0.013}$	$3.196^{\pm 0.010}$	—	$2.413^{\pm 0.079}$
T2M-GPT (Zhang et al., 2023a)	$0.492^{\pm 0.003}$	$0.679^{\pm 0.002}$	$0.775^{\pm 0.002}$	$0.141^{\pm 0.005}$	$3.121^{\pm 0.009}$	$9.722^{\pm 0.082}$	$1.831^{\pm 0.048}$
ReMoDiffuse (Zhang et al., 2023c)	$0.510^{\pm 0.005}$	$0.698^{\pm 0.006}$	$0.795^{\pm 0.004}$	$0.103^{\pm 0.004}$	$2.974^{\pm 0.016}$	$9.018^{\pm 0.075}$	$1.795^{\pm 0.043}$
MoMask (Guo et al., 2024)	$0.521^{\pm 0.002}$	$0.713^{\pm 0.002}$	$0.807^{\pm 0.002}$	$0.045^{\pm 0.002}$	$2.958^{\pm 0.008}$	—	$1.241^{\pm 0.040}$
BAMM (Pinyoanuntapong et al., 2024b)	$0.522^{\pm 0.003}$	$0.715^{\pm 0.003}$	$0.808^{\pm 0.003}$	$0.055^{\pm 0.002}$	$2.936^{\pm 0.077}$	$9.636^{\pm 0.009}$	$1.732^{\pm 0.055}$
MoGenTS (Yuan et al., 2024)	$0.529^{\pm 0.003}$	$0.719^{\pm 0.002}$	$0.812^{\pm 0.002}$	$0.033^{\pm 0.001}$	$2.867^{\pm 0.006}$	$9.570^{\pm 0.077}$	—
Light-T2M (Zeng et al., 2025)	$0.511^{\pm 0.003}$	$0.699^{\pm 0.002}$	$0.795^{\pm 0.002}$	$0.040^{\pm 0.002}$	$3.002^{\pm 0.008}$	—	$1.670^{\pm 0.061}$
SALAD (Hong et al., 2025)	$\underline{0.581}^{\pm 0.003}$	$\underline{0.769}^{\pm 0.003}$	$\underline{0.857}^{\pm 0.002}$	$0.076^{\pm 0.002}$	$\underline{2.649}^{\pm 0.009}$	$9.696^{\pm 0.096}$	$1.751^{\pm 0.062}$
EnergyMoGen (Zhang et al., 2025a)	$0.523^{\pm 0.003}$	$0.715^{\pm 0.002}$	$0.815^{\pm 0.002}$	$0.188^{\pm 0.006}$	$2.915^{\pm 0.007}$	$9.488^{\pm 0.099}$	$2.205^{\pm 0.041}$
DC-Motion (Wang et al., 2026)	$0.528^{\pm 0.002}$	$0.720^{\pm 0.002}$	$0.812^{\pm 0.002}$	$0.041^{\pm 0.002}$	$2.945^{\pm 0.008}$	—	$\underline{2.450}^{\pm 0.070}$
MotionHiFlow (Li et al., 2026)	$0.563^{\pm 0.003}$	$0.754^{\pm 0.003}$	$0.843^{\pm 0.003}$	$\underline{0.032}^{\pm 0.002}$	$2.691^{\pm 0.009}$	$9.504^{\pm 0.071}$	—
QuanTopo	$\mathbf{0.586}^{\pm 0.002}$	$\mathbf{0.774}^{\pm 0.003}$	$\mathbf{0.861}^{\pm 0.002}$	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{2.623}^{\pm 0.008}$	$9.641^{\pm 0.074}$	$2.083^{\pm 0.058}$
KIT-ML							
MDM (Tevet et al., 2023)	—	—	$0.396^{\pm 0.004}$	$0.497^{\pm 0.021}$	$9.191^{\pm 0.022}$	—	$1.907^{\pm 0.214}$
MLD (Chen et al., 2023)	$0.390^{\pm 0.008}$	$0.609^{\pm 0.008}$	$0.734^{\pm 0.007}$	$0.404^{\pm 0.027}$	$3.204^{\pm 0.027}$	—	$\mathbf{2.192}^{\pm 0.071}$
T2M-GPT (Zhang et al., 2023a)	$0.416^{\pm 0.006}$	$0.627^{\pm 0.006}$	$0.745^{\pm 0.006}$	$0.514^{\pm 0.029}$	$3.007^{\pm 0.023}$	$10.921^{\pm 0.108}$	$1.570^{\pm 0.039}$
ReMoDiffuse (Zhang et al., 2023c)	$0.427^{\pm 0.014}$	$0.641^{\pm 0.004}$	$0.765^{\pm 0.055}$	$0.155^{\pm 0.006}$	$2.814^{\pm 0.012}$	$10.80^{\pm 0.105}$	$1.239^{\pm 0.028}$
MoMask (Guo et al., 2024)	$0.433^{\pm 0.007}$	$0.656^{\pm 0.005}$	$0.781^{\pm 0.005}$	$0.204^{\pm 0.011}$	$2.779^{\pm 0.022}$	—	$1.131^{\pm 0.043}$
BAMM (Pinyoanuntapong et al., 2024b)	$0.436^{\pm 0.007}$	$0.660^{\pm 0.006}$	$0.791^{\pm 0.005}$	$0.200^{\pm 0.011}$	$2.714^{\pm 0.016}$	$10.914^{\pm 0.097}$	$1.517^{\pm 0.058}$
MoGenTS (Yuan et al., 2024)	$0.445^{\pm 0.006}$	$0.671^{\pm 0.006}$	$0.797^{\pm 0.005}$	$0.143^{\pm 0.004}$	$2.711^{\pm 0.024}$	$10.918^{\pm 0.090}$	—
Light-T2M (Zeng et al., 2025)	$0.444^{\pm 0.006}$	$0.670^{\pm 0.007}$	$0.794^{\pm 0.005}$	$0.161^{\pm 0.009}$	$2.746^{\pm 0.016}$	—	$1.005^{\pm 0.036}$
SALAD (Hong et al., 2025)	$0.477^{\pm 0.006}$	$\underline{0.711}^{\pm 0.005}$	$\underline{0.828}^{\pm 0.005}$	$0.296^{\pm 0.012}$	$2.585^{\pm 0.016}$	$11.097^{\pm 0.095}$	$1.004^{\pm 0.040}$
EnergyMoGen (Zhang et al., 2025a)	$0.436^{\pm 0.006}$	$0.651^{\pm 0.006}$	$0.772^{\pm 0.006}$	$0.495^{\pm 0.020}$	$2.861^{\pm 0.020}$	$11.06^{\pm 0.101}$	$1.256^{\pm 0.024}$
DC-Motion (Wang et al., 2026)	$0.442^{\pm 0.005}$	$0.665^{\pm 0.005}$	$0.788^{\pm 0.005}$	$0.148^{\pm 0.008}$	$2.745^{\pm 0.018}$	—	$\underline{1.950}^{\pm 0.060}$
MotionHiFlow (Li et al., 2026)	$\underline{0.482}^{\pm 0.005}$	$0.704^{\pm 0.005}$	$0.825^{\pm 0.005}$	$\underline{0.135}^{\pm 0.007}$	$\underline{2.552}^{\pm 0.014}$	$10.894^{\pm 0.117}$	—
QuanTopo	$\mathbf{0.489}^{\pm 0.004}$	$\mathbf{0.715}^{\pm 0.004}$	$\mathbf{0.833}^{\pm 0.004}$	$\mathbf{0.122}^{\pm 0.006}$	$\mathbf{2.523}^{\pm 0.012}$	$10.972^{\pm 0.101}$	$1.612^{\pm 0.066}$

Implementation and statistics. We compress normalized RIC input to a stride-four plan and a 128-D continuous code, and train the representation/planner/flow for 300k/200k/200k AdamW steps with global batches 256/64/64. All stages use a peak learning rate of 2×10^{-4}, 2k warmup steps, $0.2\times$ decays at 50% and 75%, and BF16 distributed training on four A800 GPUs; generation uses 20 Euler steps. Component results are means and sample standard deviations over three seeds under matched data, budgets, and selection rules. Appendix A.6 gives complete settings.

4.2 Comparison with State-of-the-Art Methods

Quantitative results. Table 1 compares representative continuous, discrete, and hybrid generators. Among source-reported results, QuanTopo leads FID on both datasets and HumanML3D retrieval, with a smaller KIT-ML alignment gain. Appendix Table S3 gives checkpoint-based results under matched records, lengths, evaluator, and sample counts.

Qualitative comparison. Figure 4 compares compositional, cyclic, and coordinated actions with MoMask, MotionHiFlow, and SALAD under shared motion lengths and rendering. Keyframes are selected within each output to expose its progression while limiting overlap. Red phrases mark action and event content, while blue phrases mark spatial and trajectory content. QuanTopo preserves the late return in the first two examples; in the figure-eight example, its keyframes traverse both lobes and revisit the crossing, aligning local poses with the root path.

4.3 Component Analysis

Plan learning and conditional planning. Table 2 evaluates plan content and use. Full MPJPE measures $D(q^*, r^*)$, e_{anchor} measures $D(q^*, 0)$, and e_q evaluates descriptor predictability after freezing the representation. Topology supervision makes the plan more predictive of the trajectory descriptor. Removing the anchor degrades plan-only reconstruction despite improving full MPJPE, showing that joint reconstruction alone can shift information into the continuous channel. The last two rows share the same frozen Stage-A representation; geometric modulation then converts the predicted trajectory target into a generation gain. In the matched intervention, the predicted condition

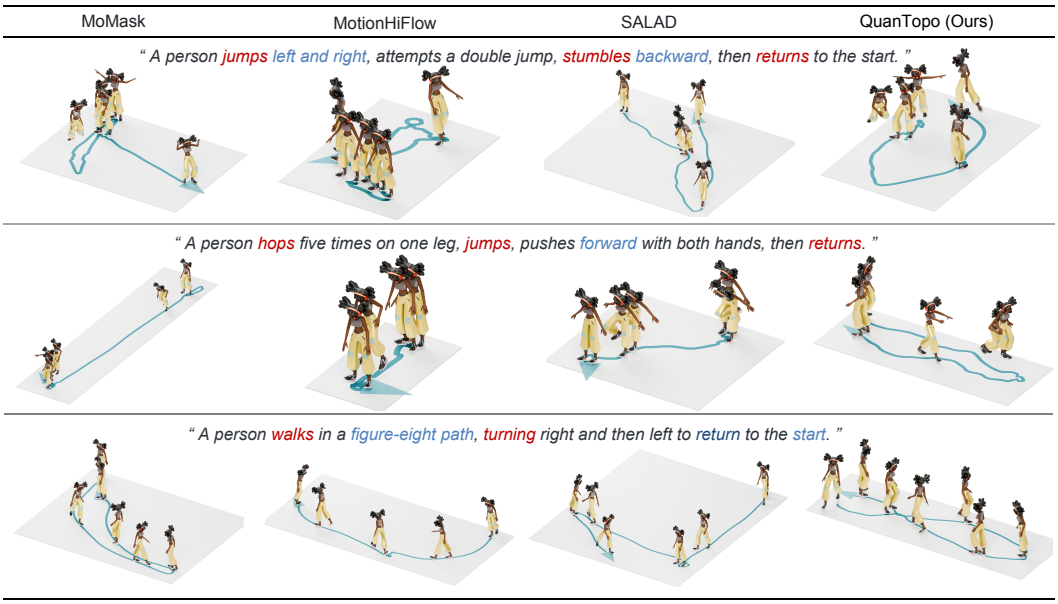

Figure 4: Visual comparisons on three HumanML3D test captions. Only key frames are displayed; overlaid curves show world-space root trajectories.

raises Seq. from 0.756 to 0.804 relative to zero modulation, whereas a mismatched condition lowers it to 0.708 (Appendix Table S6).

Table 2: Plan ablations on HumanML3D. Base-Hybrid contains none of the proposed components; each "w/o" variant removes the named component from the full model. Reconstruction errors are in mm. For w/o Topology Modulation and QuanTopo, the final three metrics are identical by construction because both use the same frozen Stage-A representation.

Variant	FID↓	R@1↑	$e_q \downarrow$	$e_{\text{anchor}} \downarrow$	MPJPE↓
Base-Hybrid	$0.061^{\pm 0.004}$	$0.541^{\pm 0.004}$	$0.284^{\pm 0.009}$	$712.4^{\pm 18.6}$	$231.6^{\pm 4.8}$
w/o Plan Topology	$0.043^{\pm 0.003}$	$0.568^{\pm 0.003}$	$0.219^{\pm 0.007}$	$\underline{428.9}^{\pm 12.7}$	$228.7^{\pm 4.1}$
w/o Anchor	$0.039^{\pm 0.003}$	$0.573^{\pm 0.003}$	$\underline{0.149}^{\pm 0.005}$	$812.5^{\pm 21.9}$	$\mathbf{215.9}^{\pm 3.7}$
w/o Topology Modulation	$\underline{0.035}^{\pm 0.002}$	$\underline{0.578}^{\pm 0.003}$	$\mathbf{0.141}^{\pm 0.004}$	$\mathbf{409.6}^{\pm 10.8}$	$\underline{224.8}^{\pm 3.9}$
QuanTopo	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{0.586}^{\pm 0.002}$	$\mathbf{0.141}^{\pm 0.004}$	$\mathbf{409.6}^{\pm 10.8}$	$\underline{224.8}^{\pm 3.9}$

Matched trajectory summaries. Table 3 compares persistence with global summaries under matched states, descriptor width, objectives, and training budget. ACF, PSD, EDM, and PH-vector use the same vector head, while Learned Set matches the persistence-interface parameter count. Every nonempty summary improves over no global supervision; persistence remains strongest, with a smaller but consistent capacity-matched margin over Learned Set and the wide-vector control (Appendix Table S9). Semantic order remains supplied by the ordered plan and generator.

Anatomical construction and continuous generation. Table 4 separates hierarchy, anatomy, operator choice, and continuous sampling with matched controls. Shared and Flat-MPS remove the part hierarchy, Conditional-MLP preserves routing, and the non-anatomical control uses a fixed random partition. Hierarchical MPS modestly improves on its parameter-matched MLP counterpart, while anatomical routing improves on random routing. Under fixed conditions, deterministic regression produces the same output on every repeated draw, so its $V_{\text{joint}} = 0$ by construction; the nonzero flow value measures stochastic within-plan variation. With representation and decoder fixed, flow improves multimodality and FID without changing paired reconstruction MPJPE, supporting variation within a plan.

Table 3: Matched trajectory-summary controls on HumanML3D.

Summary	FID↓	R@1↑	Match↓	Seq.↑	MM↑
None	$0.055^{\pm 0.004}$	$0.549^{\pm 0.004}$	$2.755^{\pm 0.012}$	$0.681^{\pm 0.018}$	$1.820^{\pm 0.061}$
ACF	$0.041^{\pm 0.003}$	$0.565^{\pm 0.003}$	$2.701^{\pm 0.010}$	$0.728^{\pm 0.018}$	$1.897^{\pm 0.060}$
PSD	$0.043^{\pm 0.003}$	$0.568^{\pm 0.003}$	$2.689^{\pm 0.010}$	$0.716^{\pm 0.017}$	$1.934^{\pm 0.058}$
EDM	$0.035^{\pm 0.003}$	$0.571^{\pm 0.003}$	$2.674^{\pm 0.009}$	$0.758^{\pm 0.016}$	$1.925^{\pm 0.059}$
Learned Set	$0.033^{\pm 0.003}$	$0.574^{\pm 0.003}$	$2.661^{\pm 0.009}$	$0.766^{\pm 0.016}$	$1.961^{\pm 0.058}$
PH-vector	$\underline{0.031}^{\pm 0.002}$	$\underline{0.579}^{\pm 0.003}$	$\underline{2.648}^{\pm 0.009}$	$\underline{0.783}^{\pm 0.015}$	$\underline{1.992}^{\pm 0.056}$
Point-set (QuanTopo)	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{0.586}^{\pm 0.002}$	$\mathbf{2.623}^{\pm 0.008}$	$\mathbf{0.804}^{\pm 0.014}$	$\mathbf{2.083}^{\pm 0.058}$

Table 4: Anatomical-decoder and continuous-generator ablations on HumanML3D; distances are in mm.

Group	Variant	FID↓	R@1↑	MM↑	MPJPE↓	Dec. params	V_{joint}
Decoder	Shared	$0.037^{\pm 0.003}$	$0.573^{\pm 0.003}$	$1.910^{\pm 0.052}$	$239.8^{\pm 5.2}$	14.3M	—
	Flat-MPS	$0.033^{\pm 0.003}$	$0.581^{\pm 0.003}$	$1.984^{\pm 0.057}$	$232.6^{\pm 4.6}$	15.1M	—
	Hier. Conditional-MLP	$0.031^{\pm 0.002}$	$0.579^{\pm 0.003}$	$\underline{2.018}^{\pm 0.054}$	$\underline{229.7}^{\pm 4.2}$	15.3M	—
	Hier. MPS, non-anatomical	$\underline{0.030}^{\pm 0.003}$	$\underline{0.583}^{\pm 0.003}$	$1.997^{\pm 0.059}$	$230.4^{\pm 4.5}$	15.4M	—
	Hier. decoder (QuanTopo)	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{0.586}^{\pm 0.002}$	$\mathbf{2.083}^{\pm 0.058}$	$\mathbf{224.8}^{\pm 3.9}$	15.4M	—
Generator	Regression	$\underline{0.048}^{\pm 0.003}$	$\underline{0.565}^{\pm 0.003}$	$\underline{1.362}^{\pm 0.044}$	$224.8^{\pm 3.9}$	15.4M	0.0
	Flow (QuanTopo)	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{0.586}^{\pm 0.002}$	$\mathbf{2.083}^{\pm 0.058}$	$224.8^{\pm 3.9}$	15.4M	$84.7^{\pm 3.1}$

Composite-action evaluation. Table 5 evaluates four methods on a prespecified 180-caption HumanML3D test subset whose captions contain at least two identifiable events with explicit precedence. Before viewing outputs, we retain one caption per motion ID and sample 60 motions from each train-set duration tertile, balancing duration within a fixed annotation budget. Each method contributes one output per caption from a prespecified checkpoint. Three method-blind annotators yield edgewise Seq., Coverage, and Order | Covered. Relative to the strongest baseline, SALAD, QuanTopo improves paired Seq. by $+0.022$ (95% CI $[+0.006, +0.039]$). On the 149 motions covered by both, Order | Covered improves by $+0.015$ ($[+0.002, +0.029]$), ruling out coverage alone; standalone Coverage is inconclusive.

Table 5: Composite-action organization on 180 captions; brackets are 95% paired-bootstrap confidence intervals.

Method	Seq.↑	Coverage↑	Order|Cov.↑	N_{cov}
MoMask	$0.724\,[0.697, 0.751]$	$0.806\,[0.750, 0.856]$	$0.867\,[0.831, 0.901]$	145/180
MotionHiFlow	$0.759\,[0.735, 0.782]$	$0.833\,[0.780, 0.878]$	$\underline{0.891}\,[0.859, 0.920]$	150/180
SALAD	$\underline{0.782}\,[0.759, 0.804]$	$\underline{0.856}\,[0.806, 0.894]$	$0.886\,[0.855, 0.915]$	154/180
QuanTopo	$\mathbf{0.804}\,[\mathbf{0.783}, \mathbf{0.825}]$	$\mathbf{0.872}\,[\mathbf{0.824}, \mathbf{0.911}]$	$\mathbf{0.904}\,[\mathbf{0.875}, \mathbf{0.930}]$	157/180

Taken together, the matched controls separate plan content from use: topology supervision and the anchor shape what q retains, while anatomical routing and flow govern how the plan is realized.

5 Conclusion

QuanTopo frames text-to-motion generation as the coordination of planning and realization. Its topology-directed discrete plan captures global motion organization, while anatomical synthesis and conditional flow translate that organization into diverse motions. Results across standard benchmarks, matched controls, and composite-action evaluation support jointly designing what an intermediate plan represents and how it is used by the generator. More broadly, this perspective suggests that structured intermediate representations can preserve long-range organization without sacrificing realization-level flexibility.

AI USE STATEMENT

We used generative AI tools only for language polishing and internal consistency checks of the manuscript. We did not use generative AI tools to generate model implementations, experimental data, annotations, participant responses, figures, or any core research content. All final interpretations, analyses, and conclusions are made by the authors.

ETHICS STATEMENT

The user study involved adult volunteers recruited through student and personal networks. Participation was unpaid and voluntary; participants were informed of the task and data use before participation and could withdraw at any time. Responses were stored under anonymous IDs, and no identifying or demographic information was collected. Participants evaluated only rendered motion videos; no sensitive or biometric data were collected.

REPRODUCIBILITY STATEMENT

Model architecture and objectives are specified in Section 3 and Appendix A; implementation settings appear in Appendix A.6; and evaluation, sampling, and user-study protocols are documented in Appendix C. Additional controls and diagnostics are reported in the remaining appendices. An anonymized implementation is included in the supplementary material.

REFERENCES

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18 (8):1–35, 2017.

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal action compositions for 3D humans. In *International Conference on 3D Vision*, 2022.

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9984–9995, 2023.

Sheldon Axler. *Linear Algebra Done Right*. Springer, 4th edition, 2024.

Nikita Balabin, Daria Voronkova, Ilya Trofimov, Evgeny Burnaev, and Serguei Barannikov. Disentanglement learning via topology. In *International Conference on Machine Learning*, pp. 2474–2504, 2024.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. PersLay: A neural network layer for persistence diagrams and new graph topological signatures. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2786–2796. PMLR, 2020.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18000–18010, 2023.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007. doi: 10.1007/s00454-006-1276-5.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. MotionLCM: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pp. 390–408, 2024. doi: 10.1007/978-3-031-72640-8_22.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533, 2002.

Rickard Brüel Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, and Primoz Skraba. A topology layer for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1553–1563, 2020.

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. 5152–5161, 2022.

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.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. SALAD: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202, 2016.

Guillermo Legarda Herranz, Emanuele Garone, and Mauro Birattari. Persistent homology as a universal metric for collective motion. *Research Square preprint*, 2026. doi: 10.21203/rs.3.rs-8672197/v1.

Heng Li, Xiaotong Lin, Ling-An Zeng, Yulei Kang, Shuai Li, and Jian-Fang Hu. MotionHiFlow: Text-to-motion via hierarchical flow matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9352–9363, 2026.

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

Jiahao Liu, Huiyang Xiao, Yuhang Li, Yuning Qiu, and Guoxu Zhou. PB-MLD: Part-based motion latent diffusion model. *Image and Vision Computing*, 174:106118, 2026. doi: 10.1016/j.imavis.2026.106118.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jianxin Ma, Shuai Bai, and Chang Zhou. Pretrained diffusion models for unified human motion synthesis. *arXiv preprint arXiv:2212.02837*, 2022.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*, 2024.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *International Conference on Machine Learning*, pp. 7045–7054, 2020.

Ivan V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011. doi: 10.1137/090752286.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, 2022.

Mathis Petrovich et al. Multi-track timeline control for text-driven 3D human motion generation. *arXiv preprint arXiv:2401.08559*, 2024.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. MMM: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1546–1555, 2024a.

Ekkasit Pinyoanuntapong et al. BAMM: Bidirectional autoregressive motion model. In *European Conference on Computer Vision*, 2024b.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. doi: 10.1089/big.2016.0028.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Tencent Hunyuan 3D Digital Human Team. HY-Motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

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.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Hequan Wang, Xuean Chen, Jiaxu Zhang, Zhengbo Zhang, and Zhigang Tu. DC-Motion: Decoupling structure and details via discrete-continuous tokens for human motion generation. *arXiv preprint arXiv:2606.14721*, 2026. URL https://arxiv.org/abs/2606.14721v2.

Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision*, pp. 3–19, 2018.

Zhenyu Xie, Yang Wu, Xuehao Gao, Zhongqian Sun, Wei Yang, and Xiaodan Liang. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6252–6260, 2024. doi: 10.1609/aaai.v38i6.28443.

Weihao Yuan et al. MoGenTS: Motion generation based on spatial-temporal joint modeling. In *Advances in Neural Information Processing Systems*, 2024.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. Light-T2M: A lightweight and fast model for text-to-motion generation. In *AAAI Conference on Artificial Intelligence*, 2025.

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.

Jianrong Zhang, Hehe Fan, and Yi Yang. EnergyMoGen: Compositional human motion generation with energy-based diffusion model in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. FineMoGen: Fine-grained spatio-temporal motion generation and editing. In *Advances in Neural Information Processing Systems*, 2023b.

Mingyuan Zhang et al. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023c.

Pengfei Zhang, Pinxin Liu, Pablo Garrido, Hyeongwoo Kim, and Bindita Chaudhuri. KinMo: Kinematic-aware human motion understanding and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11187–11197, 2025b.

Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. EMDM: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, 2024.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019.

Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22968–22977, 2025.

Supplementary Material

Contents

Appendix A: Model and Optimization Details
Appendix B: Interpretive Analysis
Appendix C: Evaluation and Sampling Protocols
Appendix D: Extended Experiments
Appendix E: Representation Design Studies
Appendix F: Event Coverage and Conditional-Order Diagnostics
Appendix G: Supplementary Mechanism Analysis

A Model and Optimization Details

This appendix follows the dependency chain trajectory target → motion representation → anatomical synthesis → text-conditioned generation → optimization. Trajectory descriptors, motion features, and continuous latents use separate coordinate systems and statistics connected by explicit transformations.

A.1 Trajectory Filtration and Persistence Readout

Let the state dimension in Equation 2 be d_s. We uniformly select $N = \min(T, 32)$ valid states, rounding equally spaced indices when $T > N$. With selected states s_i,

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i, \qquad \sigma_s = \sqrt{\frac{1}{Nd_s}\sum_{i=1}^{N}\|s_i - \bar{s}\|_2^2}, \qquad \tilde{s}_i = \frac{s_i - \bar{s}}{\max(\sigma_s, \epsilon)}. \tag{A.1}$$

Pose, root, and displacement blocks share one RMS scale with $\epsilon = 10^{-6}$. The normalization is invariant to a shared positive scale and coordinatewise constant translation when the floor is inactive, but not necessarily to a local limb rescaling or a change in speed. It preserves the blocks' original relative scale, so both dimension and variance affect their distance contributions. Assigning evenly spaced $\tau_i \in [0, 1]$, we use

$$d_{ij} = \sqrt{\frac{\|\tilde{s}_i - \tilde{s}_j\|_2^2}{d_s} + \lambda_t^2|\tau_i - \tau_j|^2}. \tag{A.2}$$

We symmetrize the k-nearest-neighbor graph and add edges between adjacent sampled times, with effective degree $\min(k, N-1)$. Vertices enter at zero, edges at their distances, and triangles at the maximum value of their three edges. At the largest original edge scale, a terminal cone connects the first vertex to all remaining vertices and adds newly completed triangles, making residual one-dimensional classes finite. Simplices are ordered by filtration value, dimension, and vertex indices, and H_0/H_1 are computed over $\mathbb{F}_2$; the unique infinite H_0 class is removed. We add 10^{-12} before square roots. Neighbor ties use ascending target index; persistence ties use birth, death, then reduction-output order.

For each homology dimension, points below a lifetime threshold are removed and at most the M longest-lived points are retained, with zero padding masked out. Let $\mathcal{P}_h$ contain birth b_i^h, death d_i^h, lifetime $\ell_i^h = d_i^h - b_i^h$, and weight w_i^h. With $g(a) = a/(1+a)$, grid centers $\mu_{uv} \in [0,1]^2$, and bandwidth σ_P,

$$z_h[u, v] = \sum_i w_i^h \ell_i^h \exp\left[-\frac{\|[g(b_i^h), g(\ell_i^h)] - \mu_{uv}\|_2^2}{2\sigma_P^2}\right]. \tag{A.3}$$

Flattening and concatenating $h = 0, 1$ produces d_z coordinates. Reference points have $w_i^h = 1$; text-predicted points have soft existence weights. Reference and generated motions share sampling, normalization, filtration, selection, grid, and bandwidth. Training-motion diagrams are precomputed for representation and text supervision. Validation and test diagrams are computed only for evaluation. Gradients pass through the plan readout and the differentiable embedding of predicted points.

A.2 Motion Features and Paired Encoders

The network input is a dataset-normalized RIC sequence containing root angular/horizontal increments, root height, root-relative joint positions, non-root 6D rotations, local joint velocities, and four foot-contact channels. Its width is $12J - 1$: 263 for HumanML3D and 251 for KIT-ML. Physical joints X_t used by topology and evaluation are recovered from these features. Motion-feature and continuous-latent normalization are distinct transformations.

E_q uses a width-5 input convolution, two width-4 stride-2 convolutions, temporal residual blocks with dilation 1/2/4, and LayerNorm. E_r concatenates motion features with a length-aware linear upsampling of quantized embeddings, then uses a width-3 convolution, dilation 1/2/4/8 blocks, and a pointwise output projection. Residual blocks use padding-aware GroupNorm (Wu & He, 2018), SiLU, two temporal convolutions, and dropout. Neither encoder directly reads text or sample topology. For $T \geq 4$, $L = \lfloor T/4 \rfloor$. The decoder uses native r, while flow uses

$$y = (r - \mu_r) \oslash \sigma_r, \qquad r = \mu_r + \sigma_r \odot y. \tag{A.4}$$

Statistics are fitted to valid training frames with a positive scale floor. Thus the anchor $r = 0$ corresponds to $y = -\mu_r \oslash \sigma_r$, whereas $y = 0$ corresponds to $r = \mu_r$. Clean targets, accepted roll-in targets, and samples use the same frozen transform.

A.3 Anatomical Operators and Whole-Body Readout

Each part site is conditioned on plan and part identity; each joint site on r_t, its routed part state, and ι_j. For local core $A^{(b)} \in \mathbb{R}^{\chi_{b-1} \times d \times \chi_b}$, let v_b be a learned unit-normalized environment probe. A site-specific affine map produces (γ_b, β_b) and

$$M_b = \sum_k A^{(b)}_{:,k,:}(1 + \gamma_{b,k})v_{b,k} + \langle \beta_b, v_b \rangle \mathbf{1}\mathbf{1}^\top. \tag{A.5}$$

$b = 1, \ldots, B$ indexes chain sites; $B = 6$ for parts and $B = J$ for joints. Boundary bonds satisfy $\chi_0 = \chi_B = 1$. With $\mathcal{N}(a) = a/\max(\|a\|_2, 10^{-6})$, excluding-site environments are

$$L_1 = 1, \quad L_{b+1} = \mathcal{N}(L_b M_b), \qquad R_B = 1, \quad R_{b-1} = \mathcal{N}(M_b R_b), \tag{A.6}$$

and the pre-normalized local physical output and its vector normalization are

$$\tilde{f}_{b,k} = (1 + \gamma_{b,k}) L_b A^{(b)}_{:,k,:} R_b + \beta_{b,k}(L_b \mathbf{1})(\mathbf{1}^\top R_b), \qquad f_b = \mathrm{LN}(\tilde{f}_b), \quad \tilde{f}_b = (\tilde{f}_{b,1}, \ldots, \tilde{f}_{b,d}). \tag{A.7}$$

LayerNorm acts across the physical vector. No extra projection or residual is hidden inside this readout. The construction retains adjacent MPS bonds while conditional maps and normalization make it a nonlinear feature operator.

Part assignment is fixed and separate from contraction order. In zero-based HumanML3D indices, head is $(12, 15)$, left arm $(13, 16, 18, 20)$, right arm $(14, 17, 19, 21)$, left leg $(1, 4, 7, 10)$, and right leg $(2, 5, 8, 11)$; remaining joints form the trunk. For KIT-ML, these are (4), $(8, 9, 10)$, $(5, 6, 7)$, $(16, 17, 18, 19, 20)$, and $(11, 12, 13, 14, 15)$. Joint contraction follows a depth-first traversal with children sorted by index.

A learned score over $F_{t,j} + e^{\mathrm{readout}}_j$ gives jointwise softmax weights. Their weighted projected features form a frame representation, processed by temporal residual blocks and separate LayerNorm MLPs for root channels and non-root rotations. The four root channels are half-yaw increment ω_t, local horizontal velocity (v^x_t, v^z_t), and height h_t. Under the dataset quaternion convention,

$$\theta_1 = 0, \qquad \theta_t = \sum_{i<t} \omega_i, \qquad Q_t = (\cos\theta_t, 0, \sin\theta_t, 0),$$
$$p^{xz}_t = \sum_{i<t} [\mathrm{Rot}(Q^{-1}_{i+1})(v^x_i, 0, v^z_i)]_{xz}, \qquad p^y_t = h_t. \tag{A.8}$$

Local FK starts from $G_{t,0} = I_3$ and $X^{\mathrm{local}}_{t,0} = 0$, recursively applies Equation 7, and then applies recovered root orientation and translation. Canonical offsets come from the dataset skeleton and training statistics. Joint differences are transformed back to root-local coordinates for velocity channels; contacts are reconstructed by smooth gates over foot height and squared displacement. This training-time RIC contact output is distinct from the near-ground skating metric in Appendix C.2.

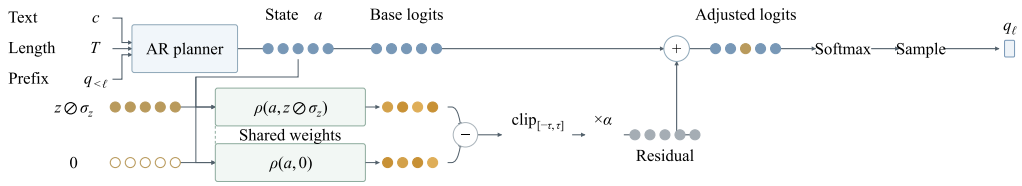

Figure S1: Zero-preserving planner modulation. The clipped difference between the shared modulator's responses to $z \oslash \sigma_z$ and zero corrects the base logits. This residual vanishes exactly when $z = 0$.

A.4 Text Predictor, Planner, and Conditional Flow

Frozen CLIP supplies token and sentence features. A two-layer SiLU MLP embeds $T/196$ and adds it to six learned part queries before they attend to valid tokens and undergo two gated message-passing rounds on a trunk-centered star graph. Mean/max query pools and the sentence feature feed independent H_0/H_1 point-set heads. For raw slot outputs (η_i, δ_i, o_i),

$$\hat{b}_i = \text{softplus}(\eta_i), \quad \hat{d}_i = \hat{b}_i + \text{softplus}(\delta_i) + 10^{-5}, \quad \hat{w}_i = \text{sigmoid}(o_i). \tag{A.9}$$

This yields $\widehat{\mathcal{P}}_h = \{(\hat{b}_i^h, \hat{d}_i^h, \hat{w}_i^h)\}_{i=1}^M$ and $\hat{z} = \Psi(\widehat{\Phi}(c, T))$.

The causal Transformer planner consumes a shifted token prefix, start symbol, and learned positions, cross-attending to text. External T fixes token count and masks and conditions the topology head; duration is not inferred from text. Equation 9's two-layer SiLU modulator is shared across both calls and contains no independent stochastic layer. The full-resolution flow Transformer has width 384, six heads, and eight blocks. A framewise projection embeds 128-D y_u; flow time modulates blocks by AdaLN (Peebles & Xie, 2023). Text tokens and linearly upsampled plan embeddings form type-tagged cross-attention memory. Padding is excluded from attention and loss. Inference uses 20 uniform Euler steps and Equation A.4.

A.5 Objectives and Gradient Paths

The three training stages use

$$\begin{aligned} \mathcal{L}_A &= \mathcal{L}_{\text{rec}} + \lambda_v \mathcal{L}_{\text{vel}} + \lambda_{\text{vq}} \mathcal{L}_{\text{commit}} + \lambda_q \mathcal{L}_{q\text{-topo}} + \lambda_a \mathcal{L}_{\text{anchor}}, \\ \mathcal{L}_{B1} &= \lambda_z \mathcal{L}_{\text{diagram}} + \mathcal{L}_{\text{AR}} + \lambda_s \mathcal{L}_{\text{soft-plan}}, \qquad \mathcal{L}_{B2} = \mathcal{L}_{\text{flow}}. \end{aligned} \tag{A.10}$$

Stage A updates E_q, E_r, the plan readout f, the decoder D, and the EMA codebook; the persistence readout Ψ remains fixed. Stage B1 freezes the representation and learns the topology predictor and planner. Stage B2 further freezes the predictor and planner and learns only the flow. CLIP is always frozen; codebook update/reset stops after Stage A.

$\mathcal{L}_{\text{rec}}$ is valid-element mean smooth-L_1 error on normalized RIC plus mean absolute physical-joint error in units of the median non-root canonical bone length b_{ref}. The anchor uses the identical reduction at native $r = 0$. $\mathcal{L}_{\text{vel}}$ applies the same normalized absolute error to consecutive physical-joint displacements. EMA-VQ uses data initialization, dead-code reset, and commitment to stop-gradient codewords. Descriptor and reconstruction gradients reach E_q through $h_\ell + \text{sg}[e_{q_\ell} - h_\ell]$ (Bengio et al., 2013).

For diagram supervision, we apply the following construction separately to $h = 0, 1$ and suppress the homology superscript. Predicted $P = \{(p_i, \hat{w}_i)\}_{i=1}^M$ and reference $Q = \{(q_j, v_j)\}_{j=1}^M$ use $p_i, q_j \in \mathbb{R}^2$, predicted existence mass $\hat{w}_i \in [0, 1]$, and reference validity mass $v_j \in \{0, 1\}$. Each set is augmented by the other's diagonal projections $\pi_\Delta(b, d) = ((b + d)/2, (b + d)/2)$. Let $m = \sum_i \hat{w}_i + \sum_j v_j$ and, for $m > 10^{-12}$, normalized marginals $\boldsymbol{\nu}_P = [\hat{w}; v]/m$ and $\boldsymbol{\nu}_Q = [v; \hat{w}]/m$. Log-domain Sinkhorn iterations (Cuturi, 2013) produce a unit-mass coupling Π_ε with row/column marginals $\boldsymbol{\nu}_P, \boldsymbol{\nu}_Q$ under squared Euclidean cost C; diagonal-copy-to-copy cost is zero. For each homology dimension, $n_h = \sum_j v_j^h$ is the number of valid reference points. With the fixed stabilizer

Table S1: Network, target, and sampling configuration.

Component	Setting
Motion input	RIC 263/251 channels; 4–196 frames; deterministic seeded crop for longer motions; right padding excluded.
Persistence target	$N = \min(T, 32)$, $k = 6$, $\lambda_t = 0.15$; $M = 16$ points each for H_0/H_1; lifetime threshold 10^{-5}.
Descriptor	8×8 grid per homology dimension over $0, 1/7, \ldots, 1$; $\sigma_P = 0.12$; $d_z = 128$.
Plan	One-level EMA-VQ, $K = 512$, $d_e = 256$, stride 4; E_q width 256; f: 256–256–128 SiLU MLP.
Continuous encoder	Width 256, $d_r = 128$; dropout 0.1; GroupNorm with 32 groups.
Anatomical decoder	Part/joint physical widths 384; bonds 8/16; identity widths 256; shared frame width 384 and dilation 1/2/4 temporal blocks.
Text predictor	Frozen CLIP ViT-B/32 (512-D); two-layer length MLP; six queries, one 8-head text-attention layer, two graph rounds; two 512-hidden heads each output 16×3.
AR planner	Six layers, width 256, eight heads, FFN 1024, dropout 0.1; 64 plan positions (up to 256 frames).
Modulation	Shared two-layer 256-hidden SiLU network, zero-initialized final layer, output 512; no dropout; $\alpha = 0.1$, $\tau = 5$.
Conditional flow	128-D input/output, width 384, eight blocks, six heads, FFN 1536, dropout 0.1; time AdaLN and text/plan cross-attention.
Sampling	AR temperature 1 with no top-k/nucleus truncation; standard Gaussian; 20 uniform Euler steps; guidance scale 1.

$\delta = 10^{-8}$, define $\mathcal{T}_\varepsilon = \sqrt{m\langle \Pi_\varepsilon, C\rangle + \delta} - \sqrt{\delta}$, or zero when $m \le 10^{-12}$. Then

$$\mathcal{L}_{\text{diagram}} = \frac{\lambda_{\text{desc}}}{d_z}\|\hat{z} - S(x)\|_2^2 + \frac{1}{2}\sum_{h=0}^{1}\left[\lambda_{\text{OT}}\mathcal{T}_\varepsilon(\widehat{\mathcal{P}}_h, \mathcal{P}_h) + \lambda_{\text{mass}}\frac{\text{smoothL1}(\sum_i \hat{w}_i^h - n_h)}{M}\right]. \tag{A.11}$$

We use $\lambda_{\text{desc}} = 0.5$, $\lambda_{\text{OT}} = 0.1$, $\lambda_{\text{mass}} = 0.05$, $\varepsilon = 0.05$, 30 Sinkhorn row/column updates, and FP32. In Equation A.10, $\lambda_v = 0.5$, $\lambda_{\text{vq}} = 0.25$, $\lambda_a = 0.5$, $\lambda_z = 1$, and $\lambda_q = \lambda_s = 0.1/d_z$.

The predicted descriptor is not detached in Stage B1, allowing the AR objective to update the topology predictor through the planner. For teacher-forced probability $p_\theta(q_\ell = k)$, let $\bar{e}_\ell = \sum_k p_\theta(q_\ell = k)e_k$. The frozen readout and codebook retain input derivatives in

$$\mathcal{L}_{\text{soft-plan}} = \left\| f\left(\frac{\sum_\ell m_\ell \bar{e}_\ell}{\sum_\ell m_\ell}\right) - \text{sg}[S(x)]\right\|_2^2. \tag{A.12}$$

After 50,000 clean flow steps, 20% of samples attempt roll-in. The frozen planner proposes $\tilde{q}$ and the encoder recomputes $\tilde{r}^* = E_r(x, \tilde{q})$. With

$$e_{\text{clean}} = \mathcal{L}_{\text{rec}}(D(q^*, r^*), x), \qquad e_{\text{prop}} = \mathcal{L}_{\text{rec}}(D(\tilde{q}, \tilde{r}^*), x), \tag{A.13}$$

we accept if $e_{\text{prop}} \le 1.25 e_{\text{clean}}$ and otherwise retain the clean pair. Accepted and clean targets use the same frozen latent transform; the 20% figure is a proposal probability, not an acceptance rate.

A.6 Implementation Configuration

Both datasets use the same capacity and optimization budgets; skeletons, input widths, and training statistics differ. Representation and flow budgets follow the scale used by MotionHiFlow (Li et al., 2026); planning uses an independent 200k-step budget. AdamW is used throughout (Loshchilov & Hutter, 2019).

Training samples motion IDs uniformly and then one associated caption. RIC statistics are fitted on the full training split. After freezing the representation, σ_z is fitted with equal motion weighting and (μ_r, σ_r) over valid frames. Validation/test records never fit statistics. Under DDP, each loss is reduced over valid elements within a sequence and then equally over sequences; gradient

Table S2: Optimization and statistical configuration. Steps are optimizer updates; batch sizes are totals over four GPUs.

Item	Setting
Hardware/precision	4×A800, DDP, BF16; optimizer state, diagram matching, and MPS environments in FP32.
Budgets	Representation 300k steps, batch 256; planning 200k, batch 64; flow 200k, batch 64; initial peak LR 2×10^{-4}.
Optimizer	AdamW, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01 excluding bias/norm, global gradient clipping 1.
Schedule	2k-step linear warm-up; multiply LR by 0.2 at 50% and 75% of each stage; optimizer reinitialized per stage.
Codebook EMA	Decay 0.99, count smoothing 10^{-5}; 512 data-initialized codes; globally synchronized counts and feature sums.
Dead-code reset	From step 100, every 100 steps; codes with EMA count < 1 reset from valid global-batch features; disabled after Stage A.
Statistics	RIC/r population-standard-deviation floors $10^{-6}/10^{-4}$; σ_z floor 10^{-3}; descriptor modulation scales but does not center.
Condition dropout	Flow only: probability 0.1 replaces text memory with a learned null token while retaining plan, length, and flow time.
Roll-in	First 50k steps clean; then proposal probability 0.2 and Equation A.13 acceptance factor 1.25.
Selection	Validate/save every 5k; minimize $\mathcal{L}_A/\mathcal{L}_{B1}$ for first stages; select flow by mean FID on fixed validation records and five generation seeds.
Randomness	Training seeds 1/2/3; 20 standard evaluation repeats; 10k paired bootstrap samples; separate AR and Gaussian random streams.

accumulation is one. Frozen modules are in evaluation mode for statistics, validation, roll-in, and inference.

Equation A.2 bounds the temporal contribution by 0.15 while temporal-adjacency edges guarantee connectivity. Sixteen points per homology dimension cap descriptor capacity and make the augmented matching matrix at most 32×32. With grid spacing $1/7$ and bandwidth 0.12, a Gaussian centered on one grid point retains about 0.49 response at an adjacent point. Equation 9 guarantees $\|\ell_\theta - \ell_0\|_\infty \le 0.5$; at temperature 1, any pairwise probability ratio changes by a factor in $[e^{-1}, e]$. For roll-in, the acceptance bound gives $\mathbb{E}[e_{\text{train}} \mid x] \le (0.8 + 0.2 \times 1.25) e_{\text{clean}} = 1.05 e_{\text{clean}}$.

A.7 Tensor Interfaces and End-to-End Algorithms

Tensor interfaces, omitting batch dimensions.

Object	Meaning	Shape/domain
x	physical joints	$T \times J \times 3$
RIC	normalized motion input	$T \times (12J - 1)$
H_c	text context	$N_c \times d_c$
q	plan indices	$\{1, \dots, K\}^L$
$\bar{q}$	upsampled embeddings	$T \times d_e$
r, y	native/normalized realization	$T \times d_r$
$\mathcal{P}_h$	points and mask	$M \times 2, M$
$\widehat{\mathcal{P}}_h$	birth, death, existence	$M \times 3$
$z, \hat{z}, z_q$	reference, predicted, and plan descriptors	d_z
U, F	part/joint features	$T \times 6 \times d_p, T \times J \times d_f$
root, $\{R_{t,j}\}$	root channels and local rotations	$T \times 4, T \times (J-1) \times 3 \times 3$

Algorithm 1 Stagewise training and text-to-motion inference.

Require: Training motions and captions $\mathcal{D}$; test text c and length T

Fit statistics and cache persistence targets $S(x)$ using training motions only; compute validation/test descriptors only for evaluation.

Stage A: topology-directed representation

Update E_q, the EMA codebook, E_r, f, and D with $\mathcal{L}_A$, including $D(q^*, r^*)$ and $D(q^*, 0)$.

Select and freeze the representation; fit realization and descriptor scales.

Stage B1: text-conditioned planning

Predict $\widehat{\Phi}(c, T)$ and $\hat{z}$; teacher-force q^* and update only the topology predictor and planner with $\mathcal{L}_{B1}$.

Select and freeze the predictor and planner.

Stage B2: continuous realization

Use clean pairs initially; thereafter propose and reconstruction-gate roll-in plans by Equation A.13.

Normalize the accepted target, sample (ξ, u), form y_u, and update only ϕ with Equation 10.

Inference

Compute $(H_c, \widehat{\Phi}(c, T), \hat{z})$, set $L = \lfloor T/4 \rfloor$, and autoregressively sample $q_{1:L}$ using Equation 9.

Draw $y_0 = \xi \sim \mathcal{N}(0, I)$ and apply n Euler updates $y_{i+1} = y_i + n^{-1} v_\phi(y_i, i/n; H_c, q, T)$.

Inverse-normalize $r = \mu_r + \sigma_r \odot y_n$ and return $D(q, r)$ after anatomical routing, temporal readout, and forward kinematics.

B Interpretive Analysis

B.1 What Pooled Plan Supervision Targets

For fixed L, K, random plan Q satisfies $H(Q) \le L \log_2 K$. For square-integrable $Z = S(X)$ and $B = b(Q)$,

$$\inf_f \mathbb{E}\|Z - f(B)\|_2^2 = \mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(Z \mid B). \tag{B.1}$$

This is the conditional-mean principle for squared loss (Hastie et al., 2009): the optimal readout is $\mathbb{E}[Z \mid B]$. Because B is permutation-invariant, the objective constrains aggregate plan content; it does not alone establish what the full ordering contains. That value is tested by sequential-event experiments.

B.2 Plan-Readout Agreement and Within-Plan Variation

Fix (c, q, T) and let $R \sim p_\phi(r \mid c, q, T)$, $Y = S(D(q, R))$, $\mu_q = \mathbb{E}[Y]$, and $m_q = R_q(q)$. Then

$$\mathbb{E}\|Y - m_q\|_2^2 = \underbrace{\|\mu_q - m_q\|_2^2}_{\text{plan-readout discrepancy}} + \underbrace{\mathbb{E}\|Y - \mu_q\|_2^2}_{\text{within-plan dispersion}}. \tag{B.2}$$

For an independent conditional draw Y',

$$\mathbb{E}\|Y - Y'\|_2^2 = 2\mathbb{E}\|Y - \mu_q\|_2^2 \le 2\mathbb{E}\|Y - m_q\|_2^2. \tag{B.3}$$

Both follow by expanding around the conditional mean. We report these descriptor quantities with joint-space variation and text alignment to distinguish geometrically coherent yet diverse families from collapsed outputs.

B.3 Local Geometric Intuition

Fix q on a differentiable branch of decoder and summary. With $A = J_D(r)$, $B = J_S(D(q, r))$, $r_A = \mathrm{rank}(A)$, and $r_{BA} = \mathrm{rank}(BA)$, rank–nullity gives $\dim(A(\ker BA)) = r_A - r_{BA}$ (Axler, 2024). Thus if $r_A > r_{BA}$, nonzero locally expressible directions preserve the summary to first order. This local result motivates within-plan variation but does not imply finite-amplitude invariance.

C Evaluation and Sampling Protocols

C.1 General Evaluation, Reconstruction, and Runtime

Each record stores split, motion/caption ID, length, model/evaluator versions, training seed, and generation seed. Training seeds measure training variability; repeated generation from one checkpoint measures sampling variability. Original-paper and unified local results are labeled separately.

Table S3: Unified evaluation using identical records, ground-truth motion lengths, evaluator, and generation counts. $\pm$ indicates a 95% confidence interval.

Method	FID↓	R@1↑	Match↓
HumanML3D			
MoMask	$0.045^{\pm 0.003}$	$0.521^{\pm 0.003}$	$2.958^{\pm 0.010}$
SALAD	$0.073^{\pm 0.004}$	$0.579^{\pm 0.004}$	$2.654^{\pm 0.011}$
MotionHiFlow	$0.032^{\pm 0.003}$	$0.563^{\pm 0.004}$	$2.691^{\pm 0.011}$
QuanTopo	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{0.586}^{\pm 0.002}$	$\mathbf{2.623}^{\pm 0.008}$
KIT-ML			
MotionHiFlow	$0.135^{\pm 0.007}$	$0.482^{\pm 0.005}$	$2.552^{\pm 0.014}$
QuanTopo	$\mathbf{0.122}^{\pm 0.006}$	$\mathbf{0.489}^{\pm 0.004}$	$\mathbf{2.523}^{\pm 0.012}$

Unified reevaluation uses identical record lists, ground-truth lengths, and retrieval candidates. Three QuanTopo checkpoints use random streams 1/2/3; a baseline with one public checkpoint generates three independent streams. Differences are first computed within paired repeats and then averaged. A two-level paired bootstrap resamples the three repeats and, within each, motion IDs with replacement for 10,000 replicates; all captions of a motion remain grouped. FID is recomputed for both methods on every resample. For one-checkpoint baselines, the outer level combines their sampling variability with QuanTopo training variability and is not interpreted as baseline training uncertainty.

On HumanML3D, paired differences from MotionHiFlow are $\Delta\text{FID} = -0.005$ (95% CI $[-0.008, -0.002]$), $\Delta\text{R@1} = +0.023$ ($[+0.016, +0.030]$), and $\Delta\text{Match} = -0.068$ ($[-0.087, -0.049]$); differences from MoMask are -0.018, $+0.065$, and -0.335. On KIT-ML, the corresponding MotionHiFlow differences are -0.013, $+0.007$, and -0.029, with the R@1 interval $[-0.001, +0.015]$. Re-evaluated SALAD differences on HumanML3D are -0.046, $+0.007$, and -0.031, with R@1 CI $[+0.001, +0.013]$.

For reference pairing, $d_{\text{ref}} = \|S(x_{\text{gen}}) - S(x_{\text{ref}})\|_2$ uses the specified record, not a nearest post-hoc reference; condition agreement uses $d_{\text{cond}} = \|S(x_{\text{gen}}) - z_{\text{cond}}\|_2$, where z_{cond} is the descriptor supplied to the planner. We average over samples within caption, captions within motion, then motion IDs. MPJPE is mean valid-frame/joint Euclidean error in the shared initial RIC world frame, converted to mm and averaged over sequences, without per-frame root alignment or Procrustes fitting. End-to-end time includes text encoding, topology prediction, planning, flow, and decoding after warm-up and synchronization, excluding loading and rendering. Offline persistence cost is separate.

On one A800 in BF16 at batch size 1 and 196 frames, 100 warm-up and 1,000 synchronized trials give end-to-end latency $0.48^{\pm 0.01}$ s (95th percentile 0.50 s) and 4.1 GB peak memory; flow integration costs $0.35^{\pm 0.01}$ s. With 16 CPU workers, offline persistence extraction costs $12.6^{\pm 4.8}$ ms per motion and adds no inference latency. We do not treat raw latency reported by diffusion accelerators such as EMDM (Zhou et al., 2024) as directly comparable because hardware, sequence length, and timing boundaries differ.

C.2 Plan, Summary, and Motion-Structure Evaluation

R_q^{eval} is distinct from the generative readout. After each representation is frozen, a 256–256–128 SiLU MLP is initialized with seed 17 and trained for 10k steps with AdamW (LR 10^{-3}, decay 0.01, batch 256), selected every 500 steps by validation descriptor MSE. We define $e_q = \mathbb{E}\|R_q^{\text{eval}}(q^*) - S(x^*)\|_2$ and $e_z = \mathbb{E}\|\hat{z} - S(x^*)\|_2$, averaging per-sequence Euclidean descriptor errors. Shared representations reuse one evaluation readout.

Matched summary controls. ACF, PSD, EDM, and PH begin from the same sampled normalized states. ACF uses per-channel lags 1–16 divided by zero-lag second moment with floor 10^{-6} and appends a valid-lag mask. PSD uses orthonormal real FFT power $|F_{c,k}|^2$, log-transformed and interpolated at nonzero frequencies $m/32$, $m = 1, \ldots, 16$. EDM uses the pairwise distances in Equation A.2, bilinearly resizes the matrix to 32×32, and takes 496 strict upper-triangle entries. ACF/PSD/EDM are flattened and reduced to 128 dimensions by training-fitted centered PCA (Jolliffe & Cadima, 2016), with deterministic eigenvector signs, no whitening, and total variance

Table S4: Seq. by prespecified caption strata on the 180-caption subset.

Stratum	N	SALAD	QuanTopo	Δ
Two required events	104	0.805	0.821	+0.016
Three or more events	76	0.748	0.781	+0.033
Short / medium / long	60/60/60	0.804/0.786/0.756	0.818/0.807/0.787	+0.014/+0.021/+0.031
Locomotion–turn	64	0.788	0.812	+0.024
Upper/lower composition	61	0.779	0.802	+0.023
Posture transition	55	0.778	0.798	+0.020

Table S5: Physical quality on unified outputs. Support rate reports the fraction of valid near-ground transitions; real motion is shown as an unranked reference.

Method	Skating↓ (cm/s)	Support rate	Penetration↓ (mm)
Real motion	$1.21^{\pm 0.05}$	$0.294^{\pm 0.005}$	0.0
MoMask	$2.08^{\pm 0.09}$	$0.281^{\pm 0.007}$	$4.7^{\pm 0.3}$
MotionHiFlow	$\underline{1.93}^{\pm 0.08}$	$0.287^{\pm 0.006}$	$\mathbf{4.1}^{\pm 0.3}$
QuanTopo	$\mathbf{1.79}^{\pm 0.07}$	$0.290^{\pm 0.006}$	$\underline{4.3}^{\pm 0.2}$

matched to PH. All transforms and modulation scales use training data only. ACF, PSD, EDM, and PH-vector use the same 1536–512–128 vector prediction head and supervision locations. Learned Set uses its parameter-matched set encoder, while Point-set uses separate birth–death and existence heads; Table S9 reports the capacity-matched wide-vector control.

Event-order satisfaction. We form the subset from the official HumanML3D test split before generating or viewing method outputs. Eligible captions contain at least two visually identifiable events with an explicit precedence relation. We randomly retain at most one eligible caption per motion ID, divide motions into short, medium, and long bins using duration-tertile boundaries estimated on the training split, and sample 60 motions without replacement from each bin with seed 2026. The fixed size reflects the annotation budget while balancing motion duration. Required event instances and a DAG are annotated before viewing outputs; unrelated events may occur in either order. Three method-blind annotators mark onset/offset or absence. A required edge $a \to b$ receives a positive vote only when both appear and a ends no later than b begins; missing endpoints fail that edge. Majority voting gives edge labels, a sample score is its fraction of satisfied edges, and Seq. averages samples. All-event Coverage requires every event; Order | Covered averages order only over fully covered motions. They diagnose distinct error sources and do not satisfy a product identity. For the external four-method study, one prespecified checkpoint and one generated motion per caption are used for each method; code, weights, seeds, rendering, and anonymization are locked before rating. Fleiss' $\kappa = 0.74$.

On the 149 motions covered by both QuanTopo and SALAD, the paired Order | Covered gain is $+0.015$ (95% CI $[+0.002, +0.029]$), so the ordering gain is not explained only by event omission.

Foot skating and physical quality. All motion is in meters with dataset ground at zero and no per-output floor fit. A candidate support transition has both endpoint foot heights within 5 cm. Horizontal displacement divided by frame interval is averaged within each sequence and then across sequences with at least one candidate, in cm/s. Support rate is the candidate fraction among valid foot-joint transitions, not the fraction of frames labeled as foot contact. Ground penetration averages $\max(0, -y_{t,j})$ over four foot joints and valid frames, then sequences, in mm.

C.3 Conditional Sampling and Geometric Control

For each of three checkpoints we keep caption c and duration T fixed, sample $N_q = 4$ plans, and draw $N_r = 8$ continuous-noise samples per plan (Figure S2). Let $r_{ij} = R_\phi(c, q_i, T, \xi_j)$ be the realization obtained by integrating the flow from noise ξ_j, $Y_{ij} = S(D(q_i, r_{ij}))$, $\bar{Y}_i = \frac{1}{N_r}\sum_{j=1}^{N_r} Y_{ij}$,

Table S6: Topology-condition intervention using one checkpoint. Seq. uses 180 captions and FID uses the full test set. Reference is a test-motion oracle unavailable at inference.

Condition	FID↓	Match↓	Seq.↑	d_{ref} ↓	d_{cond} ↓
Zero	0.034	2.681	0.756	0.151	—
Mean	0.032	2.660	0.771	0.139	0.126
Mismatched	0.041	2.742	0.708	0.184	0.128
Predicted	**0.027**	**2.623**	**0.804**	**0.118**	**0.112**
Reference (oracle)	0.026	2.610	0.821	0.096	0.096

Table S7: Complete plan/realization decomposition. Brackets are 95% two-level checkpoint/caption bootstrap CIs.

Statistic	Value
V_{joint} (fixed plan)	84.7 mm $[80.1, 89.3]$
V_{local} (fixed plan)	61.8 mm $[58.2, 65.5]$
V_{root} (fixed plan)	47.3 mm $[43.7, 50.9]$
Within-family Matching Distance	2.657 $[2.638, 2.676]$
$\|\bar{Y}_i - m_i\|^2$	0.036 $[0.032, 0.040]$
Within-plan descriptor dispersion	0.021 $[0.018, 0.024]$
Between-plan mean dispersion	0.064 $[0.057, 0.071]$
V_{joint} (change plan, reuse ξ)	131.6 mm $[124.8, 138.5]$
V_{local} (change plan, reuse ξ)	78.9 mm $[74.6, 83.2]$
V_{root} (change plan, reuse ξ)	103.7 mm $[96.5, 110.9]$
$D(q_{\text{gen}}, 0)$ Seq.	0.773 $[0.748, 0.797]$
$D(q_{\text{gen}}, 0)$ Coverage	0.844 $[0.789, 0.888]$
$D(q_{\text{gen}}, \mu_r)$ MPJPE	438.7 mm $[426.1, 451.6]$
$D(q_{\text{gen}}, \mu_r)$ Seq.	0.761 $[0.735, 0.786]$
$D(q_{\text{gen}}, \mu_r)$ Coverage	0.828 $[0.771, 0.874]$

and $m_i = R_q(q_i)$. We report $\|\bar{Y}_i - m_i\|^2$, $\frac{1}{N_r}\sum_{j=1}^{N_r}\|Y_{ij} - \bar{Y}_i\|^2$, and variance among the N_q family means. Statistics are equally reduced over plans and captions within checkpoint and then over checkpoints. V_{joint}, V_{local}, and V_{root} are mean pairwise distances between, respectively, world-space joints, root-relative joints, and root trajectories. Fixed-plan variation averages all 28 realization pairs per plan; change-plan variation averages six plan pairs for each shared noise. Motions share initial frame and length with no framewise alignment. The organization rows below use generated plans q_{gen}; the encoded-plan anchor MPJPE $D(q^*, 0)$ is reported only in Table 2.

Condition interventions share seed-2 checkpoint, record list, stepwise AR random numbers, and Gaussian initial states. Zero removes modulation in scaled coordinates; Mean uses the equally motion-weighted training mean; Predicted uses text and supplied length; Reference and Mismatched read current/donor reference summaries. Donors are chosen within valid-length bins $[4, 64]$, $[65, 128]$, $[129, 196]$ with seed 2026 and a distinct motion ID. Only paired distances to the same donor support the directional conclusion. All nonzero conditions use the same scale transform; AR and flow have separate random streams.

Against Predicted, Zero increases Match by $+0.058$ (95% CI $[+0.041, +0.075]$) and reduces Seq. by 0.048 ($[-0.067, -0.029]$). Comparing Predicted and Mismatched outputs to the same donor, donor distance decreases from 0.176 to 0.128, a paired -0.048 ($[-0.057, -0.039]$), while text and order scores worsen.

Both fixed realization inputs retain meaningful plan-conditioned synthesis, although the trained native origin performs better. This comparison supports an operational synthesis role for q but not a coordinate-invariant claim that $r = 0$ denotes an absence of realization information.

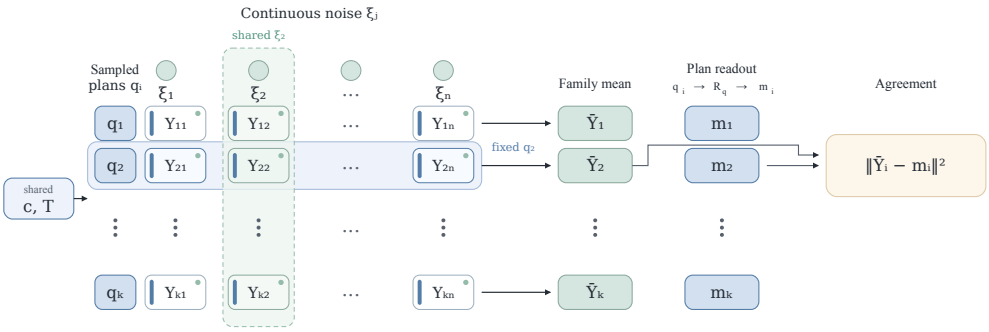

Figure S2: Conditional-family sampling under fixed caption c and duration T. Rows vary plans and columns vary continuous noise; the highlighted row and column show fixed-plan and shared-noise comparisons. We use $N_q = 4$ plans and $N_r = 8$ noise draws.

C.4 User-Study Protocol

We preselect 60 HumanML3D captions: 20 each for compositional, cyclic, and whole-body coordination, assigning each record to one primary stratum before viewing outputs. Each record fixes caption/motion ID, length, checkpoint, and random seed. QuanTopo is paired with SALAD, MotionHiFlow, and the record's ground-truth motion. Skeleton, material, camera, ground, frame rate, and controls are identical; only the motion differs.

We recruit 33 participants and exclude three who fail at least two of three prespecified attention checks, retaining 30. Two practice trials are not scored. Each video pair is judged for Text Alignment and Motion Realism, with left, right, or no clear difference as the response; question order and video side are randomized independently. A balanced incomplete block assigns the three comparator pairs for each of the 60 captions to four participants, yielding 720 video-pair presentations (24 per valid participant) and 1,440 dimension-level judgments.

Participants and data handling. Participants were adult volunteers recruited through student and personal networks. Participation was unpaid and voluntary; before beginning, participants were informed of the task and data use and could stop at any time. Responses were stored under anonymous IDs, and no identifying or demographic information was collected.

For each opponent and dimension we record win/tie/loss. The main statistic is $W/(W + L)$; ties remain in the full composition. A two-way participant/caption cluster bootstrap independently resamples both identifiers and weights observations by the product of their multiplicities. We use 10,000 replicates, seed 2026, and percentile 95% intervals.

The decisive Text Alignment win rates are 52.8% against SALAD (95% CI $[47.3, 58.2]$) and 58.6% against MotionHiFlow ($[53.3, 63.8]$); Motion Realism rates are 60.7% ($[55.5, 65.8]$) and 55.2% ($[49.8, 60.6]$). Against Ground Truth, they are 46.1% and 37.9%.

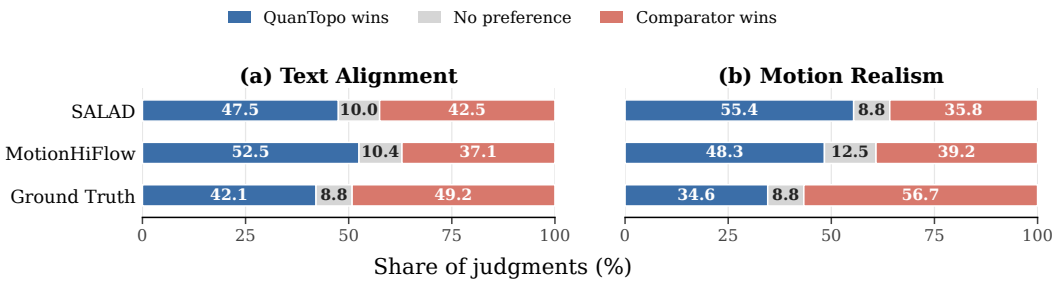

Figure S3: Blind pairwise user study. Each panel shows the shares of all judgments preferring QuanTopo, expressing no clear preference, or preferring the comparator. Decisive win rates and two-way participant/caption cluster-bootstrap confidence intervals are reported in the text.

D Extended Experiments

D.1 Additional Qualitative Comparisons

Figure S4 extends the main-text comparison with four additional HumanML3D test captions. The examples stress complementary sequence-level challenges: ordered turns and strikes, opposing curved paths, repeated direction changes followed by a return, and lateral steps that revisit the starting region. Within each row, methods use the same motion duration and renderer. Because independently generated motions need not realize events at identical times, keyframes are selected separately for each output to summarize its temporal progression while limiting overlap, rather than being forced to shared absolute frame indices or a fixed display count.

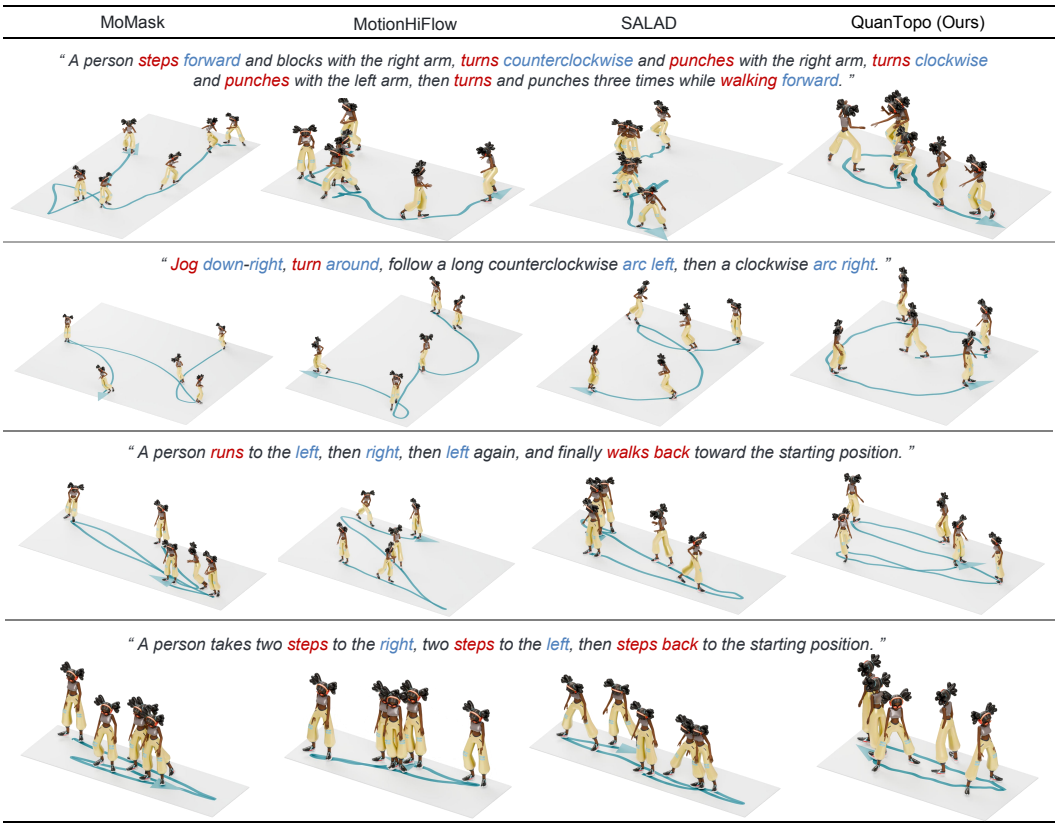

Figure S4: Additional visual comparisons on four HumanML3D test captions. Columns show MoMask, MotionHiFlow, SALAD, and QuanTopo. Each row uses a common motion duration and consistent rendering; keyframes are selected within each output to expose its progression while limiting overlap. Cyan curves show world-space root trajectories, and arrowheads indicate the final direction. Only key frames are displayed.

Across these examples, QuanTopo keeps direction reversals and late-stage returns visible while maintaining locally plausible action poses. The curved-route cases emphasize global path organization, whereas the striking and side-step cases test whether local actions remain synchronized with turns and displacement. These examples complement Figure 4; they are qualitative illustrations rather than replacements for the aggregate metrics and event-order diagnostics.

D.2 Cumulative Architecture Comparison

Table S8 expands the Base-Hybrid and full-model endpoints of Table 2. "Directed plan" adds topology and anchor objectives together with soft-plan supervision; "conditioned planner" adds text point-set prediction and geometric modulation. Planner-only changes share the preceding frozen representation, while representation changes receive matched downstream training. Directed-plan learning reduces reconstruction FID from 0.083 to 0.048 and generation FID from 0.061 to 0.049. Adding the conditioned planner at fixed representation further reduces generation FID to 0.037 and raises R@1 to 0.573, although MM drops slightly from 1.923 to 1.910. Adding the anatomical decoder reaches FID 0.027 and MPJPE 224.8 mm and recovers greater multimodal variation. The gains are cumulative on central quality/alignment metrics but not strictly monotone on every statistic.

Table S8: Cumulative architecture comparison. Plan combines topology and anchor objectives; the main text ablates them separately. Rec. metrics evaluate paired reconstruction, and the remaining columns evaluate text-conditioned generation.

Variant	Plan	Cond.	Anat. dec.	Rec. FID	MPJPE	Gen. FID	R@1	MM
Base-Hybrid				$0.083^{\pm 0.006}$	$231.6^{\pm 4.8}$	$0.061^{\pm 0.004}$	$0.541^{\pm 0.004}$	$1.790^{\pm 0.063}$
+ directed plan	✓			$0.048^{\pm 0.004}$	$239.8^{\pm 5.2}$	$0.049^{\pm 0.003}$	$0.557^{\pm 0.004}$	$1.923^{\pm 0.061}$
+ conditioned planner	✓	✓		$0.048^{\pm 0.004}$	$239.8^{\pm 5.2}$	$0.037^{\pm 0.003}$	$0.573^{\pm 0.003}$	$1.910^{\pm 0.052}$
QuanTopo	✓	✓	✓	$\mathbf{0.034}^{\pm 0.003}$	$\mathbf{224.8}^{\pm 3.9}$	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{0.586}^{\pm 0.002}$	$\mathbf{2.083}^{\pm 0.058}$

Table S9: Persistence point-set versus capacity-matched vector prediction. Descriptor error is e_z in the shared descriptor coordinates.

Interface	Params	Desc.↓	FID↓	R@1↑	MM↑
Vector-standard	0.853M	$0.129^{\pm 0.005}$	$0.031^{\pm 0.002}$	$0.579^{\pm 0.003}$	$1.992^{\pm 0.056}$
Vector-wide	1.648M	$0.116^{\pm 0.004}$	$0.029^{\pm 0.002}$	$0.584^{\pm 0.003}$	$\mathbf{2.102}^{\pm 0.061}$
Point-set	1.623M	$\mathbf{0.112}^{\pm 0.004}$	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{0.586}^{\pm 0.002}$	$\underline{2.083}^{\pm 0.058}$

D.3 Persistence-Point versus Vector Prediction

The following comparison fixes the PH target, representation, decoder, descriptor coordinates, modulation location, and training budget. Parameter counts cover biased linear layers in the prediction head only. Vector-standard uses 1536–512–128, Vector-wide 1536–832–384–128, and Point-set two independent 1536–512–48 heads. Vector-wide and Point-set differ by 1.5% in this count; each receives a retrained planner and flow. Widening the vector head absorbs part of the original gap, but the point-set interface still improves descriptor error, FID, and R@1 at matched capacity; Vector-wide has slightly higher MM. The comparison jointly changes interface and supervision and therefore supports their combined design rather than attributing the effect to point coordinates alone.

D.4 Sampling Efficiency and Motion Conditions

Quality–speed trade-off. QuanTopo has 62.5M trainable parameters, counting each shared parameter once and excluding the frozen CLIP text encoder and non-parameter EMA buffers, and uses 4.1 GB at inference for a 196-frame motion on one A800. Persistence targets are computed offline and add no inference cost. Holding the flow checkpoint, texts, lengths, plans, and Gaussian states fixed, 5/10/20/40 Euler steps give FID 0.049/0.033/0.027/0.026, R@1 0.575/0.582/0.586/0.587, and MM 2.020/2.041/2.083/2.090. End-to-end latencies are 0.22/0.31/0.48/0.82 s, of which integration costs 0.09/0.18/0.35/0.69 s. Moving from 20 to 40 steps improves FID by only 0.001 while adding 0.34 s, motivating the 20-step default.

Roll-in. With a 20% proposal rate in the enabled phase, 12.7% of training samples actually use re-encoded generated plans; accounting for the initial clean-only phase yields 9.5% over the full budget. The improvement is smaller than the central architecture ablations but consistent across all three metrics, so roll-in is retained as a robustness component rather than a principal contribution.

Length and compositional complexity. Using train-set tertiles, short/medium/long motions obtain R@1 0.601/0.588/0.565 and Matching Distance 2.540/2.612/2.781. Single-event, concurrent multi-event, and explicitly ordered captions obtain R@1 0.598/0.579/0.568; the ordered group has Seq. 0.804. Group definitions are fixed before inspecting generations. Late-sequence semantics, transitions, and scene contact remain designated failure categories for traceable qualitative review.

D.5 State Contributions and Topology Sensitivity

Pose, root position, and displacement account for 62.1%, 17.4%, and 20.5% of average squared state distance in training. Whole-vector RMS therefore does not erase root/displacement, although the higher-dimensional pose block dominates. Block-balanced uses training statistics to equalize mean contributions and retrains the complete system. H_0 and H_1 individually underperform their

Table S10: Roll-in ablation. Acceptance rates are reported among proposals, enabled-stage samples, and all flow-training samples.

Setting	FID↓	R@1↑	Match↓	Accept/Proposed	Accept/Enabled	Accept/Full
No roll-in	$0.030^{\pm 0.002}$	$0.582^{\pm 0.003}$	$2.651^{\pm 0.010}$	—	—	—
Gated roll-in	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{0.586}^{\pm 0.002}$	$\mathbf{2.623}^{\pm 0.008}$	$63.4^{\pm 1.8}\%$	$12.7^{\pm 0.4}\%$	$9.5^{\pm 0.3}\%$

Table S11: Topology-target sensitivity. $e_q/\sqrt{d_z}$ and $e_z/\sqrt{d_z}$ are mean per-sample normalized L_2 errors.

Target	$e_q/\sqrt{d_z}\downarrow$	$e_z/\sqrt{d_z}\downarrow$	FID↓	Seq.↑
H_0 only	0.0148	0.0120	0.032	0.772
H_1 only	0.0162	0.0131	0.037	0.748
$N = 16$	0.0134	0.0106	0.030	0.791
$N = 32$ (default)	0.0125	0.0099	**0.027**	<u>0.804</u>
$N = 64$	<u>0.0123</u>	<u>0.0098</u>	<u>0.028</u>	**0.806**
$\lambda_t = 0$	0.0133	0.0105	0.031	0.786
No terminal cone	0.0129	0.0101	0.029	0.796
Block-balanced	**0.0122**	**0.0097**	0.029	0.798

combination. $N = 64$ slightly improves descriptor error but not FID, so $N = 32$ controls preprocessing and target capacity. Removing temporal distance leaves adjacency edges in place; its Seq. decrease indicates distinct roles for edge inclusion and the metric's time term. Under the default, 18.6% of retained H_1 classes are closed by the terminal cone. Removing it and dropping infinite H_1 reduces Seq. from 0.804 to 0.796; the cone contributes information but does not solely explain the gain. Equalizing state blocks lowers descriptor error but not generation metrics.

E Representation Design Studies

These short-budget studies examine discrete-vocabulary utilization and plan-only decodability before training the full generator. Their values are not directly compared with the full-budget ablations.

E.1 Discrete Vocabulary Utilization

Each HumanML3D quantizer was screened for 1,000 steps, reporting window means over steps 800–1000. Under the same short budget, EMA-VQ produces more distributed code use and lower quantization and hybrid reconstruction errors than the tested SimVQ and FSQ configurations (Zhu et al., 2025; Mentzer et al., 2024). The KIT-ML EMA run obtains perplexity 132.929, rolling coverage 502.457/512, and largest-code share 0.0475.

E.2 Plan Decodability and Reconstruction

Four representation variants are evaluated on the same 2,451 held-out HumanML3D records over three seeds. Plan-only anchoring lowers plan-only MPJPE but raises full-reconstruction error for both decoders, demonstrating that optimizing joint reconstruction alone and requiring the plan itself to reconstruct motion allocate information differently. All channels are encoded and decoded with the frozen benchmark-mean topology $\bar{z}$. Here Full is $D(q^*, r^*; \bar{z})$, q-only is $D(q^*, 0; \bar{z})$, and mean-plan+r is $D(\bar{q}, r^*; \bar{z})$, where $\bar{q}$ is the masked mean of quantized plan embeddings; $\bar{q}$ and $\bar{z}$ are frozen and shared by every evaluated record. The accompanying probes find stronger phase and cycle predictability in r. Together with the reconstruction interventions, they support complementary computational biases but not an identifiable structure/detail factorization.

Table S12: HumanML3D quantizer screening. Active/512 is rolling coverage averaged across windows; Rec. is the logged hybrid reconstruction objective.

Quantizer	PPL	Active	Max share	Quant.	Rec.	q-topo
EMA-VQ	160.000	493.657	0.0789	0.2653	15.135	0.2187
SimVQ	1.868	170.690	0.7780	1.2474	17.020	0.2324
FSQ	21.572	452.600	0.2985	1.1249	17.941	0.5468

Table S13: Stage-A reconstruction records. MPJPE in mm, mean $\pm$ sample standard deviation over three seeds.

Representation	Full↓	q-only↓	Mean-plan+r ↓
Shared	$256.8^{\pm 17.3}$	$878.9^{\pm 23.4}$	$283.4^{\pm 20.6}$
Shared + plan-only anchor	$303.7^{\pm 77.8}$	$361.3^{\pm 67.2}$	$501.7^{\pm 62.5}$
Routed	$227.5^{\pm 7.2}$	$1183.7^{\pm 157.8}$	$254.7^{\pm 13.6}$
Routed + plan-only anchor	$244.8^{\pm 25.8}$	$570.7^{\pm 24.0}$	$461.0^{\pm 66.2}$

F Event Coverage and Conditional-Order Diagnostics

Table S14 reuses event-level majority votes without changing outputs or event definitions. Coverage requires all events. Order | Covered first averages DAG edges within fully covered samples. Main-text Seq. instead scores every required edge over all samples, with a missing endpoint failing only that edge; the three quantities therefore do not obey a product identity. Table S6 has denominator 180 for one checkpoint. Summary comparisons calculate each metric within three checkpoints and then average; N_{cov} sums actual counts over 540 outputs. From Zero to Predicted, Coverage rises 0.839 to 0.872 and Order 0.878 to 0.904; Mismatched lowers both. On the 148 motions covered by both Predicted and Zero, Predicted's paired order score is higher by 0.022 (95% CI $[0.008, 0.036]$); on the 139 covered by both Predicted and Mismatched, the gain is 0.034 ($[0.018, 0.050]$). ACF covers more events than PSD while PSD has slightly higher conditional order. EDM and Learned Set improve both diagnostics; PH-vector further raises coverage, and the full point-set interface improves both coverage and conditional order.

G Supplementary Mechanism Analysis

G.1 Internal Trajectory-Summary Controls

Table 3 compares no summary, autocorrelation (ACF), power spectrum (PSD), the spatiotemporal Euclidean distance matrix (EDM), a learned permutation-invariant set encoder, and persistence (PH). EDM directly represents Equation A.2; PH aggregates connected and cyclic relationships across its filtration. The Learned Set baseline applies a shared 256–256 SiLU point MLP to the same sampled states, concatenates masked mean and max pools, and projects to 128 dimensions; its trainable parameter count is matched to the PH prediction/readout interface within 1.8%. Nonempty summaries share records, preprocessing, 128-D output, downstream objectives, and retraining budget. PH-vector directly regresses the fixed persistence readout, while Point-set predicts birth–death coordinates and existence masses. Across the internal summary-control outputs, Fleiss' κ for necessary-edge judgments is 0.76. Table S9 separately isolates point-set versus vector prediction at matched head capacity.

G.2 Order-Sensitive Plan Controls

The pooled readout in Equation 4 constrains global content but is exactly invariant to plan-token permutation. We therefore compare it with a parameter-matched two-layer temporal Transformer readout and perform an evaluation-only permutation that preserves every token multiset while shuffling valid positions before decoding. The temporal readout yields only a small improvement, whereas permutation substantially harms reconstruction and Seq. despite identical pooled supervision. Thus

Table S14: Event coverage and conditional-order diagnostics. Reference is an unranked oracle in the intervention block.

Source	Variant	Seq.↑	Coverage↑	Order|Cov.↑	N_{cov}
Condition intervention	Zero	0.756	0.839	0.878	151/180
	Mean	0.771	0.856	0.884	154/180
	Mismatched	0.708	0.800	0.861	144/180
	Predicted	**0.804**	**0.872**	**0.904**	157/180
	Reference (oracle)	0.821	0.883	0.912	159/180
Summary comparison	None	$0.681^{\pm 0.018}$	$0.789^{\pm 0.019}$	$0.834^{\pm 0.016}$	426/540
	ACF	$0.728^{\pm 0.018}$	$0.824^{\pm 0.018}$	$0.862^{\pm 0.017}$	445/540
	PSD	$0.716^{\pm 0.017}$	$0.811^{\pm 0.017}$	$0.868^{\pm 0.015}$	438/540
	EDM	$0.758^{\pm 0.016}$	$0.844^{\pm 0.015}$	$0.881^{\pm 0.014}$	456/540
	Learned Set	$0.766^{\pm 0.016}$	$0.850^{\pm 0.015}$	$\underline{0.883}^{\pm 0.015}$	459/540
	PH-vector	$\underline{0.783}^{\pm 0.015}$	$\underline{0.867}^{\pm 0.016}$	$0.878^{\pm 0.017}$	468/540
	Point-set	$\mathbf{0.804}^{\pm 0.014}$	$\mathbf{0.876}^{\pm 0.013}$	$\mathbf{0.905}^{\pm 0.012}$	473/540

Table S15: Order controls on HumanML3D. Permutation leaves the pooled descriptor unchanged by construction.

Setting	$e_q \downarrow$	Plan-only MPJPE↓	Seq.↑
Mean readout (default)	$0.141^{\pm 0.004}$	$409.6^{\pm 10.8}$	$0.804^{\pm 0.014}$
Temporal readout	$0.136^{\pm 0.004}$	$405.2^{\pm 10.5}$	$0.809^{\pm 0.014}$
Permute q before decoder	$0.141^{\pm 0.004}$	$623.7^{\pm 16.9}$	$0.642^{\pm 0.020}$

topology targets global trajectory organization; positional order is carried by the ordered code sequence and learned through reconstruction and autoregressive prediction rather than being identified by $\mathcal{L}_{q\text{-topo}}$ alone.

G.3 Duration Conditioning

We retrain the predictor/planner with either text alone or the supplied-length embedding described in Appendix A.4; representation and flow capacity are unchanged.

Table S16: Length conditioning on HumanML3D. e_z is descriptor error; short/long use train-set length tertiles.

Predictor	$e_z \downarrow$	Short↓	Long↓	FID↓	Seq.↑
$P_{\theta_z}(c)$	$0.119^{\pm 0.004}$	$0.116^{\pm 0.005}$	$0.131^{\pm 0.006}$	$0.029^{\pm 0.002}$	$0.792^{\pm 0.015}$
$P_{\theta_z}(c, T)$	$\mathbf{0.112}^{\pm 0.004}$	$\mathbf{0.111}^{\pm 0.005}$	$\mathbf{0.115}^{\pm 0.005}$	$\mathbf{0.027}^{\pm 0.002}$	$\mathbf{0.804}^{\pm 0.014}$

Length conditioning mainly reduces the long-motion descriptor error and is used in the complete model. In a counterfactual probe that holds text fixed and varies valid supplied length over $0.75/1.0/1.25$ of the record length, the predicted descriptor changes by mean normalized distance 0.041; the topology reference is therefore not treated as duration-invariant.

G.4 Frozen Probes of Plan and Continuous Channels

We encode reference motions to (q, r) and train equal-capacity two-layer temporal Transformers (width 128) while freezing the encoder, codebook, and decoder. The event probes reuse the human event and DAG annotations for the 180-caption composite-action subset in Section C. Each latent sequence is paired with the corresponding event or directed event-pair query; the same learned query embeddings are used for the q and r probes. For order prediction, each annotated edge $a \to b$ is a positive query and its reversal $b \to a$ is a negative query, and balanced accuracy averages the two labels. Root-path targets concatenate eight horizontal root-displacement control points sampled uniformly in normalized time after subtracting the initial position and rotating the initial heading to the canonical forward axis. On periodic records, cycle phase is obtained from the dominant nonzero-lag autocorrelation peak of root-free joint velocity, mapped to $[0, 2\pi)$, and quantized into eight equal bins; records without a valid peak are excluded from this target. We report motion-ID-grouped five-fold cross-validation, with 144 captions for fitting and 36 for evaluation in each fold. Motion-derived regression and phase targets follow the official train/validation/test split. All values average three independent probe initializations.

Table S17: Frozen functional probes. Classification uses accuracy/macro-F1 and regression test R^2; each row compares identical targets and capacity.

Target	Metric	Plan q	Realization r
Required event set	macro-F1	$\mathbf{0.781}^{\pm 0.014}$	$0.744^{\pm 0.016}$
Required order edges	Bal. Acc.	$\mathbf{0.742}^{\pm 0.018}$	$0.651^{\pm 0.020}$
Root path control points	R^2	$\mathbf{0.684}^{\pm 0.026}$	$0.592^{\pm 0.030}$
Cycle phase (8 bins)	Acc.	$0.631^{\pm 0.022}$	$\mathbf{0.784}^{\pm 0.019}$
Root-free joint velocity	R^2	$0.412^{\pm 0.029}$	$\mathbf{0.668}^{\pm 0.027}$

q better predicts required events, precedence edges, and root path, whereas r better predicts cycle phase and local velocity. The channels therefore exhibit complementary biases rather than a hard separation: plans favor sequence-level organization, and continuous variables favor concrete realization.