CHAINMO: CHAIN-SEGMENT MOTION MODELING ACROSS ARBITRARY SKELETON TOPOLOGIES

Anonymous authors
Paper under double-blind review

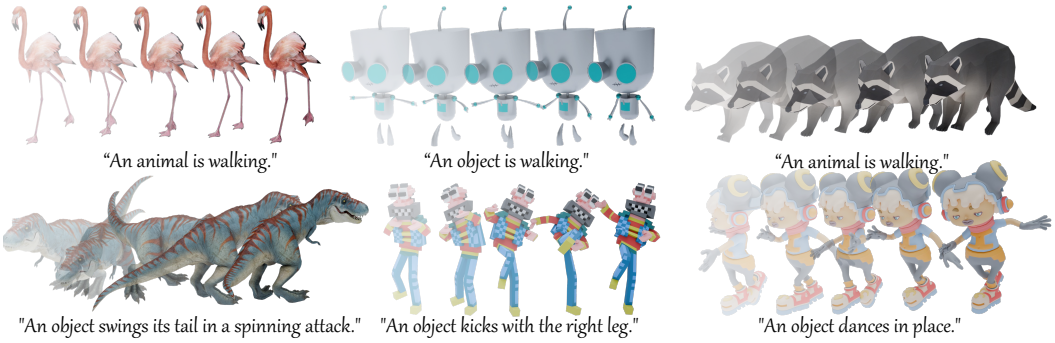


Figure 1: **ChainMo enables text-driven motion generation across arbitrary skeleton topologies.** By modeling motion with topology-adaptive chain segments and separating shared dynamics from target-rig-specific rotation realization, ChainMo generates coherent motions across characters with diverse morphologies, topologies, and joint discretizations. Representative frames are shown under the corresponding text prompts.

ABSTRACT

Generating motion for articulated characters with diverse skeleton topologies remains challenging because existing motion representations are often tightly coupled to the skeletal structures on which they are defined. Position-based representations are relatively flexible across skeletons, but remain sensitive to rig discretization, as the same articulated chain may contain different numbers of joints across rigs. Moreover, converting generated joint trajectories into animation-ready local rotations requires inverse kinematics, introducing ambiguities in under-constrained degrees of freedom such as bone twist. Directly modeling local rotations avoids this post-processing ambiguity, but further entangles motion generation with rig-specific kinematics and local rotation conventions. We address these limitations with **ChainMo**, a unified framework that separates shared motion modeling from target-rig-specific rotation realization. ChainMo represents motion using a dedicated root token and a variable set of disjoint, non-branching kinematic chain-segment tokens, forming a shared motion abstraction that is robust to variations in within-chain joint resolution. A target-rig-conditioned decoder realizes the abstract motion as coherent local joint rotations without relying on external inverse kinematics. For motion generation, we further introduce a flow-matching model with global spatiotemporal attention over the variable set of chain-segment tokens, enabling shared generative modeling across heterogeneous articulated characters. Extensive experiments on characters with diverse and previously unseen skeletons demonstrate strong generalization across kinematic structures and rig resolutions, accurate reconstruction of local joint rotations, and effective generative modeling in the learned chain-segment latent space. See our anonymous project page for more demos: https://anonymous1778.github.io/chainmo/.

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.

1 INTRODUCTION

Motion generation for articulated characters is fundamental to character animation, gaming, virtual reality, and embodied intelligence. Recent advances (Gat et al., 2025; Pu et al., 2026; Zhang et al., 2026) have extended motion modeling beyond fixed human skeletons toward heterogeneous characters with diverse and even previously unseen skeletal topologies. To accommodate structural variation, several recent approaches (Huang et al., 2025; Gong et al., 2026a; Zhang et al., 2026; Tao et al., 2026) represent motion using joint positions, which provide a relatively skeleton-agnostic description of motion. However, their joint-level formulation remains sensitive to rig discretization, as changes in joint count or chain resolution alter the motion representation even when the underlying articulated behavior is similar. Moreover, obtaining animation-ready local rotations requires inverse kinematics, where under-constrained degrees of freedom, particularly bone twist, can lead to unstable or implausible rotations (Gong et al., 2026b; Zhang et al., 2026). Directly modeling local rotations avoids this ambiguity, but instead couples motion generation more strongly with rig-specific rotation realization. This coupling becomes particularly challenging across heterogeneous skeletons, where similar motions may be expressed by different numbers and configurations of joint rotations.

As illustrated in Fig. 2, articulated rigs vary substantially in both skeletal topology and within-chain joint discretization. We therefore separate *motion modeling* from *rotation realization*. Our generative model operates in an abstract motion space that captures shared articulated dynamics, while a target-rig-conditioned decoder realizes these dynamics as animation-ready local joint rotations. We construct this abstract space using a variable set of non-branching kinematic chain segments rather than individual joints.

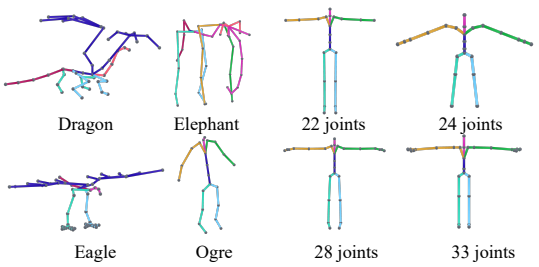


Figure 2: **Motivation for topology-adaptive motion representation.** Diverse rigs vary in both topology and joint discretization. We decompose them into non-branching kinematic chain segments *indicated by different colors*, yielding topology-adaptive motion units across heterogeneous skeletons.

Specifically, we treat the root joint separately and partition all non-root joints into disjoint, non-branching chain segments, such that every non-root joint belongs to exactly one segment. These segments are defined primarily by the branching structure of the kinematic tree and therefore remain unchanged when intermediate joints are inserted or removed within a chain. They consequently provide a topology-adaptive abstraction that reflects the branching structure of the skeleton while reducing sensitivity to within-chain rig resolution.

Based on this representation, we introduce **ChainMo**, a unified motion modeling framework for articulated characters with arbitrary skeleton topologies. ChainMo encodes motion into a shared latent representation comprising a dedicated root token and a variable set of chain-segment tokens. A target-rig-conditioned decoder then realizes the encoded motion as local joint rotations without relying on external inverse kinematics. For motion generation, we develop a flow-matching model (Lipman et al., 2023) with global spatiotemporal attention over a variable set of chain-segment tokens to capture global dependencies across segments and time. Experiments on characters with diverse and previously unseen skeletons demonstrate strong generalization across kinematic structures and rig resolutions, accurate reconstruction of local joint rotations, and effective generative modeling in the learned chain-segment latent space.

Our contributions are summarized as follows.

- We introduce a disjoint chain-segment representation that adapts to arbitrary skeleton topologies and remains robust to variations in within-segment joint resolution;
- We propose a target-rig-conditioned rotation decoder that realizes shared motion dynamics as coherent, animation-ready local joint rotations; and

- We develop a flow-matching model that operates on variable sets of chain-segment tokens and captures global dependencies across segments and time, enabling shared motion generation across heterogeneous articulated characters.

2 Related Work

Motion Modeling with Fixed Skeletons. Most existing motion generation methods assume a fixed skeletal topology with predefined joint correspondence. This setting underlies widely used human motion benchmarks such as HumanML3D (Guo et al., 2022) and KIT Motion-Language (Plappert et al., 2016), and has supported diffusion-based approaches (Zhang et al., 2024b; Tevet et al., 2023; Zhang et al., 2023c;d), continuous latent models (Petrovich et al., 2022; Tevet et al., 2022; Chen et al., 2023; Dai et al., 2024), and discrete token-based methods (Zhang et al., 2023a; Jiang et al., 2023; Zhong et al., 2023; Pinyoanuntapong et al., 2024; Guo et al., 2024). The same assumption is also common in animal motion generation, where methods such as OmniMotionGPT (Yang et al., 2024), AniMo (Wang et al., 2025), and X-MoGen (Wang et al., 2026b) model diverse species using canonical or shared skeletal representations. Despite substantial progress across both human and animal motion, these methods ultimately learn motion over predefined skeletal structures and therefore do not directly accommodate arbitrary joint counts and kinematic topologies.

Motion Modeling across Arbitrary Skeleton Topologies. Cross-skeleton motion modeling has long been studied in motion retargeting, where methods such as Neural Kinematic Networks (Villegas et al., 2018), PMnet (Lim et al., 2019), Skeleton-Aware Networks (Aberman et al., 2020), contact-aware retargeting (Villegas et al., 2021), R2ET (Zhang et al., 2023b), and HuMoT (Mourot et al., 2023) develop neural representations and architectures for transferring motion across different skeletal structures. More recent generative approaches directly target heterogeneous and previously unseen skeletons. AnyTop (Gat et al., 2025) retains joints as the basic motion elements and uses topology-aware attention to process variable skeletal graphs, while AnimaX (Huang et al., 2025), MoCapAnything (Gong et al., 2026a), and SkelMo (Tao et al., 2026) primarily model motion in Cartesian joint space. Although these formulations accommodate varying skeletal structures, joint-level representations remain sensitive to rig discretization, while position-based motion does not fully determine local joint rotations. MoCapAnything V2 (Gong et al., 2026b) addresses the latter issue with a learned pose-to-rotation module conditioned on the target rig.

Shared Motion Abstractions across Skeletons. Recent methods have explored higher-level motion representations shared across heterogeneous skeletons. MotionS (Zhang et al., 2024a) introduces a canonical motion space for cross-skeleton motion stylization, while TopoCap (Pu et al., 2026) encodes heterogeneous motions into a fixed-length latent space with target-skeleton-conditioned decoding. Concurrently, SAMoR (Zhang et al., 2026) represents motion using a fixed set of learned functional part tokens shared across skeletons. While these approaches abstract away from joint-level representations, they rely on either fixed-length global embeddings or a fixed number of learned parts. In contrast, ChainMo derives a variable set of disjoint, non-branching motion units directly from skeleton topology, preserving articulated structure while adapting to heterogeneous skeletal discretizations.

3 Proposed Method

Overview. Given a skeleton $\mathcal{S} = (\mathcal{J}, \mathcal{E})$, where $\mathcal{J}$ contains J joints and $\mathcal{E}$ defines their parent–child connectivity, ChainMo separates shared motion modeling from target-rig-specific rotation realization. Our key observation is that motion semantics, such as the movement of a limb or spine segment, can remain similar even when the underlying rigs differ in topology or joint discretization, whereas conventional joint-level representations are directly tied to the number and arrangement of joints. To reduce this dependence, as shown in Fig. 3, we first partition each skeleton into disjoint, non-branching chain segments and encode joint motion into a shared chain-segment latent representation (Secs. 3.1–3.2). This abstraction allows motion to be modeled at the level of kinematic structures rather than rig-specific joints, making the representation more adaptable to heterogeneous skeletons and different joint resolutions. A target-rig-conditioned decoder then maps the shared latent motion back to animation-ready local joint rotations according to the geometry and topology of the tar-

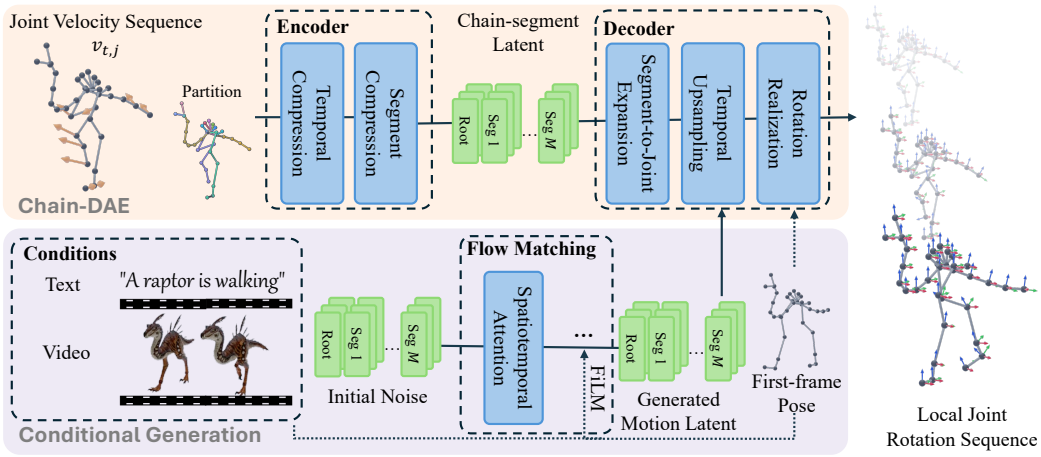


Figure 3: **Overview of ChainMo**. ChainMo represents motion on arbitrary skeletal topologies using a variable-size set of chain-segment latent tokens. **Top**: Given a joint-velocity sequence, the Chain-DAE partitions the non-root joints into disjoint, non-branching chain segments, temporally compresses each segment, and encodes the motion into one root token and a variable number of segment tokens. Its target-rig-conditioned decoder expands segment tokens back to joints and reconstructs a sequence of local joint rotations conditioned on the target skeleton and first-frame pose. **Bottom**: For conditional generation, the flow-matching model transforms noise into motion latents conditioned on the first-frame skeletal state and an optional task-specific condition, such as text or video. The generated latents are then decoded by the same Chain-DAE decoder into motion on the target skeleton.

get skeleton, thereby recovering the rig-specific details that are intentionally abstracted away in the shared representation. Finally, a conditional flow-matching model learns motion generation directly in this topology-adaptive latent space (Sec. 3.3). The training objectives and optimization strategy are described in Sec. 3.4.

3.1 Chain-Segment Motion Representation

Chain-Segment Decomposition. We treat the root joint separately and partition all non-root joints into disjoint, non-branching kinematic chain segments. Let j_{root} denote the root joint, $\mathcal{J}^{\text{nr}} = \mathcal{J} \setminus \{j_{\text{root}}\}$, and let $d^+(j)$ denote the number of child joints of joint j. We define the set of segment-start joints as

$$\mathcal{A} = \left\{ j \in \mathcal{J}^{\text{nr}} \;\middle|\; \text{par}(j) = j_{\text{root}} \text{ or } d^+(\text{par}(j)) \neq 1 \right\}. \tag{1}$$

Starting from each $j_{m,0} \in \mathcal{A}$, we follow the unique child until reaching a branching or leaf joint, producing a maximal segment

$$\mathcal{G}_m = (j_{m,0}, \ldots, j_{m,L_m}), \qquad d^+(j_{m,\ell}) = 1 \;\; (0 \leq \ell < L_m). \tag{2}$$

The resulting set $\mathcal{G} = \{\mathcal{G}_m\}_{m=1}^{M}$ forms a disjoint partition of the non-root joints,

$$\bigcup_{m=1}^{M} \mathcal{G}_m = \mathcal{J}^{\text{nr}}, \qquad \mathcal{G}_m \cap \mathcal{G}_n = \varnothing \quad (m \neq n). \tag{3}$$

Thus, the number of motion units follows the branching structure of the skeleton rather than its within-chain joint resolution.

Physical Arc-Length Coordinate. To characterize joint locations within a segment without relying on discrete joint indices, we use a normalized arc-length coordinate. For $\mathcal{G}_m = (j_{m,0}, \ldots, j_{m,L_m})$, let $j_{m,-1} = \text{par}(j_{m,0})$ denote its upstream geometric anchor. We define the bone length $d_{m,n}$ and the normalized coordinate $a_{m,l}$ of joint $j_{m,l}$ as

$$d_{m,n} = \left\| \mathbf{x}^{\text{rest}}_{j_{m,n+1}} - \mathbf{x}^{\text{rest}}_{j_{m,n}} \right\|_2, \qquad a_{m,l} = \sum_{n=-1}^{l-1} d_{m,n} \Big/ \sum_{n=-1}^{L_m-1} d_{m,n} \in (0, 1]. \tag{4}$$

Here, $\mathbf{x}^{\text{rest}}_j$ denotes the rest-pose position of joint j. Unlike a discrete joint index, $a_{m,l}$ is determined by the physical geometry of the chain and remains comparable when the same chain is discretized using different numbers of joints.

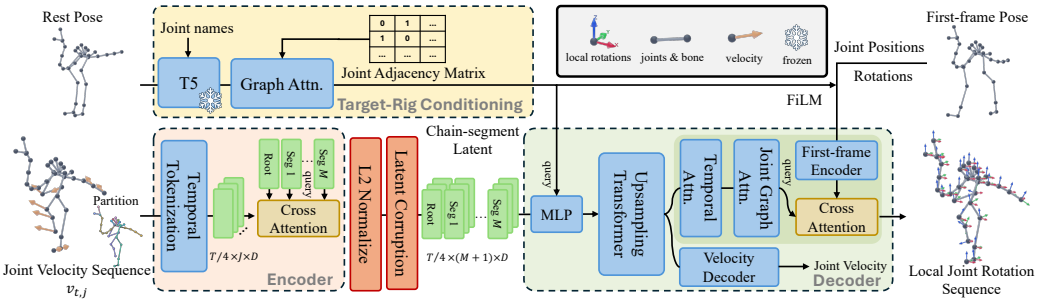


Figure 4: **Overview of Chain-DAE.** Chain-DAE compresses joint-level motion into a shared latent representation consisting of a root token and a variable set of chain-segment tokens. The encoder temporally tokenizes joint velocities and aggregates joints within each kinematic chain segment, while a target-rig conditioning branch provides target-rig structural information. The decoder expands each segment latent back to its target joints and combines the resulting motion features with the first-frame pose to realize animation-ready local joint rotations. Latent corruption and the auxiliary velocity decoder are used only during training to improve latent robustness and preserve motion dynamics.

3.2 CHAIN-DAE

Chain-DAE maps root-translation-normalized joint motion into a shared chain-segment latent space and realizes the encoded dynamics as local joint rotations conditioned on the target skeleton.

Temporal Tokenization and Segment Encoding. For non-root joints, we first remove the global root translation and represent articulated dynamics using joint velocities in the resulting translated coordinate frame. We do not canonicalize root orientation; therefore, the rotational dynamics of the root remain encoded in the non-root joint trajectories, while root translation is handled by a dedicated branch. A shared temporal tokenizer is applied independently to each joint to aggregate short-range motion patterns and reduce the temporal resolution from T to $K = T/4$, producing joint features $\mathbf{h}_{k,j}$. This temporal compression yields a compact motion representation before aggregation across joints within each chain segment. For each chain segment, we instantiate a slot from a shared learnable query. At temporal step k, the slot attends only to the joints belonging to that segment:

$$\mathbf{z}_{k,m} = \mathrm{Attn}\left(\mathbf{s}, \left\{\mathbf{h}_{k,j_{m,l}} + \phi(a_{m,l})\right\}_{l=0}^{L_m}\right), \tag{5}$$

where $\mathbf{s}$ is the shared segment query and $\phi(\cdot)$ encodes the normalized arc-length coordinate. This operation compresses a variable number of joint features into a single chain-segment token. Root translation is represented separately using the root velocity sequence, which is temporally encoded into a dedicated root latent $\mathbf{z}_k^{\mathrm{root}}$. The root token explicitly represents root translation but does not contain a dedicated root-rotation channel. Instead, the temporal evolution of root orientation is inferred during rotation realization from the articulated dynamics encoded by the non-root chain-segment tokens, while the first-frame root orientation provides the absolute orientation anchor. All root and segment tokens are projected to the latent dimension and normalized token-wise using ℓ_2 normalization.

Target-Rig-Conditioned Joint Expansion. To realize the shared latent motion on a target skeleton, we construct a joint-specific structural query $\mathbf{q}_j$ from the target rest-pose geometry, kinematic topology, physical arc coordinate, and a semantic joint-name embedding obtained from a pretrained T5 encoder. To prevent over-reliance on semantic joint names and improve generalization to unconventional rigs, we randomly drop the joint-name embedding during training. The corresponding segment latent is expanded into a joint-level feature by

$$\mathbf{h}_{k,j}^{\mathrm{dec}} = f_{\mathrm{dec}}\left(\mathbf{z}_{k,s(j)}, \mathbf{q}_j\right), \tag{6}$$

where $s(j)$ denotes the segment containing joint j. This expansion allows the same segment-level motion to be instantiated on rigs with different joint counts and within-chain discretizations. The resulting joint features are temporally upsampled from K back to T before rotation realization.

Graph-Temporal Rotation Realization. After topology-conditioned joint expansion and temporal upsampling, we obtain dense joint features $\widehat{\mathbf{H}} \in \mathbb{R}^{T \times J \times D}$, which are mapped to animation-ready

local rotations through a graph-temporal decoder. The decoder alternates temporal attention along each joint trajectory with graph attention over the skeletal edges $\mathcal{E}$, allowing it to capture both temporal motion dependencies and kinematic interactions among connected joints. Since the latent motion representation is velocity-based and does not uniquely specify the absolute skeletal configuration, we additionally condition the decoder on the first-frame skeletal state. Specifically, the first-frame joint positions and local rotations, $\{\mathbf{x}_{1,j}, \mathbf{R}_{1,j}\}_{j=1}^{J}$, are encoded into reference-pose features, which condition the graph-temporal decoder through both feature-wise modulation (FiLM) and cross-attention. A rotation head then predicts local joint rotations in the 6D representation for all frames and joints, including the root. In particular, the time-varying root orientation is inferred from the non-root articulated dynamics, with the first-frame orientation serving only as an absolute reference. In parallel, a separate root-translation head predicts the root velocity, which is integrated to recover the global trajectory.

3.3 Conditional Flow Matching over Chain-Segment Latents

Given the latent representation learned by Chain-DAE, we model its distribution using conditional flow matching. Let $\mathbf{z}_1$ denote an encoded motion latent and $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ a Gaussian sample of the same shape. For $t \sim \mathcal{U}(0, 1)$, we define

$$\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\mathbf{z}_1, \qquad \mathbf{u}_t = \mathbf{z}_1 - \mathbf{z}_0. \tag{7}$$

We learn a conditional velocity field $\mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c})$ with

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_1, t}\left[\|\mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c}) - \mathbf{u}_t\|_2^2\right], \tag{8}$$

where $\mathbf{c}$ denotes the available generation condition. At inference time, integrating the learned velocity field transforms Gaussian noise into a chain-segment motion latent, which is decoded by Chain-DAE into animation-ready motion.

Global Spatiotemporal Backbone. We adopt a DiT-style Transformer operating on the root and variable set of chain-segment tokens. Global spatiotemporal attention jointly models dependencies across segments and time without restricting interactions to kinematically neighboring chains, allowing coordinated motion patterns to emerge across the articulated system. The first-frame skeletal state and semantic joint-name embeddings are provided as structural context for all generation modes. We encode each joint name using a pretrained T5 encoder and combine the resulting semantic feature with the corresponding first-frame joint feature. The joint-level structural features are then aggregated according to the same chain decomposition as the motion representation, producing segment-aligned conditioning features that are included in $\mathbf{c}$ together with any task-specific condition.

Text Conditioning. For text-conditioned generation, we encode the input description using a pretrained T5 encoder (Raffel et al., 2020). The resulting language representation modulates the Transformer features through feature-wise affine transformations conditioned jointly on the text embedding and flow-time embedding.

Video Conditioning. To demonstrate the extensibility of the learned latent space beyond text conditioning, we additionally instantiate a video-conditioned variant. Following the joint-aligned visual conditioning strategy used in SkelMo, we extract dense DINOv3 (Siméoni et al., 2025) features, associate them with target joints using rendered rest-pose geometry, and aggregate them according to the chain decomposition. Additional qualitative results are provided in Appendix A.5.

3.4 Training Strategy

Training Chain-DAE. To encourage the latent representation to preserve motion dynamics before rotation realization, we optimize Chain-DAE in two stages. In *Stage 1*, the encoder is trained together with an auxiliary velocity decoder to reconstruct root-translation-normalized joint velocities:

$$\mathcal{L}_{\mathrm{vel}} = \ell_{\mathrm{SL1}}\left(\widehat{\mathbf{V}}, \mathbf{V}\right). \tag{9}$$

Here, ℓ_{SL1} denotes the Smooth L1 loss (Girshick, 2015), $\mathbf{V}$ is the ground-truth root-translation-normalized joint velocity, and $\widehat{\mathbf{V}}$ is its reconstruction. In *Stage 2*, we discard the auxiliary decoder, freeze the motion encoder, and train the target-rig-conditioned decoder for rotation realization. We

Table 1: **Motion reconstruction on HumanML3D.** We evaluate reconstruction fidelity using the standard HumanML3D evaluator. For R-Precision and MM-Dist, values closer to the ground-truth statistics indicate better semantic preservation. The best, second-best, and third-best results among compared methods are highlighted with different background colors.

Method	FID ↓	Semantic Statistics → Ground Truth			
		R@1	R@2	R@3	MM Dist.
Ground Truth	–	0.513 ± 0.002	0.703 ± 0.002	0.797 ± 0.003	2.969 ± 0.007
T2M-GPT (Zhang et al., 2023a)	0.070	0.489	0.676	0.775	3.130
MoMask (Guo et al., 2024)	0.019 ± 0.000	0.508 ± 0.003	0.701 ± 0.002	0.795 ± 0.002	2.999 ± 0.006
MoGenTS (Yuan et al., 2024)	0.005 ± 0.000	0.504 ± 0.002	0.702 ± 0.002	0.797 ± 0.002	2.978 ± 0.006
ChainMo	0.010 ± 0.000	0.514 ± 0.002	0.703 ± 0.002	0.797 ± 0.003	2.969 ± 0.008

supervise 6D local rotations, forward-kinematic joint positions, root-translation-normalized joint velocities, and the integrated root trajectory:

$$\mathcal{L}_{\text{DAE}} = \ell_{\text{SL1}}(\widehat{\mathbf{R}}^{\text{6D}}, \mathbf{R}^{\text{6D}}) + \ell_{\text{SL1}}(\widehat{\mathbf{P}}, \mathbf{P}) + \ell_{\text{SL1}}(\widehat{\mathbf{V}}, \mathbf{V}) + \ell_{\text{SL1}}(\widehat{\mathbf{T}}^{\text{root}}, \mathbf{T}^{\text{root}}). \tag{10}$$

Here, $\mathbf{R}^{\text{6D}}$, $\mathbf{P}$, $\mathbf{V}$, and $\mathbf{T}^{\text{root}}$ denote the ground-truth local rotations, joint positions, root-translation-normalized joint velocities, and root trajectory, respectively. The predicted positions $\widehat{\mathbf{P}}$ are obtained by forward kinematics from $\widehat{\mathbf{R}}^{\text{6D}}$, and $\widehat{\mathbf{V}}$ is computed from the resulting FK trajectories using the same temporal differencing and root-translation normalization as the ground truth.

To improve decoder robustness to imperfect latents produced by the generative model, we perturb each normalized latent during training:

$$\widetilde{\mathbf{z}}_\lambda = (1-\lambda)\bar{\mathbf{z}} + \lambda\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \lambda \in [0, \lambda_{\max}]. \tag{11}$$

Here, $\bar{\mathbf{z}}$ is the normalized encoder latent, $\boldsymbol{\epsilon}$ is standard Gaussian noise, λ controls the perturbation strength, and $\lambda_{\max}$ specifies its maximum value.

Training Conditional Flow Matching. Within each evaluation setting, we first train the flow backbone on all available training motions without text or video conditioning, while retaining the target skeletal structure and initial state as structural context. This stage learns a shared motion prior for the corresponding training setting. We then initialize separate text- and video-conditioned models from the pretrained backbone. The backbone parameters are frozen, while modality-specific conditioning branches are optimized to inject text or video control into the corresponding Transformer layers. Both stages use the flow-matching objective defined in Sec. 3.3.

4 Experiments

4.1 Experiment Settings

Training Details. Chain-DAE is trained on 64-frame clips with $4\times$ temporal downsampling and 16-dimensional latents. The flow model operates on 196-frame sequences using a 12-layer Transformer with hidden dimension 512. Full architecture and optimization details are provided in Appendix A.1.

Benchmarks. We evaluate text-conditioned motion generation on HumanML3D (Guo et al., 2022) using its official train/validation/test splits and pretrained evaluator. To assess generation across heterogeneous skeletal structures, we additionally construct a diverse evaluation subset from SkelMo (Tao et al., 2026), for which we report qualitative comparisons and a user study.

Baselines. On HumanML3D, we compare with representative fixed-skeleton methods including MLD (Chen et al., 2023), T2M-GPT (Zhang et al., 2023a), MoMask (Guo et al., 2024), and MoGenTS (Yuan et al., 2024); the latter three are also used for motion reconstruction comparison. For heterogeneous text-driven animation, we compare with AnimateAnyMesh (Wu et al., 2025), BiMotion (Wang et al., 2026a), and AnyTop-T, our text-conditioned adaptation of AnyTop (Gat et al., 2025) that retains its original joint-level topology-aware architecture.

Metrics. On HumanML3D, we report FID, R-Precision, Matching Distance (MM-Dist), and Diversity following the standard evaluation protocol; reconstruction is evaluated using reconstruction FID,

R-Precision, and MM-Dist. For heterogeneous skeletons, we measure local rotation fidelity using mean geodesic rotation error and spatial accuracy using MPJPE after forward kinematics. Since no established cross-topology evaluator exists for text-driven generation on heterogeneous articulated characters, we additionally use qualitative comparisons and a user study.

4.2 Evaluation of Motion Representation

Motion Reconstruction. Tab. 1 evaluates the fidelity of the learned motion representation on HumanML3D. Despite compressing multiple joints into chain-segment tokens, ChainMo achieves a reconstruction FID of 0.010 while closely matching the ground-truth R-Precision and MM-Dist statistics. Although MoGenTS attains a lower reconstruction FID with a joint-level representation, ChainMo preserves comparable motion fidelity while supporting topology-adaptive modeling. On heterogeneous skeletons, reconstruction errors increase only moderately from seen to unseen rigs, while the local rotation error remains nearly unchanged (Tab. 2).

Table 2: **Generalization to unseen skeletons.** Reconstruction performance on seen and unseen skeletons.

Split	MPJPE ↓	Root-aligned MPJPE ↓	Rot. Err. ↓
Seen	0.0287	0.0208	6.99°
Unseen	0.0347	0.0267	6.96°

Effect of Chain-Segment Representation. We construct a joint-level variant by removing the segment aggregation in Eq. (5), while keeping the same backbone, decoder, and heterogeneous training setting. Under 3× within-chain densification, joint tokens achieve lower original-rig reconstruction error but are substantially more sensitive to joint resolution (Tab. 3). Chain segments reduce the decoded-motion change by 61.6% (0.666% to 0.256%) and the rig-macro absolute reconstruction-error change by 73.0% (0.237 to 0.064 pp). For latent consistency, joint tokens are compared at retained original joints, while chain-segment tokens are matched one-to-one under the unchanged chain decomposition; the resulting cosine similarity is 0.9180 and 0.9908, respectively.

Table 3: **Joint-level vs. chain-segment representations under rig densification.** Both variants use the same backbone, decoder, and heterogeneous training setting. Orig. Err. measures normalized root-aligned reconstruction error, while Decoded Motion Δ directly measures the discrepancy between motions decoded from the original and 3×-densified rigs. Detailed metric definitions are provided in Appendix A.2.

Representation	Orig. Err. (%) ↓	Decoded Motion Δ (%) ↓	Mean |Δ Err.| (pp) ↓	Latent Cos.↑
Joint tokens	**2.049**	0.666	0.237	0.9180
Chain segments	2.416	**0.256**	**0.064**	**0.9908**

4.3 Evaluation on Text-Conditioned Motion Generation

Text-Conditioned Generation on HumanML3D. As shown in Tab. 4, ChainMo achieves competitive motion quality while obtaining the strongest text–motion alignment among the compared methods. Its FID of 0.034 is comparable to MoGenTS (0.033), while achieving the best R-Precision and MM-Dist. These results show that the topology-adaptive chain-segment representation retains strong generative fidelity even on the fixed-skeleton HumanML3D benchmark.

Generation across Heterogeneous Skeletons. We further evaluate text-driven generation on a diverse subset of SkelMo (Tao et al., 2026), comparing against AnimateAnyMesh (Wu et al., 2025), BiMotion (Wang et al., 2026a), and AnyTop-T (Gat et al., 2025). As shown in Fig. 5, ChainMo produces coherent motions across diverse body structures and previously unseen skeletons. A pairwise user study further supports the perceptual benefits of ChainMo, with overall preference rates of 91.86% for semantic alignment and 91.71% for motion naturalness; detailed results and protocols are provided in Appendix A.4.

4.4 Ablation Studies

Ablation on Latent Dimension. Tab. 5 studies the effect of the chain-segment latent dimension. A small latent space ($d = 8$) creates a strong information bottleneck and degrades reconstruction quality, while increasing the dimension to 16 substantially improves both reconstruction and generation. Under the same normalization and flow-training configuration, $d = 16$ provides the best empirical balance between reconstruction fidelity and downstream generation quality.

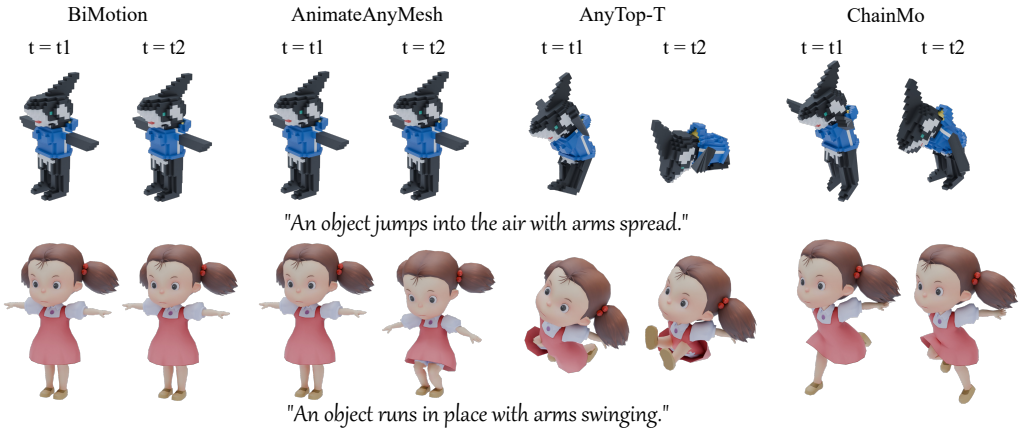


Figure 5: **Qualitative comparison of text-conditioned motion generation across heterogeneous skeletons.** We compare ChainMo with BiMotion, AnimateAnyMesh, and AnyTop-T, showing two representative frames for each generated motion. ChainMo produces semantically consistent and coherent motions across diverse skeletal structures. Additional qualitative results are provided in Appendix A.5.

Table 4: **Text-Conditioned generation on HumanML3D.** We report the mean and standard deviation over repeated evaluations. ↑/↓ indicate higher/lower is better, while → indicates closer to ground truth is better. Colors denote the top-three results.

Method	Motion Quality	Text–Motion Alignment				Diversity
	FID ↓	MM Dist. ↓	R@1 ↑	R@2 ↑	R@3 ↑	Diversity →
Ground Truth	0.002 ± 0.000	2.969 ± 0.007	0.513 ± 0.002	0.703 ± 0.002	0.797 ± 0.003	9.503 ± 0.065
MLD (Chen et al., 2023)	0.473 ± 0.013	3.196 ± 0.010	0.481 ± 0.003	0.673 ± 0.003	0.772 ± 0.002	9.724 ± 0.082
T2M-GPT (Zhang et al., 2023a)	0.116 ± 0.004	3.118 ± 0.011	0.491 ± 0.003	0.680 ± 0.003	0.775 ± 0.002	9.761 ± 0.081
MoMask (Guo et al., 2024)	0.045 ± 0.002	2.958 ± 0.008	0.521 ± 0.002	0.713 ± 0.002	0.807 ± 0.002	–
MoGenTS (Yuan et al., 2024)	0.033 ± 0.001	2.867 ± 0.006	0.529 ± 0.003	0.719 ± 0.002	0.812 ± 0.002	9.570 ± 0.077
ChainMo	0.034 ± 0.005	2.671 ± 0.020	0.565 ± 0.004	0.759 ± 0.005	0.848 ± 0.004	9.674 ± 0.163

Ablation on Rotation Decoder. As shown in Tab. 5, both first-frame rotation conditioning and graph-based joint interaction contribute to accurate rotation recovery. The first-frame rotations provide an orientation anchor for recovering absolute joint rotations, while graph attention restores cross-segment dependencies that are not explicitly modeled by the disjoint chain representation. Together, they enable locally represented motion units to be realized as globally coherent articulated motion.

Ablation on Arc Encoding. Removing the physical arc encoding degrades all evaluated metrics (Tab. 5). Since joints within a segment share the same latent representation, the arc coordinate provides joint-specific positional cues for segment-to-joint realization. Its removal makes this mapping ambiguous, confirming the importance of geometry-aware positional encoding.

Table 5: **Ablation studies of ChainMo.** Effects of latent dimension, rotation realization, and physical arc encoding. Lower is better for all metrics.

Ablation	Variant	Recon. FID ↓	Rot. Err. ↓	MPJPE ↓	Gen. FID ↓
Latent Dim.	$d = 8$	0.036	7.961°	0.063	0.070
	$d = 64$	0.009	6.646°	0.035	0.528
Rotation	w/o First-frame Rot.	0.015	5.963°	0.042	0.068
	w/o Graph Attn.	0.698	11.535°	0.095	0.795
Arc Encoding	w/o Physical Arc	0.013	7.002°	0.039	0.049
Full Model	$d = 16$ + Full Decoder	**0.010**	5.770°	**0.027**	**0.034**

5 Conclusion

We introduced ChainMo, a unified framework for motion modeling across heterogeneous skeleton topologies. By representing motion with topology-adaptive chain segments and separating shared dynamics from target-rig-specific rotation realization, ChainMo supports animation-ready motion generation across diverse rigs and joint discretizations. Experiments on HumanML3D and heterogeneous articulated characters show strong reconstruction, generation, and cross-topology generalization. We hope this representation provides a useful foundation for general motion generation and control across arbitrary articulated characters.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and improving the readability of the manuscript. We did not use generative AI tools to generate research results, conduct experiments, or make scientific claims. All AI-assisted edits were reviewed and verified by the authors. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the model architecture and training objectives in Secs. 3.2–3.4, implementation and optimization details in Appendix A.1, and dataset preparation and evaluation protocols in Appendix A.2. Additional ablations and qualitative analyses are included in the appendix. We plan to release the source code, trained models, and evaluation scripts publicly upon acceptance.

REFERENCES

Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM ToG*, 39(4):62–1, 2020. 3

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 *CVPR*, pp. 18000–18010. IEEE, 2023. 3, 7, 9

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 *ECCV*, pp. 390–408. Springer, 2024. 3

Inbar Gat, Sigal Raab, Guy Tevet, Yuval Reshef, Amit Haim Bermano, and Daniel Cohen-Or. Anytop: Character animation diffusion with any topology. In *SIGGRAPH*, pp. 1–10, 2025. 2, 3, 7, 8, 16

Ross Girshick. Fast r-cnn. In *ICCV*, pp. 1440–1448, 2015. 6

Kehong Gong, Zhengyu Wen, Weixia He, Mingxi Xu, Qi Wang, Ning Zhang, Zhengyu Li, Dongze Lian, Wei Zhao, Xiaoyu He, et al. Mocapanything: Unified 3d motion capture for arbitrary skeletons from monocular videos. In *CVPR*, pp. 7089–7099, 2026a. 2, 3

Kehong Gong, Zhengyu Wen, Dao Thien Phong, Mingxi Xu, Weixia He, Qi Wang, Ning Zhang, Zhengyu Li, Guanli Hou, Dongze Lian, et al. Mocapanything v2: End-to-end motion capture for arbitrary skeletons. *arXiv preprint arXiv:2604.28130*, 2026b. 2, 3

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 *CVPR*, pp. 5152–5161, June 2022. 3, 7, 14

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, pp. 1900–1910. IEEE, 2024. 3, 7, 9

Zehuan Huang, Haoran Feng, Yang-Tian Sun, Yuan-Chen Guo, Yan-Pei Cao, and Lu Sheng. Animax: Animating the inanimate in 3d with joint video-pose diffusion models. In *SIGGRAPH Asia*, pp. 1–13, 2025. 2, 3

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *NeurIPS*, volume 36, pp. 20067–20079, 2023. 3

Jongin Lim, Hyung Jin Chang, and Jin Young Choi. Pmnet: Learning of disentangled pose and movement for unsupervised motion retargeting. In *BMVC*. British Machine Vision Association, BMVA, 2019. 3

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t. 2

Lucas Mourot, Ludovic Hoyet, François Le Clerc, and Pierre Hellier. Humot: human motion representation using topology-agnostic transformers for character animation retargeting. *arXiv preprint arXiv:2305.18897*, 2023. 3

Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, pp. 480–497. Springer, 2022. 3

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *CVPR*, pp. 1546–1555. IEEE, 2024. 3

Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 3

Cheng-Feng Pu, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. TopoCap: Learning topology-agnostic motion priors for monocular video-to-animation. In *SIGGRAPH*, pp. 184:1–184:12, 2026. doi: 10.1145/3799902.3811159. 2, 3

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 6

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 6

Ye Tao, Yuxin Yao, Kendong Liu, Dapeng Wu, and Junhui Hou. Skelmo: Universal skeletal motion generation for 3d rigged shapes. In *ECCV*, 2026. 2, 3, 7, 8, 14

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, pp. 358–374. Springer, 2022. 3

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human Motion Diffusion Model. In *ICLR*, 2023. URL https://mlanthology.org/iclr/2023/tevet2023iclr-human/. 3

Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *CVPR*, pp. 8639–8648. IEEE, 2018. 3

Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *ICCV*, pp. 9700–9709. IEEE, 2021. 3

Miaowei Wang, Qingxuan Yan, Zhi Cao, Yayuan Li, Oisin Mac Aodha, Jason J Corso, and Amir Vaxman. Bimotion: B-spline motion for text-guided dynamic 3d character generation. *arXiv preprint arXiv:2602.18873*, 2026a. 7, 8, 16

Xuan Wang, Kai Ruan, Xing Zhang, and Gaoang Wang. Animo: Species-aware model for text-driven animal motion generation. In *CVPR*, pp. 1929–1939. IEEE, 2025. 3

Xuan Wang, Kai Ruan, Liyang Qian, Guo Zhi Zhi, Chang Su, and Gaoang Wang. X-mogen: Unified motion generation across humans and animals. In *AAAI*, volume 40, pp. 10234–10242, 2026b. 3

Zijie Wu, Chaohui Yu, Fan Wang, and Xiang Bai. Animateanymesh: A feed-forward 4d foundation model for text-driven universal mesh animation. *arXiv preprint arXiv:2506.09982*, 2025. 7, 8, 16

Zhangsihao Yang, Mingyuan Zhou, Mengyi Shan, Bingbing Wen, Ziwei Xuan, Mitch Hill, Junjie Bai, Guo-Jun Qi, and Yalin Wang. Omnimotiongpt: Animal motion generation with limited data. In *CVPR*, pp. 1249–1259. IEEE, 2024. 3

Weihao Yuan, Yisheng He, Weichao Shen, Yuan Dong, Xiaodong Gu, Zilong Dong, Liefeng Bo, and Qixing Huang. Mogents: Motion generation based on spatial-temporal joint modeling. *NeurIPS*, 37:130739–130763, 2024. 7, 9

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 *CVPR*, pp. 14730–14740, 2023a. 3, 7, 9

Jiaxu Zhang, Junwu Weng, Di Kang, Fang Zhao, Shaoli Huang, Xuefei Zhe, Linchao Bao, Ying Shan, Jue Wang, and Zhigang Tu. Skinned motion retargeting with residual perception of motion semantics & geometry. In *CVPR*, pp. 13864–13872. IEEE, 2023b. 3

Jiaxu Zhang, Xin Chen, Gang Yu, and Zhigang Tu. Generative motion stylization of cross-structure characters within canonical motion space. In *ACM MM*, pp. 7018–7026, 2024a. 3

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, pp. 364–373. IEEE, 2023c. 3

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. FineMoGen: Fine-grained spatio-temporal motion generation and editing. In *NeurIPS*, volume 36, pp. 13981–13992, 2023d. 3

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*, 46(6):4115–4128, 2024b. 3

Yuhao Zhang, Gerard Pons-Moll, and Tolga Birdal. SAMoR: Motion modelling for articulated objects of any skeleton and topology. *arXiv preprint arXiv:2607.02148*, 2026. URL https://arxiv.org/abs/2607.02148. 2, 3

Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *ICCV*, pp. 509–519. IEEE, 2023. 3

A Additional Details and Results

This appendix provides additional implementation details, evaluation protocols, and supplementary results for ChainMo. The contents are organized as follows:

- Section A.1 provides additional architecture and implementation details.
- Section A.2 describes datasets, evaluation protocols, and baseline adaptations.
- Section A.3 reports computational efficiency, including inference time and GPU memory usage.
- Section A.4 presents the user-study protocol and results.
- Section A.5 provides additional analyses and qualitative results.
- Section A.6 discusses limitations and future directions.

A.1 Architecture and Implementation Details

Chain-DAE. During Chain-DAE training, we apply joint-name dropout with probability $p_j = 0.1$, and we process motion clips of $T_{\mathrm{DAE}} = 64$ frames. The temporal tokenizer reduces the temporal resolution by a factor of 4, resulting in $K_{\mathrm{DAE}} = 16$ latent steps. For each temporal step, joint features within the same non-branching kinematic chain are aggregated into a single chain-segment token using a shared learnable query and the normalized arc-length encoding introduced in the main paper. The latent dimension of each root or chain-segment token is set to 16, and all latent tokens are normalized using token-wise ℓ_2 normalization.

The decoder first expands each chain-segment latent to the joints belonging to the corresponding target-rig segment. The expanded features are temporally upsampled to the original frame rate and processed by a graph-temporal rotation decoder, which alternates temporal attention along individual joint trajectories with graph attention over the target skeletal connectivity. The decoder is additionally conditioned on the first-frame joint positions and local rotations through feature-wise modulation and cross-attention. A separate root-translation branch predicts the global root trajectory, whereas root orientation is produced by the graph-temporal rotation decoder from the non-root chain-segment dynamics and the first-frame orientation anchor.

Flow-Matching Backbone. All flow-matching stages are trained on $T_{\mathrm{FM}} = 196$-frame motion sequences. After temporal encoding with the same factor of 4, each sequence contains $K_{\mathrm{FM}} = 49$ latent steps. Our generative backbone is a 12-layer Transformer with hidden dimension 512. The network operates directly on the root and variable set of chain-segment latent tokens and employs global spatiotemporal attention to capture dependencies across segments and time. To provide semantic correspondence across heterogeneous skeletons, we additionally encode the target joint names using the frozen T5 encoder. The joint-name features are combined with the first-frame structural features and aggregated according to the chain decomposition, yielding segment-aligned structural conditioning for the flow model. The motion-prior backbone is first trained without text or video conditioning. Separate text- and video-conditioned models are then initialized from this backbone. During conditional training, the pretrained backbone is kept frozen and modality-specific conditioning modules are optimized.

For text conditioning, we use a frozen pretrained T5 encoder. For video conditioning, we use a frozen pretrained DINOv3 encoder to extract dense visual features. The visual features are associated with the target joints using rendered rest-pose geometry and subsequently aggregated according to the same chain-segment decomposition used by the motion representation.

Optimization. All models are implemented in PyTorch and trained on a single NVIDIA RTX 5880 Ada Generation GPU. We use AdamW with a learning rate of 1×10^{-4} and batch size 64. The two Chain-DAE stages are trained for 162.5k and 140k iterations, respectively, with $\lambda_{\max} = 0.05$ for latent corruption. The flow-matching backbone is trained for approximately 110k iterations. The pretrained T5 and DINOv3 encoders remain frozen. During conditional training, we apply condition dropout with probability 0.1. At inference, classifier-free guidance is used with scale 9.

Inference and Classifier-Free Guidance. As shown in Tab. 6, increasing CFG substantially improves both generation fidelity and text–motion alignment at low guidance strengths, while the performance largely saturates around CFG = 7–9. Further increasing the guidance strength to 12 pro-

Table 6: **Effect of classifier-free guidance on HumanML3D.** We vary the CFG scale while keeping the trained model fixed. Performance improves rapidly at low guidance strengths and largely saturates around CFG = 7–9. We use CFG = 9 in all main experiments as a favorable trade-off between generation fidelity, text–motion alignment, and diversity.

CFG	FID ↓	R@1 ↑	R@3 ↑	MM Dist. ↓	Diversity →
0	1.266	0.181	0.368	6.049	9.165
1	0.148	0.467	0.760	3.052	9.703
2.5	0.082	0.518	0.812	2.770	9.863
5	0.047	0.533	0.826	2.700	9.783
7	**0.035**	0.540	0.828	2.684	9.708
9	**0.034**	**0.565**	**0.848**	**2.671**	**9.674**
12	0.056	0.536	0.829	2.697	9.527

Table 7: **Statistics of the heterogeneous motion dataset.** We report the number of motion sequences, unique rig identities, and distinct kinematic-tree topologies for each split.

Split	Motions	Rig IDs	Tree Topologies
Train	20,738	11,933	2,957
Val Seen	74	74	63
Val Unseen	220	132	51
Test Seen	74	74	61
Test Unseen	215	96	65

vides little additional alignment benefit while degrading FID. We therefore adopt CFG = 9, which provides a favorable overall balance between motion fidelity, text alignment, and diversity.

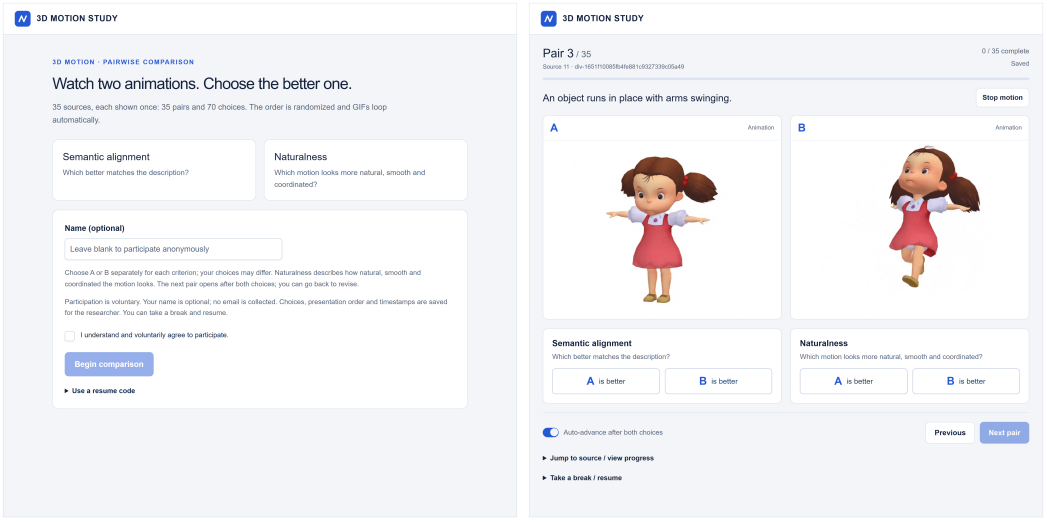
A.2 Datasets and Evaluation Protocols

HumanML3D. We train a dedicated ChainMo model on HumanML3D (Guo et al., 2022) and follow its official train, validation, and test splits and pretrained evaluator. Motion sequences are converted to our velocity-based representation. For Chain-DAE training, we sample 64-frame clips, whereas all flow-matching stages use 196-frame sequences. Generated and reconstructed motions are converted back to the representation required by the HumanML3D evaluator.

Heterogeneous-Skeleton Splits. Separately, we train the heterogeneous-skeleton ChainMo model on motion data from SkelMo (Tao et al., 2026). Tab. 7 summarizes the training, validation, and test splits. The training set contains 20,738 motion sequences spanning 11,933 unique rigs and 2,957 distinct kinematic-tree topologies, providing substantial structural variation for learning a shared motion representation. For validation and testing, we consider two complementary settings. The seen split contains unseen motion sequences performed on rigs that also appear in the training set, and therefore evaluates motion generalization on known skeletal structures. For the unseen split, we group rigs by rooted-tree isomorphism of their kinematic graphs, disregarding joint names, rest-pose geometry, and character identity. We then hold out entire tree-shape groups from training, such that no rig with the same parent–child tree structure appears in the training set. This setting evaluates structural generalization to previously unseen skeletal configurations.

Heterogeneous Text-Conditioned Generation Evaluation. For text-conditioned evaluation, we further select a diverse subset from the SkelMo test data. The subset covers characters with different body structures, joint counts, branching patterns, and within-chain skeletal resolutions. We additionally include skeletons not observed during training to evaluate cross-topology generalization.

Rig-Densification Metrics. We evaluate sensitivity to within-chain joint discretization using the original rig and its 3×-densified counterpart. The densified rig is constructed by uniformly subdividing each bone within a non-branching chain while preserving the original branching structure and rest geometry. Original Error is the root-aligned MPJPE between the decoded motion and ground truth, normalized by the diagonal length of the original rest-pose bounding box and multiplied by



(a) Study instructions (b) Pairwise comparison

Figure 6: **User-study interface.** (a) Instructions, evaluation criteria, and consent. (b) An example trial with a shared motion description and two animations labeled A and B. Participants judge semantic alignment and motion naturalness separately. Method names are hidden; the screenshot shows one frame from each looping animation.

100. Decoded-Motion Change directly compares the motions decoded from the original and densified encodings using the same root alignment and normalization; importantly, it is not the difference between their reconstruction errors. Mean Absolute Error Change is computed by first taking the absolute difference between dense- and original-rig reconstruction errors for each evaluated motion, averaging within each rig, and then macro-averaging across rigs, thereby avoiding cancellation across positive and negative changes. For Latent Cosine, each model is compared only against its own original/dense encodings: joint-token latents are aligned at retained original joints, whereas chain-segment latents are matched one-to-one under the unchanged chain decomposition. Because the representations may differ in latent width and temporal tokenization, decoded-motion change serves as the common output-space comparison.

AnyTop-T Adaptation. We construct AnyTop-T by augmenting the original AnyTop architecture with text and initial-state conditioning, while retaining its joint-level, topology-aware motion modeling. A frozen T5-base encoder extracts caption features, which are injected through both a pooled global text embedding added to the motion features and cross-attention in each Transformer block. Per-joint semantic name embeddings are encoded by the same T5 encoder and added to the corresponding joint features, while skeleton topology and rest-pose geometry are provided through the original graph-aware and rest-pose conditioning mechanisms. We additionally provide the same target first-frame skeletal state used by ChainMo: the per-joint initial position, 6D rotation, and boundary velocity are projected and added to the corresponding joint features. At inference, the boundary velocity is set to zero, and no hard first-frame inpainting is applied.

A.3 Computational Efficiency

For the text-conditioned ChainMo model, we report the computational cost of ChainMo in terms of parameter count, inference time, and peak GPU memory usage. ChainMo contains 86.70M parameters, excluding the frozen T5 text encoder. All measurements are conducted on a single NVIDIA RTX 5880 Ada Generation GPU with a batch size of 1. For a motion sequence of 196 frames on a 34-joint skeleton, ChainMo requires 2,793 ms for generation and 0.93 GB of peak GPU memory. We use 50 Heun steps with classifier-free guidance. Runtime is averaged over 20 runs after 5 warm-up runs and includes text encoding and motion decoding. Memory measurements include the frozen text encoder.

Table 8: **Pairwise user-study results.** Results are collected from 20 completed questionnaires. Each respondent evaluated 35 cases, yielding 700 paired comparisons and 1,400 binary judgments. We report the percentage of judgments favoring ChainMo for semantic alignment and motion naturalness.

	ChainMo Preference Rate (%) ↑	
Baseline	Semantic Alignment	Motion Naturalness
AnyTop-T (Gat et al., 2025)	85.78	89.22
BiMotion (Wang et al., 2026a)	97.87	97.02
AnimateAnyMesh (Wu et al., 2025)	91.85	88.84
Overall	**91.86**	**91.71**

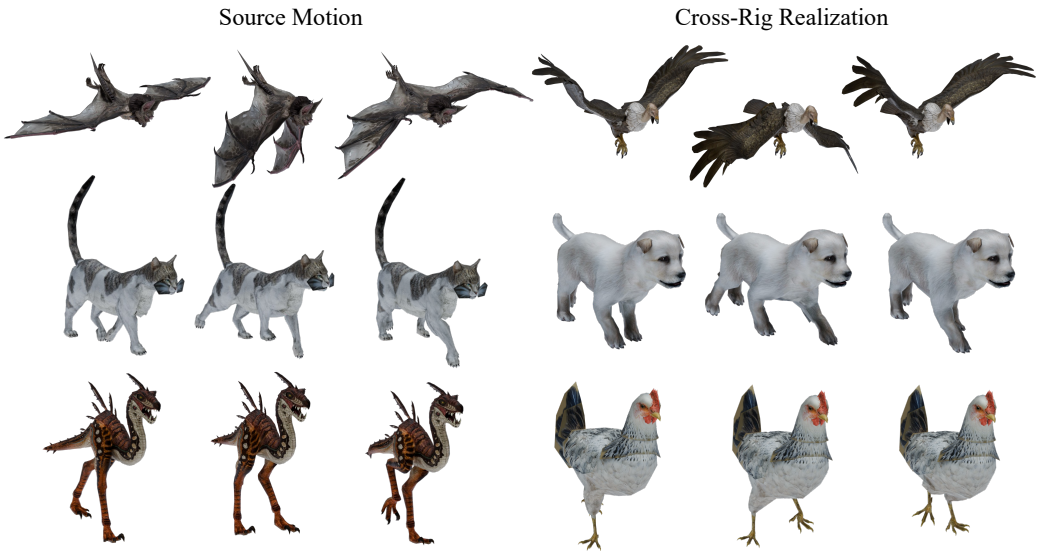


Figure 7: **Cross-skeleton latent realization.** We encode a motion from a source character and manually establish correspondence between semantically matching chain segments of the source and target skeletons. The latent sequence of each source segment is then transferred to its matched target segment and decoded using the target-rig structural conditioning and initial pose. Representative frames from the source and transferred motions are shown. Despite substantial differences in morphology, skeletal topology, and joint discretization, the transferred segment latents preserve the characteristic dynamics of the source motion on the target rig.

A.4 User Study Protocol

We compare ChainMo with three baselines on 35 cases using three questionnaire forms. Each form includes one comparison per case, with comparisons distributed nearly evenly across the baselines. Together, the forms cover all 105 case–baseline pairs exactly once. Participant assignments are balanced across forms.

After providing voluntary consent, participants evaluate 35 pairs, making separate forced-choice judgments on semantic alignment and motion naturalness. Method names are hidden, and both trial order and left/right placement are randomized for each participant. Participants may pause and revise their choices before submission. Fig. 6 shows the interface.

For each baseline and criterion, we report ChainMo's win rate as the percentage of judgments favoring ChainMo. Only fully submitted responses are included, with each participant's final choice counted once per comparison and criterion.

A.5 Additional Analyses and Qualitative Results

We provide additional qualitative results to illustrate the generalization of ChainMo across heterogeneous articulated characters.

Cross-Skeleton Latent Realization. To examine whether the learned chain-segment latents capture motion dynamics that can be realized across different rigs, we perform a segment-wise cross-skeleton transfer experiment. Given a source motion, we first encode it into the proposed chain-segment latent space. We then manually establish correspondence between semantically matching source and target chain segments and transfer each source segment latent sequence to its matched target segment. The transferred latents are decoded using the target-rig structural conditioning and initial pose, without modifying their encoded motion content. As shown in Fig. 7, the resulting motions preserve the characteristic dynamics of the source sequences across substantially different character morphologies and skeletal structures. This experiment provides qualitative evidence that individual chain-segment latents capture motion information that can be realized on different target rigs when segment correspondence is provided.

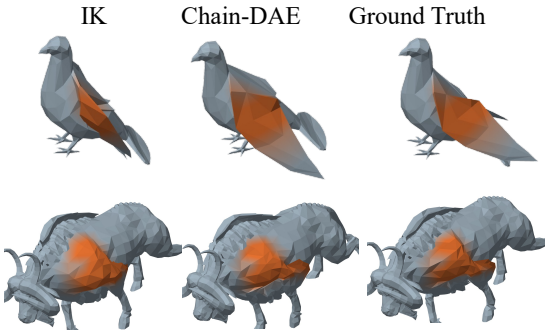


Figure 8: **Position-to-rotation ambiguity.** Even when IK is provided with the corresponding joint positions, it can produce inaccurate local orientations and bone-axis twists, whereas Chain-DAE reconstructs rotations that more closely match the ground truth.

Position-to-Rotation Ambiguity. We identify challenging intervals using the IK results only. An interval is selected when the world-space $\mathrm{SO}(3)$ orientation error of a physical internal joint exceeds 60° for at least five consecutive predicted frames, corresponding to 0.25 s at 20 fps. The provided first frame is excluded from interval selection. Chain-DAE predictions are never used to select intervals. For every selected failure interval, IK and Chain-DAE are evaluated on the same joint and the same frames against the same ground-truth orientations. On seen skeletons, Chain-DAE reduces the orientation error by 44.9% on average and improves 96.7% of the identified intervals. Importantly, the improvement also generalizes to unseen skeletons, where the error is reduced by 27.8% and 84.8% of the intervals are improved. Overall, these results show that direct rotation realization can substantially alleviate position-to-rotation ambiguities across both seen and previously unseen skeletal structures. As illustrated in Fig. 8, IK can produce noticeably different local orientations and bone-axis twists even when the articulated joint configuration is well preserved, whereas Chain-DAE recovers rotations that more closely match the reference.

Text-Conditioned Generation. Fig. 9 presents additional text-conditioned motions for characters with diverse skeletal structures. The examples include human-like characters, quadrupeds, winged characters, and articulated shapes with substantially different joint counts and branching patterns.

Video-Conditioned Generation. We additionally evaluate the extensibility of the learned chain-segment latent space under video conditioning. As shown in Fig. 10, ChainMo transfers motion cues from the input video to heterogeneous target skeletons and produces animations that closely follow the corresponding ground-truth motion. Complete motion sequences are provided on the anonymous project page.

A.6 Limitations and Future Work

Although ChainMo is designed to support heterogeneous and previously unseen skeletons, several limitations remain.

Limited Motion Diversity and Supervision. Although the proposed chain-segment representation reduces the dependence of motion modeling on a particular skeletal discretization, the generative model remains limited by the diversity and annotation quality of currently available motion data.

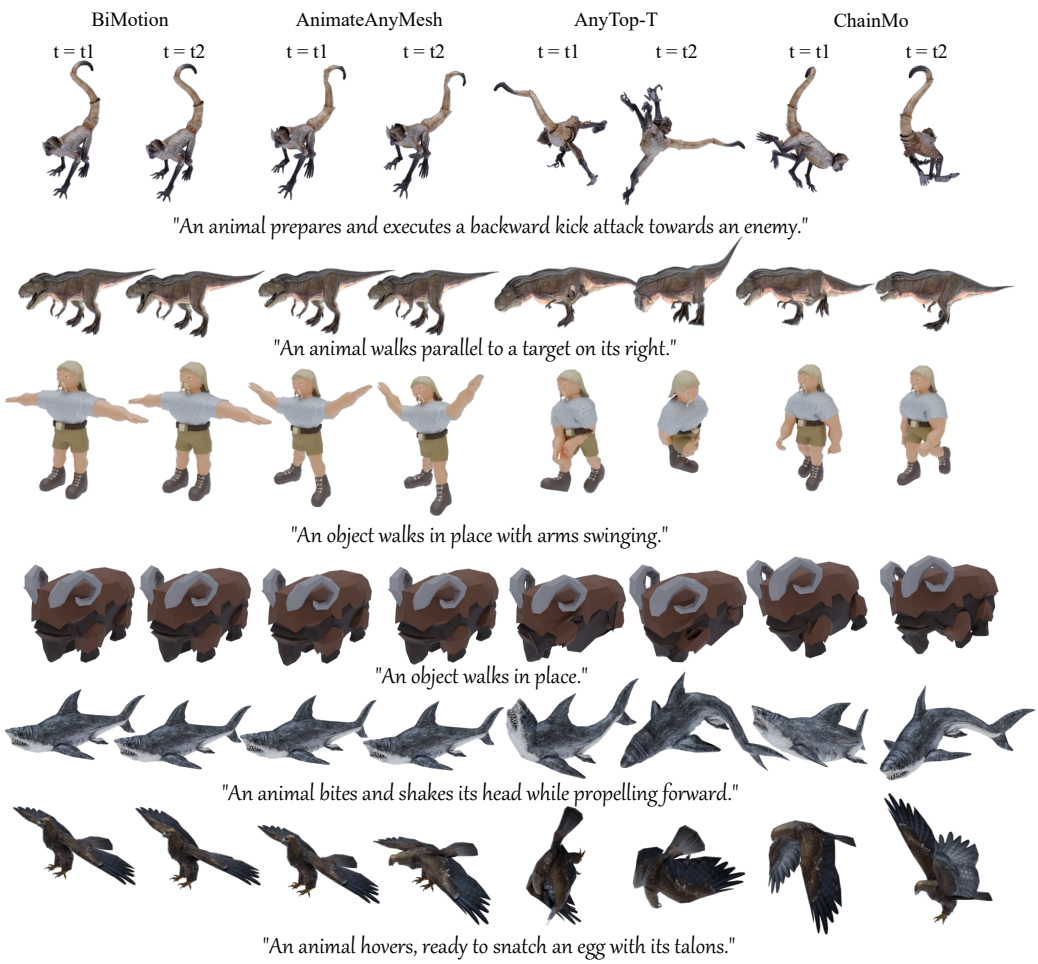


Figure 9: **Additional qualitative comparisons for text-conditioned generation.** We compare ChainMo with BiMotion, AnimateAnyMesh, and AnyTop-T on diverse articulated characters. Two representative frames are shown for each motion. In these examples, ChainMo better captures the described motion semantics while maintaining coherent articulation across heterogeneous skeletal structures. Full motion sequences and additional video results are available on the anonymous project page.

The training distribution is highly imbalanced across articulated structures and motion patterns, with many rare behaviors represented by only a small number of examples. Moreover, text annotations are often sparse, coarse-grained, or inconsistent, particularly for non-human characters and unusual body structures. Consequently, the learned motion prior may inherit dataset biases and dominate weak textual supervision, leading to repetitive motions, reduced controllability, or less appropriate generation for underrepresented characters and behaviors. Improving both motion diversity and structure-aware motion-language annotations is therefore an important direction for more general cross-topology motion generation.

Sensitivity to Structural Topology Changes. ChainMo is designed to reduce sensitivity to within-chain joint discretization rather than to be invariant to arbitrary modifications of skeletal topology. Because the chain decomposition is determined by the branching structure of the kinematic tree, structural edits that introduce or remove branches, such as additional leaf, helper, or auxiliary joints, can change the resulting set of chain segments. While the variable-size representation allows ChainMo to accommodate such structures, the corresponding motion units are no longer preserved in the same way as when joints are inserted or removed within a non-branching chain. Developing representations that are simultaneously robust to both skeletal discretization and larger structural edits remains an interesting direction for future work.

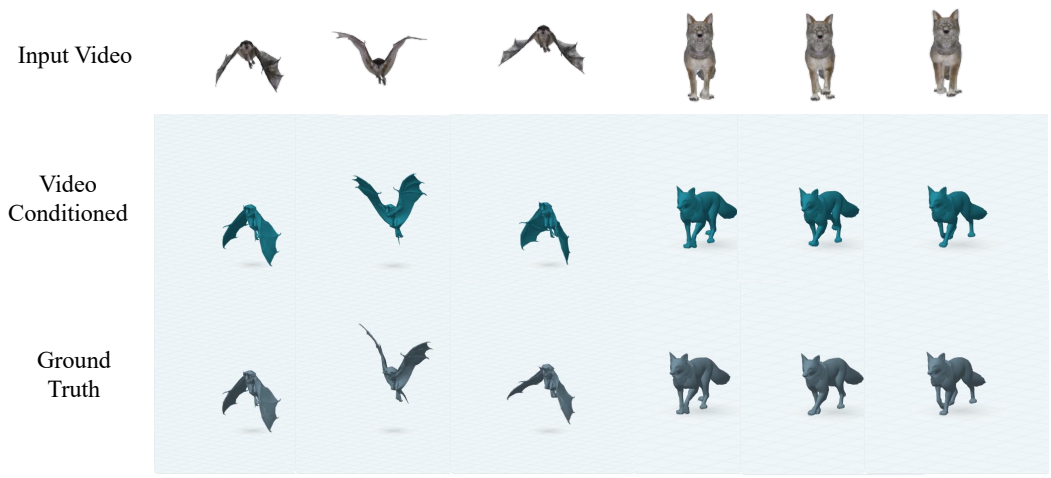


Figure 10: **Video-conditioned motion generation.** We show representative video-conditioned results on heterogeneous articulated characters. Given an input video, ChainMo transfers the observed motion dynamics to the target skeleton and produces motions that closely follow the corresponding ground-truth animation. Representative frames from two examples are shown.

Evaluation across Arbitrary Topologies. There is currently no widely adopted quantitative benchmark for text-driven motion generation across arbitrary skeletal topologies. Our heterogeneous text-generation evaluation therefore relies on qualitative comparisons and a user study in addition to standard HumanML3D evaluation. Developing standardized datasets and perceptual metrics for cross-topology motion generation would enable more comprehensive comparison in future work.