

UniMate: One Unified Model to Animate Diverse Skeletons

LINZHAN MOU, Princeton University, USA
JIAHUI LEI, University of California, Berkeley, USA
ZHIYANG DOU, Massachusetts Institute of Technology, USA
CHENYUE CAI, Princeton University, USA
CHAOYUE SONG, Nanyang Technological University, Singapore
ADAM FINKELSTEIN, Princeton University, USA
SZYMON RUSINKIEWICZ, Princeton University, USA



Fig. 1. Given a rigged 3D asset and a text prompt, UniMate generates animations for characters with *diverse skeletal topologies* within a single *unified* model.

Recent advances in automatic rigging now deliver animation-ready 3D assets at scale, yet generating the motion to drive them remains a bottleneck.

Authors' Contact Information: Linzhan Mou, Princeton University, Princeton, New Jersey, USA, linzhan@princeton.edu; Jiahui Lei, University of California, Berkeley, Berkeley, California, USA, leijh@berkeley.edu; Zhiyang Dou, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, frankdou@mit.edu; Chenyue Cai, Princeton University, Princeton, New Jersey, USA, cc4880@princeton.edu; Chaoyue Song, Nanyang Technological University, Singapore, Singapore, chaoyue002@e.ntu.edu.sg; Adam Finkelstein, Princeton University, Princeton, New Jersey, USA, af@princeton.edu; Szymon Rusinkiewicz, Princeton University, Princeton, New Jersey, USA, smr@princeton.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGGRAPH Asia '26, Kuala Lumpur, Malaysia

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

Existing learned animators are topology-constrained: they rely on category-specific templates or require per-skeleton fine-tuning and reference motions at inference. We present UniMate, a unified foundation model that synthesizes articulated motion for arbitrary skeletons from a rigged 3D asset and a text prompt, with no test-time optimization or per-skeleton retraining. UniMate relies on a topology-aware diffusion transformer, which integrates skeletal topology into attention via three mechanisms: (1) a graph-aware attention bias from pairwise joint relations and geodesic distances; (2) a spectral rotary position embedding generalizing RoPE to arbitrary kinematic trees via the graph Laplacian; and (3) a global topological conditioner attention-pooled from the rest-pose skeleton. We also curate UniML3D, 13,006 motion sequences spanning bipedal, quadrupedal, avian, marine, insectoid, serpentine, and articulated rigid objects with unified canonicalization and text pairing. Trained on this dataset, UniMate outperforms state-of-the-art baselines in quality, generalization, and efficiency, and supports zero-shot cross-topology transfer, in-betweening, expansion, and text-guided editing. Our project page is available at <https://linzhanmou.com/unimate/>.

CCS Concepts: • **Computing methodologies** → **Animation; Artificial intelligence; Machine learning.**

Additional Key Words and Phrases: Motion Synthesis, Skeletal Animation, 3D Character Animation, Diffusion Models, Topology-Aware Learning

ACM Reference Format:

Linzhao Mou, Jiahui Lei, Zhiyang Dou, Chenyue Cai, Chaoyue Song, Adam Finkelstein, and Szymon Rusinkiewicz. 2026. UniMate: One Unified Model to Animate Diverse Skeletons. In *Proceedings of SIGGRAPH Asia 2026 Conference Papers (SIGGRAPH Asia '26)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 Introduction

Character animation is fundamental to 3D content creation, central to film, gaming, virtual reality, and robotics simulation. Traditional pipelines define a *unique* skeleton per character, over which artists craft motion through manual keyframing, motion capture, or slow per-asset optimization. Recent advances in 3D content creation [Li et al. 2024c; Liu et al. 2023b,a; Nichol et al. 2022; Wang et al. 2023; Xiang et al. 2024; Zhao et al. 2025] and automatic rigging [Liu et al. 2025; Song et al. 2025a,b; Xu et al. 2020; Zhang et al. 2025b] now deliver skeleton-ready 3D assets across a broad range of categories, from humans and animals to articulated rigid objects. Yet while *rigged assets* can be generated at scale, the motion that drives them cannot—animation remains the laborious final bottleneck in an otherwise automated 3D pipeline.

The bottleneck lies in the design of existing learned animators. State-of-the-art motion generators [Karunratanakul et al. 2023; Petrovich et al. 2021; Raab et al. 2023; Rempe et al. 2026; Tevet et al. 2024, 2023; Wen et al. 2025; Zhao et al. 2024a] are typically topology-constrained, relying on fixed category-specific skeleton templates such as SMPL [Loper et al. 2023] or SMAL [Zuffi et al. 2017]. Topology-agnostic models [Gat et al. 2025; Li et al. 2022; Raab et al. 2024b] relax these constraints but still require per-skeleton fine-tuning or reference motions at inference time. Mesh-based methods avoid explicit skeleton modeling altogether, but either depend on costly per-asset distillation [Chen et al. 2025a; Jiang et al. 2024; Uzolas et al. 2025] or regress kinematically unconstrained vertex-wise deformations [Shi et al. 2025; Wu et al. 2025; Zhang et al. 2025c,a]. These limitations motivate a unified foundation model that can synthesize motion for arbitrary skeletal topologies directly from high-level descriptions. Such a model would support a broad range of applications in a versatile and feed-forward manner, while sharing data-driven motion priors across diverse subjects and prompts (Figs. 5, 6, 9 and 10 and ??).

However, developing such a unified animator poses two fundamental challenges. The first is *modeling*: real-world skeletons are highly heterogeneous: bipedal humans, multi-legged insects, winged animals, and articulated rigid objects all exhibit distinct kinematic trees, joint counts, and motion patterns. This challenge is further compounded by the diverse motion behaviors associated with different morphologies. A general-purpose model must therefore treat skeletal topology as an explicit input, rather than baking it into an architectural prior, and reason jointly over structure and motion. The second is *data*: text-paired motion corpora spanning diverse skeletal topologies remain scarce, with existing benchmarks dominated by humans [Guo et al. 2022; Mahmood et al. 2019; Plappert et al. 2016] and a limited number of quadrupeds [Yang et al. 2024]. Meanwhile, raw rigged 4D assets [Deitke et al. 2023a,b; Studios 2022]

are often noisy and inconsistent, and lack unified preprocessing and canonicalization across topologies, leaving data-driven approaches without coherent supervision for cross-topology generalization.

In this work, we propose **UniMate**, a unified foundation model that animates diverse skeletons. Given a rigged 3D asset and a natural-language prompt, UniMate synthesizes plausible articulated motion with no test-time fitting or per-skeleton specialization (see Figure 1). Joint training across a wide range of skeletons lets the model learn motion patterns that are shared and transferable across topologies, enabling stronger generalization to unseen rigs and motion transfer between heterogeneous structures.

At the core of UniMate is the Topology-Aware Diffusion Transformer (TADiT), a flow-matching architecture in which attention layers jointly reason over rest-pose kinematics and motion manifolds through a shared token stream. To encode heterogeneous topologies, we equip TADiT with three key design choices. First, vanilla self-attention is blind to the underlying kinematic graph. We therefore inject a *graph-aware attention bias* [Ying et al. 2021] derived from pairwise joint relations and geodesic distances, so anatomically nearby joints attend more strongly while the model retains its capacity for long-range, full-body coordination. Second, we introduce *Spec-RoPE*, a spectral rotary position embedding that generalizes RoPE [Su et al. 2024] to arbitrary kinematic trees by deriving rotary angles from the graph Laplacian spectrum. With provable translation invariance in spectral coordinates and equivariance under joint reindexing, Spec-RoPE adapts to skeletons of varying size and connectivity—a property that index- or coordinate-based encodings cannot provide. Third, a *global topological conditioner*, attention-pooled from the topology tokens, modulates every transformer block through AdaLN-Zero [Peebles and Xie 2023], so layer-wise feature statistics adapt to the input skeleton and provide global structural context complementing the local signals above.

To support training at scale, we curate UniML3D, a heterogeneous motion dataset of 13,006 animation sequences (roughly 20 hours) drawn from Truebones [Studios 2022], Mixamo [Adobe 2022], and Objaverse [Deitke et al. 2023a,b], pairing thousands of distinct rigs across bipedal, quadrupedal, avian, marine, insectoid, serpentine, and articulated rigid objects with 3,584 unique text prompts that span a broad action spectrum, including locomotion, combat, idle, mechanical articulation, and object manipulation. Rigorous filtering followed by unified canonicalization yields a shared representation across skeleton types, while online skeletal augmentation further broadens topological coverage during training. This scale and coverage substantially exceed prior work [Gat et al. 2025] and are essential for cross-skeleton generalization.

Extensive experiments demonstrate that UniMate achieves state-of-the-art performance on topology-agnostic motion generation and mesh animation, surpassing prior methods in quality, generalization, and efficiency. UniMate also supports zero-shot downstream tasks such as cross-topology motion transfer, in-betweening, expansion, and text-guided editing, serving as a controllable engine for scalable 3D character animation. In summary, our contributions are:

- We present UniMate, a unified foundation model that synthesizes articulated motion for skeletons of arbitrary topology from a rigged 3D asset and a text prompt.

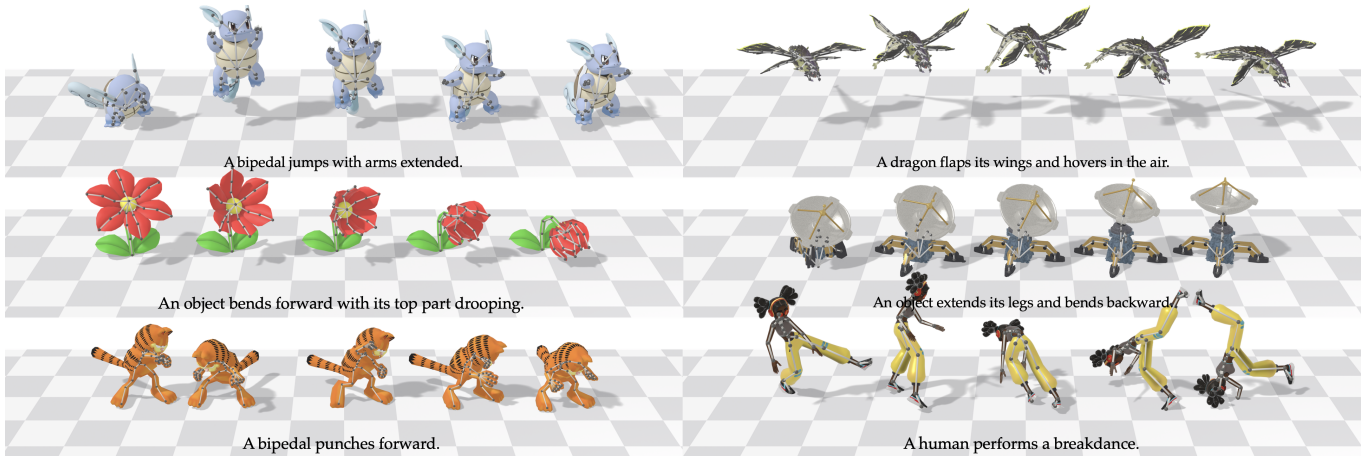


Fig. 2. **Animations generated by UniMate.** Our method generalizes across heterogeneous skeletons and diverse motion prompts.

- We propose TADiT, which couples motion and skeletal structure within shared attention layers through a graph-aware attention bias, the Spec-RoPE spectral rotary position embedding with provable structural properties, and a global topological conditioner.
- To facilitate model training and benchmarking, we curate UniML3D, comprising 13,006 motion sequences over thousands of skeletons spanning bipedal, quadrupedal, avian, marine, insectoid, serpentine, and articulated rigid objects, unified by a canonicalization pipeline and online skeletal augmentation.
- We conduct various experiments showing that UniMate improves over prior methods in quality, generalization, and runtime, and enables zero-shot cross-topology motion transfer, in-betweening, expansion, and real-time text-guided editing.

2 Related Work

2.1 3D Animation

A growing body of work animates 3D assets by distilling image- and video-generative priors [Chen et al. 2025a; Huang et al. 2025; Jiang et al. 2024; Liang et al. 2024; Lyu et al. 2026; Mou et al. 2025; Uzolas et al. 2025; Yenphraphai et al. 2025; Yun et al. 2025]. These methods typically require costly per-asset test-time optimization and easy to produce jittery, unstable trajectories, making them ill-suited to large-scale or interactive use. A second class of methods [Chen et al. 2026; Sabathier et al. 2026; Shi et al. 2025; Wang et al. 2026b; Wu et al. 2025; Zhang et al. 2025c,a] trains feed-forward networks that regress per-vertex or per-point deformations directly. Operating in raw geometry space sacrifices the compactness of skeletal representations and breaks native compatibility with the rig-driven ecosystem: linear blend skinning [Kavan et al. 2007; Li et al. 2021; Magnenat-Thalmann et al. 1988], articulated controllers [Tevet et al. 2024], physics simulators [Makovychuk et al. 2021; Todorov et al. 2012], and motion-capture pipelines [Gong et al. 2025]. Recent advances in automatic rigging [Liu et al. 2025; Song et al. 2025a,b; Xu et al. 2020; Zhang et al. 2025b] have made high-quality skeletons widely accessible, yet driving the resulting rigs still demands manual keyframing or expensive per-asset optimization [Li et al. 2025; Song et al. 2025a; Xie et al. 2025]. UniMate closes this gap with a data-driven generative model that operates directly on arbitrary skeletons.

2.2 Cross-Topology Motion Generation and Retargeting

The dominant family of learned motion generators [Chen et al. 2024; Dou et al. 2025; Karunratanakul et al. 2023; Li et al. 2024a; Meng et al. 2025; Raab et al. 2024a, 2023; Rempe et al. 2026; Sawdayee et al. 2026; Shafir et al. 2024; Sun et al. 2024; Tevet et al. 2024, 2023; Wan et al. 2024; Wang et al. 2026a, 2025; Wen et al. 2025; Wu et al. 2023; Zhao et al. 2024a; Zhou et al. 2024] is built on category-specific parametric body models such as SMPL [Loper et al. 2023; Pavlakos et al. 2019] and SMAL [Zuffi et al. 2017], inheriting their fixed joint count, hierarchy, and skinning, and therefore cannot handle characters whose topology departs from the template. A separate family of example-based methods sidesteps neural networks training by stitching patches from exemplar motion clips, originating with Drop-the-GAN [Granot et al. 2022] and later refined by generative motion matching [Li et al. 2023] and sparse-correspondence cross-topology transfer [Chen et al. 2025b]. Although topology-flexible, these methods still require an exemplar motion for every target and cannot synthesize motion from a static rigged asset alone. Closer to our setting, GANimator [Li et al. 2022] and SinMDM [Raab et al. 2024b] learn neural generators on arbitrary topologies, but train a separate model per skeleton and so do not generalize across structures. Cross-topology retargeting [Aberman et al. 2020; Lee et al. 2023; Li et al. 2024b; Liu et al. 2026; Zhao et al. 2024b] bypasses the template constraint from a different angle, but again only by transferring an existing source motion onto a target skeleton. The closest prior work, AnyTop [Gat et al. 2025], jointly trains a single diffusion model over heterogeneous skeletons, but is limited to a small animal corpus [Studios 2022], requires a user-supplied reference motion at inference, and supports only unconditional generation. In contrast, UniMate covers a much broader range of articulated topologies (humans, animals, and general articulated objects), accepts text conditioning, and maps a rigged mesh end-to-end to animated skeleton-mesh sequences without reference motion or per-skeleton training.

3 Method

Given a rigged 3D asset and a text prompt, our goal is to synthesize a plausible motion sequence that animates the input mesh. We first introduce a unified representation for heterogeneous skeletons and

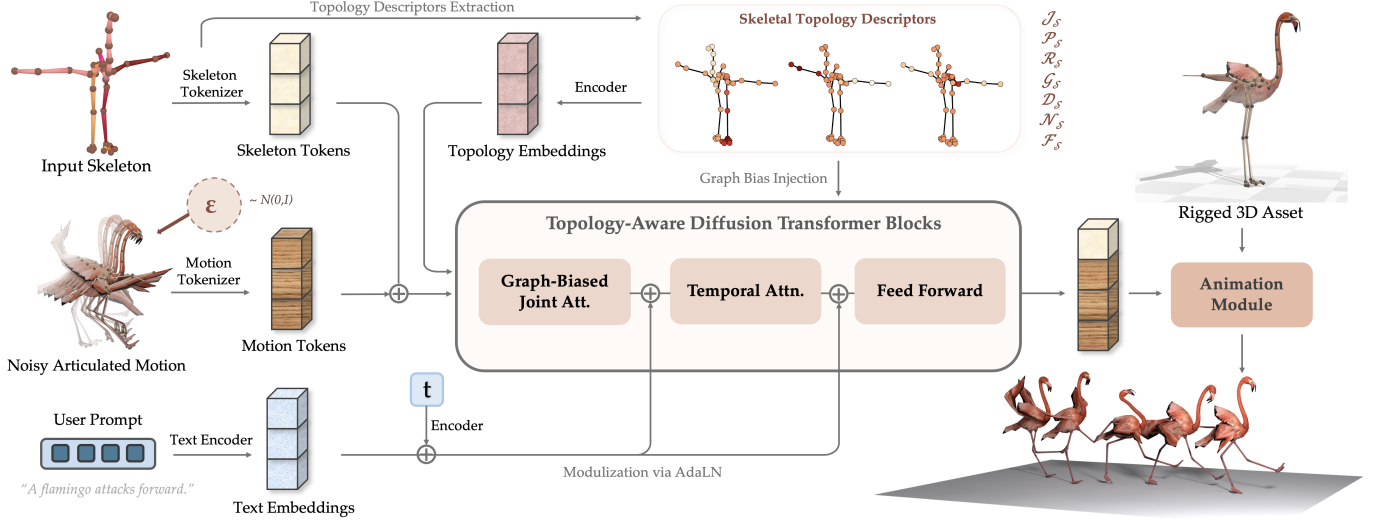


Fig. 3. **UniMate pipeline.** The proposed TADiT operates on a joint token space that combines per-frame motion features with rest-pose skeletal descriptors, while injecting skeletal graph-structured bias into both positional encoding and attention computation. Conditioned on input text prompts, our unified model produces realistic, coherent animations while demonstrating strong cross-topology generalization and efficient runtime performance.

motion sequences (Section 3.1), then present our Topology-Aware Diffusion Transformer (TADiT), and finally describe the training objective and inference procedure (Section 3.3).

3.1 Skeleton and Motion Representation

An articulated rigged 3D asset is animated by a *skeleton* – a kinematic tree with joint hierarchy and bone lengths defined, whose joints drive mesh deformation through forward kinematics. A *rest pose* of a skeleton is the neutral undeformed reference configuration. Our model takes as input the skeleton at rest-pose and build diffusion model over a motion sequence defined on it.

Skeleton definition. We model an articulated object as a rooted kinematic tree with J joints. To represent skeletons with heterogeneous topologies in a shared transformer space, we first assign each skeleton a canonical joint ordering. Concretely, we linearize the tree using a breadth-first traversal (BFS) given the known root node

$$\mathcal{J}_S = [(j_1, \text{pa}_1), (j_2, \text{pa}_2), \dots, (j_J, \text{pa}_J)], \quad (1)$$

where $j_k \in \mathbb{R}^3$ denotes the rest-pose position of the k -th joint and $\text{pa}_k \in \{1, \dots, J\}$ denotes its parent index.

Topology-diameter normalization. Skeletons in our dataset vary significantly in absolute size and topological extent. To place them into a comparable representation space, we normalize each rest-pose skeleton by its topology diameter, defined as

$$d_{\text{topo}} = \max_{i,j} \text{dist}_{\text{tree}}^{\text{geo}}(i, j), \quad (2)$$

where $\text{dist}_{\text{tree}}^{\text{geo}}(i, j)$ denotes the geodesic distance between joints i and j along the kinematic tree. This normalization removes scale differences while preserving the relative kinematic structure.

Local topology descriptors. Given the canonicalized skeleton, we describe its rest-pose geometry and local topology using several descriptors. Rest-pose joint positions are stacked in $\mathcal{P}_S \in \mathbb{R}^{J \times 3}$. Local kinematic context is encoded by a pairwise relation matrix

$\mathcal{R}_S \in \mathbb{N}_0^{J \times J}$, whose entries specify relation types such as parent, child, sibling, or ancestor; a pairwise graph-distance matrix $\mathcal{G}_S \in \mathbb{N}_0^{J \times J}$ on the kinematic tree; a per-joint depth vector $\mathcal{D}_S \in \mathbb{N}_0^J$; and a per-joint text description $\mathcal{N}_S \in \mathbb{N}_0^J$ that captures semantic identity.

Spectral coordinates. To complement the discrete topological descriptors above with a continuous encoding of skeletal structure, we additionally compute spectral features from the kinematic graph. Let $A \in \{0, 1\}^{J \times J}$ denote the adjacency matrix of the skeleton, $D = \text{diag}(\sum_j A_{ij})$ the corresponding degree matrix, and $L_S = D - A$ the graph Laplacian. Let $L_S = U\Lambda U^T$ be its eigendecomposition. Discarding the trivial constant eigenvector u_0 , we define the spectral feature of joint j using the first m non-trivial eigenvectors:

$$\mathcal{F}_S(j) = [u_1(j), u_2(j), \dots, u_m(j)] \in \mathbb{R}^m. \quad (3)$$

These spectral coordinates provide a continuous encoding of joint location on the kinematic graph (see Fig. 7), complementing the discrete relation and distance descriptors.

Overall, we represent a skeleton as

$$\mathcal{S} = \{\mathcal{J}_S, \mathcal{P}_S, \mathcal{R}_S, \mathcal{G}_S, \mathcal{D}_S, \mathcal{N}_S, \mathcal{F}_S\}, \quad (4)$$

where $\mathcal{J}_S$ specifies the ordered kinematic tree, $\mathcal{P}_S$ the rest-pose geometry, $\mathcal{R}_S, \mathcal{G}_S, \mathcal{D}_S$ the discrete topological structure, $\mathcal{N}_S$ the semantic joint identity, and $\mathcal{F}_S$ the global spectral descriptor.

Motion Representation. Given a skeleton $\mathcal{S}$, a motion sequence is $X = \{\mathbf{m}_j^t\} \in \mathbb{R}^{T \times J \times D}$ over T frames and J joints. Inspired by [Gat et al. 2025; Guo et al. 2022], each joint feature has dimension $D = 12$:

$$\mathbf{m}_j^t = [\mathbf{p}_j^t, \mathbf{r}_j^t, \mathbf{v}_j^t] \in \mathbb{R}^D. \quad (5)$$

$\mathbf{p}_j^t \in \mathbb{R}^3$: *Position.* For non-root joints, horizontal (x, z) components are taken relative to the root and rotated into the facing-canonical frame, with vertical y kept in the world frame. The root joint $\mathbf{p}_0^t$ retains its global position.

$\mathbf{r}_j^t \in \mathbb{R}^6$: *Rotation.* Represented in the continuous 6D format [Zhou et al. 2019].

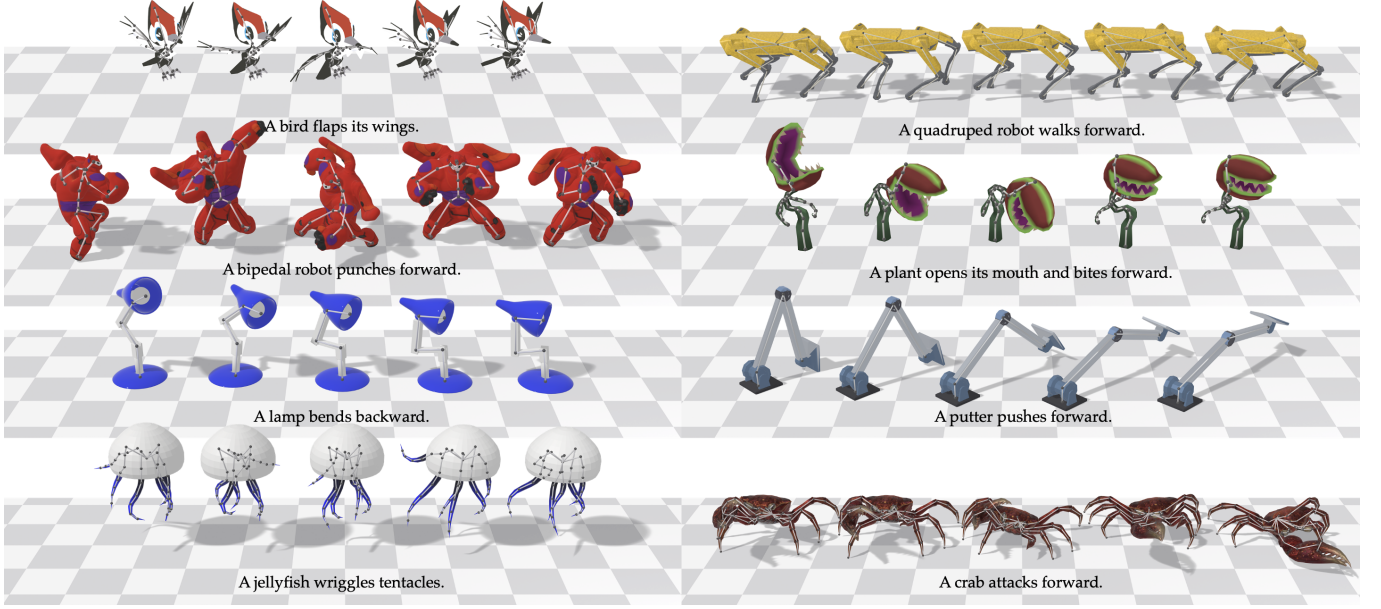


Fig. 4. **Cross-topology animation.** UniMate generates prompt-aligned motions for diverse characters and articulated objects within a single unified model.

$\mathbf{v}_j^t \in \mathbb{R}^3$: *Velocity*. Computed in the world frame as the temporal derivative of global joint positions, preserving absolute motion cues that the canonical-frame projection of $\mathbf{p}_j^t$ would otherwise discard.

3.2 Topology-Aware Diffusion Transformer (TADiT)

Our goal is to generate a plausible motion sequence $\hat{\mathbf{x}}_1$ conditioned on a rest-pose skeleton $\mathcal{S}$ and a text prompt c . This is challenging because the model must generalize across heterogeneous skeletons with different kinematic trees, joint counts, and motion patterns. To address this, we introduce the *Topology-Aware Diffusion Transformer* (TADiT), which injects topology through graph-aware attention bias, spectral rotary positional encoding (Spec-RoPE), and a global topological conditioner. We train TADiT with conditional flow matching [Liu et al. 2022], learning a velocity field $v_\theta(x_r, \tau, c, \mathcal{S})$ that transforms Gaussian noise $x_0 \sim \mathcal{N}(0, I)$ into a valid motion sequence conditioned on both text and skeletal topology.

Skeleton and Motion Tokenization. For each joint j , we form the skeleton token $\mathbf{T}_j \in \mathbb{R}^d$ by concatenating MLP-projected rest-pose positions of the joint and its parent and passing a fusion MLP:

$$\mathbf{T}_j = \text{MLP}\left(\text{Concat}(\text{MLP}(\mathbf{p}_j), \text{MLP}(\mathbf{p}_{\text{pa}(j)}))\right). \quad (6)$$

Each motion feature $\mathbf{m}_j^t \in \mathbb{R}^D$ (defined in Section 3.1) is projected to dimension d and augmented with learnable depth embedding and joint-name text embeddings as a hierarchical and semantic prior:

$$\mathbf{M}_j^t = \text{MLP}(\mathbf{m}_j^t) + E_{\text{depth}}(\mathcal{D}_S(j)) + E_{\text{name}}(\mathcal{N}_S(j)). \quad (7)$$

We prepend the skeleton tokens to the motion tokens along the temporal axis, forming $Z = \text{Concat}(\mathbf{T}, \mathbf{M}) \in \mathbb{R}^{(T+1) \times J \times d}$, such that every transformer block operates on a unified token space containing both static skeletal structure and dynamic per-frame motion.

Skeletal-Temporal Transformer Blocks. Each block operates on $Z \in \mathbb{R}^{(T+1) \times J \times d}$ and is conditioned on the diffusion timestep, text prompt, and global topology embeddings. For tractable cost on

heterogeneous skeletons, each block uses a factorized attention with a *joint* branch (across joints at each frame) and a *temporal* branch (across frames at each joint), followed by a feed-forward sublayer:

$$Z^{(\ell+1)} = \text{FFN}\left(\text{TempAttn}(\text{JointAttn}(Z^{(\ell)}, \mathcal{S}), c), c\right). \quad (8)$$

The temporal branch is a standard multi-head self-attention with 1D rotary position embedding (RoPE) [Su et al. 2024] on the frame index. Kinematic structure is exposed exclusively to the joint branch, through a graph-aware attention bias and the Spectral Rotary Position Embedding (Spec-RoPE) described next.

Graph-Aware Attention Bias. We inject the kinematic graph into attention through a learned bias added to the joint-attention logits, exposing pairwise structural relations that vanilla self-attention is hard to recover from token features alone. Inspired by [Ying et al. 2021], the pairwise graph-distance and relation descriptors $\mathcal{G}_S, \mathcal{R}_S$ are embedded by lookup tables

$$e_d(i, j) = E_d(\mathcal{G}_S(i, j)), \quad e_r(i, j) = E_r(\mathcal{R}_S(i, j)), \quad (9)$$

and projected to a per-head scalar bias

$$B_{ij}^{(h)} = (\mathbf{w}_d^{(h)})^\top e_d(i, j) + (\mathbf{w}_r^{(h)})^\top e_r(i, j), \quad (10)$$

with head-specific projection vectors $\mathbf{w}_d^{(h)}, \mathbf{w}_r^{(h)} \in \mathbb{R}^{d_e}$. Letting $\tilde{q}_i^{(h)}, \tilde{k}_j^{(h)} \in \mathbb{R}^{d_h}$ denote the Spec-RoPE-rotated queries and keys for joints i and j , the joint-attention logits read

$$\text{Attn}^{(h)}(i, j) = \frac{1}{\sqrt{d_h}} (\tilde{q}_i^{(h)})^\top \tilde{k}_j^{(h)} + B_{ij}^{(h)}. \quad (11)$$

The bias is shared across frames, so its memory cost is independent of sequence length. Because it is parameterized by graph-distance and relation-type embeddings rather than absolute joint indices, it transfers to unseen topologies (Section 5.5).

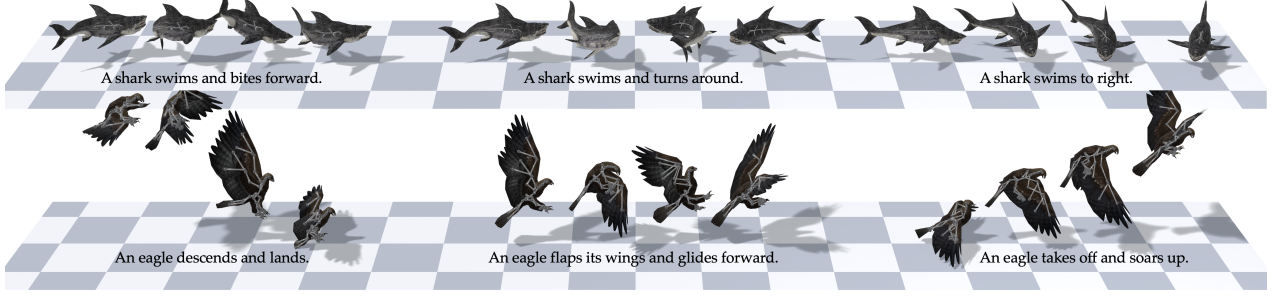


Fig. 5. **One skeleton, diverse prompts.** Given a single skeleton, UniMate synthesizes diverse motions that follow different input text prompts.

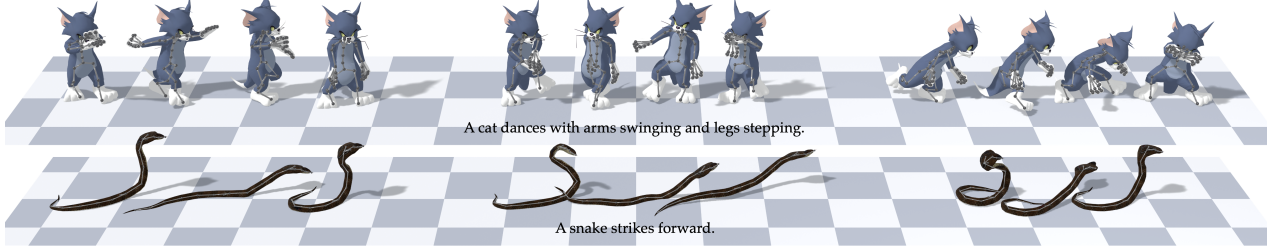


Fig. 6. **One skeleton, one prompt, diverse motions.** Given the same skeleton and text prompt, UniMate generates diverse plausible motion samples.

Spectral Rotary Position Embedding (Spec-RoPE). Kinematic trees have no canonical ordering, so the index used by 1D RoPE is ill-defined for joints. We instead derive rotary angles from the spectrum of the graph Laplacian, applied to the joint branch only:

$$\underbrace{\theta_j = \omega j}_{\text{1D RoPE}} \longrightarrow \underbrace{\theta_j = f(s_j)}_{\text{Spec-RoPE}}, \quad (12)$$

where $j \in \mathbb{N}$ is the token index, $\omega \in \mathbb{R}^{d_h/2}$ the standard frequency vector, $s_j = \mathcal{F}_S(j) \in \mathbb{R}^m$ the joint’s spectral coordinate from Section 3.1, and f a learned angle map specified below.

Intuition. RoPE relies on positional coordinates to define relative phase offsets in attention. For temporal tokens, the frame index is a natural causal position; for kinematic-tree joints, the BFS index is arbitrary: two joints adjacent in the index can lie on opposite limbs. The spectral coordinate s_j replaces it with an intrinsic position on the graph, where the leading eigenvector traces the skeleton’s principal axis (head-to-feet for a humanoid, nose-to-tail for a quadruped) and successive eigenvectors resolve finer structure (Fig. 7). The rotary phase therefore depends on *where a joint sits on the skeleton*, not on how it is serialized.

The spectral coordinate $s_j = \mathcal{F}_S(j) \in \mathbb{R}^m$ comprises the leading m non-trivial eigenvectors of L_S . Since these are determined only

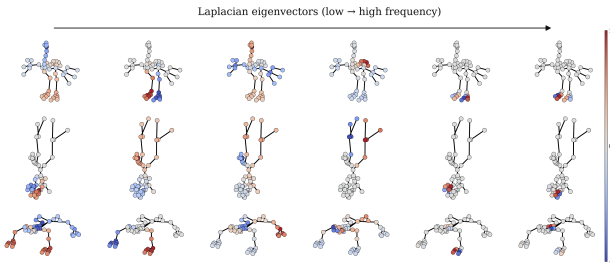


Fig. 7. **Spectral visualization.** From left to right, Laplacian eigenvectors increase in frequency. Low-frequency modes capture global kinematic structure, while higher-frequency modes encode finer local relationships.

SIGGRAPH Asia ’26, December 1–4, 2026, Kuala Lumpur, Malaysia.

up to sign, we realize f as a SignNet [Lim et al. 2023]:

$$\theta_j = \text{MLP}_{\text{ang}} \left(\text{Concat} \left(\left\{ \text{MLP}_{\text{sym}}(u_n(j)) + \text{MLP}_{\text{sym}}(-u_n(j)) \right\}_{n=1}^m \right) \right). \quad (13)$$

The angles $\theta_j \in \mathbb{R}^{d_h/2}$ then drive the standard RoPE block-diagonal rotation $\mathcal{R}(\theta_j)$ [Su et al. 2024], applied to per-joint queries and keys as $\tilde{q}_j = \mathcal{R}(\theta_j) q_j$ and $\tilde{k}_j = \mathcal{R}(\theta_j) k_j$.

Spec-RoPE satisfies two structural properties, formalized in the supplemental material: *translation invariance in spectral coordinates* and *equivariance under joint reindexing*. Together, these allow Spec-RoPE to adapt to skeletons of varying size and connectivity. We empirically validate the effect of Spec-RoPE in Section 5.5.

Global Topological Conditioner. Beyond the pairwise graph attention bias and per-joint Spec-RoPE topology signals, each transformer block also requires a single, joint-count-invariant skeleton summary. We obtain this global topological condition c_{topo} via *attention pooling* [Lee et al. 2019] over the topology token sequence T , producing a fixed-size summary independent of joint count and, after the final mean aggregation, of joint ordering. A small set of n_q learnable query tokens $Q_{\text{pool}} \in \mathbb{R}^{n_q \times d}$ serves as a content-adaptive readout: each query attends to all topology tokens through cross-attention, with T providing the keys and values,

$$H = \text{softmax} \left(\frac{1}{\sqrt{d}} (Q_{\text{pool}} W_q) (T W_k)^\top \right) T W_v, \quad (14)$$

where $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ are learnable projections. The query outputs $H \in \mathbb{R}^{n_q \times d}$ are aggregated by mean pooling and then injected via AdaLN, yielding a holistic context of the skeletal structure.

3.3 Training and Inference

Training Objective. We supervise v_θ with three complementary losses, with full expressions deferred to the appendix.

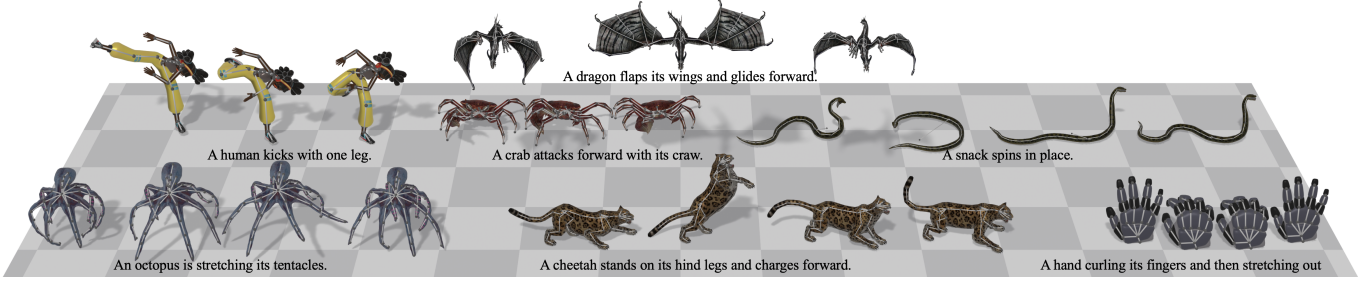


Fig. 8. **Samples from UniML3D dataset.** Our dataset spans diverse skeletons across bipedal, quadrupedal, avian, marine, insectoid, serpentine, and articulated rigid objects, with detailed skeleton annotations and coherent text prompts paired with motion sequences.

Flow-matching MSE. The base loss $\mathcal{L}_{\text{mse}}$ is a masked mean-squared error against the target velocity $v^* = x_1 - x_0$, with padded joint slots zeroed out in heterogeneous-skeleton batches.

Geodesic rotation loss. Because joint rotations live on the non-Euclidean manifold $SO(3)$, an isotropic MSE on their 6D channels is geometrically misaligned. We therefore add a geodesic loss $\mathcal{L}_{\text{geo}}$ on the one-step denoised rotations $\hat{R}_j^t \in SO(3)$ from $\hat{x}_1 = x_\tau + (1-\tau)v_\theta$, that penalizes their angular deviation from ground truth.

Velocity smoothness. To suppress high-frequency jitter, a smoothness regularizer $\mathcal{L}_{\text{smooth}}$ penalizes the temporal acceleration of the denoised velocity channels of $\hat{x}_1$.

The final objective is a weighted combination of these three losses:

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}. \quad (15)$$

Inference. At inference time, we draw $x_0 \sim \mathcal{N}(0, I)$ and integrate the learned velocity field $\dot{x}_\tau = v_\theta(x_\tau, \tau, c, \mathcal{S})$ from $\tau = 0$ to 1 using a fixed-step Euler solver, yielding the generated motion features $\hat{x}_1$. Each step applies classifier-free guidance [Ho and Salimans 2022].

4 UniML3D Dataset

Training a truly generalizable animation model for diverse object categories requires a large-scale dataset with varied skeletal structures and plausible motion sequences. However, raw 4D motion sources are noisy and inconsistent, often containing disconnected or scene-level skeletons, broken roots, non-functional joints, physically implausible motions, and mismatched coordinate frames or facing directions, making them unsuitable for direct cross-topology motion learning without rigorous preprocessing.

To address this, we curate UniML3D by unifying three complementary sources: Truebones [Studios 2022], Mixamo [Adobe 2022], and Objaverse-XL [Deitke et al. 2023a,b]. After rigorous filtering and canonicalization, UniML3D comprises 13,006 motion sequences and 2,140,232 frames over thousands of unique skeletons, spanning bipedal, quadrupedal, avian, aquatic, insectoid, serpentine, and articulated rigid objects (Fig. 8). Per-source statistics are provided in the supplemental material.

Skeleton-Based Filtering. We apply a multi-stage filter to ensure that every retained skeleton is connected, kinematically valid, and exhibits non-trivial motion:

- (1) *Single-tree pruning.* Multiple or disconnected kinematic trees are pruned, retaining only the primary skeleton: the tree with the largest cumulative skinning weight.

- (2) *Root realignment.* Spurious or misaligned root joints are corrected by propagating their global transformation onto the semantic root via forward kinematics (FK).
- (3) *Phantom-joint removal.* Non-functional joints (e.g., IK controllers and helper bones with zero skinning weight) are recursively pruned to streamline the topology.
- (4) *Static-clip removal.* Sequences with negligible activity, quantified by bone-length-normalized global displacement, are discarded.
- (5) *Implausibility filtering.* Clips exhibiting physical implausibility, such as out-of-distribution root velocities or per-joint angular jitter exceeding an anatomical threshold, are also discarded.

Motion Annotation. Joint labels in the source datasets are highly inconsistent, we standardize them to a fixed anatomical vocabulary via an LLM-based mapping. Each retained clip is then resampled to 30 FPS, rendered to four-view synchronized video, and auto-captioned by multimodal LLMs [Bai et al. 2025; Singh et al. 2025]. A final manual quality-control pass revises or discards motion-caption pairs in which the caption misidentifies the dominant action, ensuring that both the motion data and the textual supervision meet a consistent quality threshold.

Motion Canonicalization. Building on the BFS serialization and topology-diameter normalization of Section 3.1, we map every motion into a unified canonical space:

- (1) Each animation is placed in a canonical coordinate frame with y -axis up and the initial root position at the origin.
- (2) The initial facing direction is aligned with the positive z -axis, estimated as $\mathbf{f} = \text{proj}_{xz}(\mathbf{e}_y \times \mathbf{v}_{\text{hip}})$, where $\mathbf{v}_{\text{hip}}$ is the left-to-right hip (or any symmetric-joint pair) direction.
- (3) Joint rotations are expressed relative to the rest pose, yielding a unified kinematic basis across heterogeneous skeletons.
- (4) We compute global statistics $(\mu_{\text{global}}, \sigma_{\text{global}}) \in \mathbb{R}^D$ over root-joint features and local statistics $(\mu_{\text{local}}, \sigma_{\text{local}}) \in \mathbb{R}^D$ over non-root features, and apply them to normalize all motions.

5 Experiments

We evaluate UniMate from four perspectives: comparing with skeleton-based motion generation methods, benchmarking against skeleton-free mesh animation baselines, ablating our topology-aware design choices, and demonstrating broader applications.

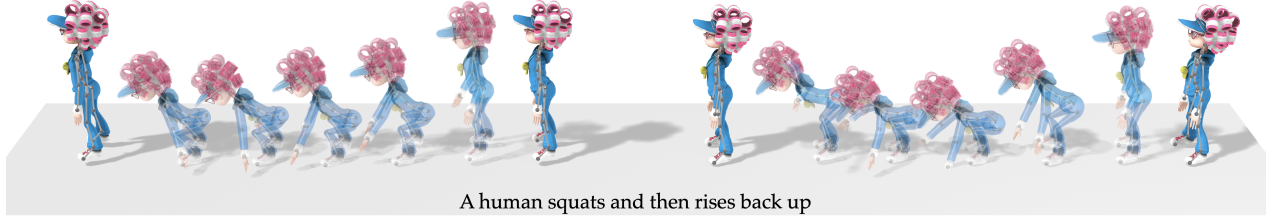


Fig. 9. **Motion in-betweening.** Given the start and end poses and a text prompt, UniMate synthesizes smooth and plausible intermediate motions.

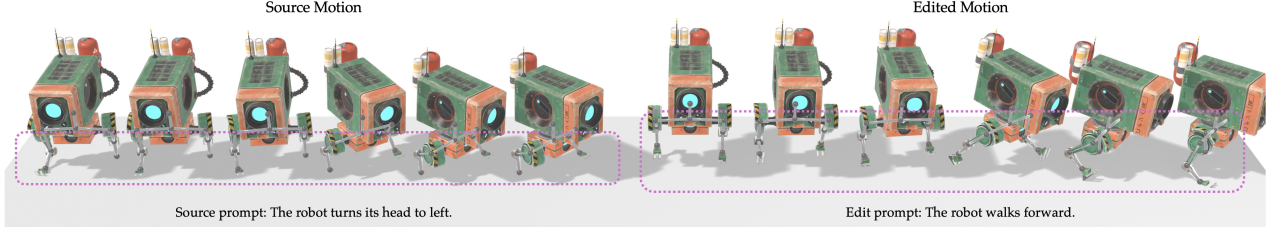


Fig. 10. **Motion editing.** Given a source motion (left), we keep a subset of joints fixed and resample the remaining joint trajectories under new prompt (right).

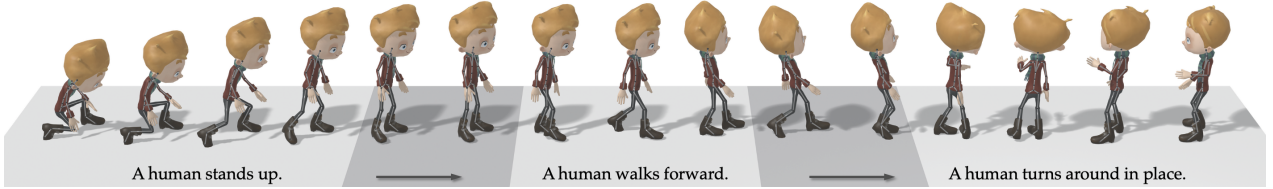


Fig. 11. **Motion expansion.** Given sequential text prompts, UniMate extends a motion with smooth transitions; dark-gray regions denote overlap.

5.1 Implementation Details

Our motion diffusion transformer consists of 8 blocks with hidden dimension 512. We use a frozen FLAN-T5 [Chung et al. 2024] text encoder and train the model with AdamW at a learning rate of 1×10^{-4} . To address the long-tailed topology distribution of UniML3D and improve generalization, we adopt square-root-balanced sampling together with four kinematics-preserving on-the-fly augmentations: joint removal, joint addition, skeleton pooling, and bone-length perturbation. Training is performed with a global batch size of 256 on 8 NVIDIA H100 GPUs for one day, and inference runs at 50 FPS.

5.2 Comparison on Topology-Aware Motion Generation

Baselines. Our task is to generate a motion sequence conditioned on an input skeleton and a text prompt. To the best of our knowledge, no existing baseline directly addresses this setting. The closest prior method is AnyTop [Gat et al. 2025], which supports *unconditional* motion generation for diverse skeletons on Truebones dataset [Studios 2022]. To adapt it to our setting and enable a fair comparison, we augment AnyTop with a cross-attention module for text conditioning, following MDM [Tevet et al. 2023]. We train and evaluate both the extended AnyTop baseline and our model on Truebones.

Metrics. We randomly hold out 7 skeleton types as unseen test skeletons, covering bipedal, quadrupedal, avian, marine, insectoid, and serpentine categories, for a total of 58 motion sequences and 5,861 frames. Specifically, (1) *Motion quality* is measured by the Fréchet Inception Distance (FID) between extracted kinematic features [Chen et al. 2025b; Siyao et al. 2022] of generated and ground-truth motions, and (2) *Diversity* is computed as the average pairwise joint distance among 5 generated samples.

Results. As shown in Table 1, UniMate outperforms AnyTop in both motion quality and diversity, while the qualitative results in ?? further demonstrate stronger generalization to unseen skeletons, where AnyTop often fails to produce dynamic, coherent motions.

5.3 Comparison on Text-Conditioned Mesh Animation

Baselines. To further evaluate UniMate in a broader animation setting, we compare against two SOTA *skeleton-free* and *vertex-wise* mesh animation baselines: (1) AnimateAnyMesh [Wu et al. 2025], a feed-forward, text-conditioned framework; and (2) V2M4 [Chen et al. 2025a], an optimization-based, monocular video-conditioned method. For a fair comparison, all methods are evaluated using the

Table 1. **Comparison on motion generation.** Our model clearly outperforms AnyTop in both quality and diversity.

Method	FID↓	Div.↑
AnyTop [2025]	2.711	8.139
UniMate (Ours)	0.757	9.200

Table 2. **Comparison on mesh animation.** Our method outperforms prior state-of-the-art approaches.

Method	OC↑	MS↑	DD↑	AQ↑	Time↓
V2M4 [2025a]	0.167	0.991	0.667	0.506	1.641h
AnimateAnyMesh [2025]	0.151	0.995	0.352	0.498	15.542s
UniMate (Ours)	0.186	0.993	0.833	0.544	1.214s

Table 3. **User study on mesh animation.** Our model achieves the best text-motion alignment and motion quality.

Method	TA↑	MP↑	ME↑	SP↑	Avg.↑
V2M4 [2025a]	2.378	2.341	2.596	2.336	2.413
AnimateAnyMesh [2025]	2.193	2.930	2.362	3.747	2.808
UniMate (Ours)	4.617	4.568	4.646	4.630	4.615

same text prompts. For V2M4, we use Wan2.2 [Wan Team, Alibaba Group 2025] to generate the driving videos from these prompts.

Metrics. Our benchmark consists of 18 randomly selected meshes spanning bipeds, quadrupeds, avians, marine life, insects, and general articulated objects. Following [Huang et al. 2025; Wu et al. 2025], we render 512×512 multi-view videos from fixed viewpoints and assess perceptual quality with VBench [Huang et al. 2024] along four axes: *overall consistency* (OC), *motion smoothness* (MS), *dynamic degree* (DD), and *aesthetic quality* (AQ). We also report the average generation time per mesh animation. We further conduct a user study with 32 participants, each reviewing 12 test cases and rating each animation on a 5-point Likert scale (1 = very poor, 5 = excellent) along *text-to-motion agreement* (TA), *motion plausibility* (MP), *motion expressiveness* (ME), and *shape preservation* (SP).

Results. In Tables 2 and 3, UniMate outperforms prior methods on most VBench metrics and achieves the highest scores across all user-study criteria, indicating stronger text–motion alignment and more expressive, plausible, and coherent motion; AnimateAnyMesh attains higher overall consistency mainly due to near-static outputs with much lower dynamic degree and expressiveness, while V2M4 relies on separately generated driving videos and is substantially slower, making it less practical than our feed-forward approach.

5.4 More Applications

Text-Mediated Motion Transfer. UniMate naturally supports cross-topology motion transfer: a source motion is first abstracted into a language prompt, which is then used to animate a target rig of different topology. Since generation is *directly conditioned on the target skeleton*, no joint correspondence, exemplar alignment, or per-skeleton optimization is required. Additional qualitative cross-topology transfer results are provided in the supplemental material.

Motion In-Betweening. UniMate further supports zero-shot motion in-betweening, as shown in Fig. 9. Given a target skeleton, a text prompt, and prescribed start and end poses, we perform sampling-time pose guidance by keeping the boundary-frame pose tokens fixed during the flow integration, while applying standard classifier-free text guidance to enforce the desired motion semantics. This produces coherent transitions that satisfy the endpoint constraints while allowing language to specify the style and intent of the transition, such as walking, jumping, or turning.

Motion Expansion. UniMate can extend an animation by chaining a sequence of text prompts, as shown in Fig. 11. For each new prompt, we generate the next motion segment using the preceding segment’s final pose as a boundary condition, producing a longer sequence with smooth temporal transitions while allowing its semantic content to evolve over time.

Text-Guided Motion Editing. Fig. 10 showcases the application of UniMate for text-guided motion editing. Starting from an existing animation, we keep the unchanged parts fixed and resample selected joints under a new text prompt. This enables localized edits such as changing action intensity, modifying limb behavior, or altering motion direction without regenerating the entire sequence. Because the model jointly reasons over motion and skeletal topology, the

edited results remain temporally smooth and structurally consistent with the input rig.

5.5 Ablation Study

In Tab. 4, we ablate the three topology-aware components of TADiT. Removing graph attention bias increases FID, showing the value of explicit structural relations in joint attention. Removing Spec-RoPE further degrades both fidelity and diversity, indicating weaker generalization to unseen

skeletons. Removing the global topological conditioner causes the largest drop in motion quality; although diversity increases, the generated motions become less stable and often exhibit jitter.

Table 4. **Ablation.** Removing topology-aware mechanisms degrades performance.

Method	FID↓	Div.↑
w/o graph attn bias	0.773	8.799
w/o spectral RoPE	0.798	8.424
w/o global topo.	0.825	9.475
Full Model	0.757	9.200

6 Limitations and Future Work

Contact and foot sliding. UniMate can produce foot sliding, drift, hovering, or ground penetration in contact-rich motions. This limitation arises because the model does not impose a unified contact model: foot–ground contact is meaningful for bipeds and quadrupeds, but ill-defined for snakes, swimming fish, birds in flight, and many articulated objects. Future work could introduce morphology-aware contact states and objectives during constraint-guided sampling; where contacts are well defined, foot locking or inverse-kinematics post-processing can also be applied directly to the predicted rig.

Rare topologies and motions. UniMate is less reliable on rare skeletal topologies and out-of-distribution motions, where results can become static, jittery, or semantically inaccurate. This limitation stems from the scarcity and long-tailed distribution of available 4D animation data, which favor humanoids and common locomotion. Future work could distill Internet-scale video or video-generative priors, using video-derived motion cues as additional supervision to broaden both topology and motion coverage while retaining UniMate’s topology-aware backbone.

References

- Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. 2020. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (ToG)* 39, 4 (2020). doi:10.1145/3386569.3392462
- Adobe. 2022. Mixamo. <https://www.mixamo.com/>.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibong Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631.
- Hongyuan Chen, Xingyu Chen, Youjia Zhang, Zexiang Xu, and Anpei Chen. 2026. Motion 3-to-4: 3D Motion Reconstruction for 4D Synthesis. arXiv preprint arXiv:2601.14253.
- Jianqi Chen, Biao Zhang, Xiangjun Tang, and Peter Wonka. 2025a. V2M4: 4D Mesh Animation Reconstruction from a Single Monocular Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11643–11653.
- Ling-Hao Chen, Yuhong Zhang, Zixin Yin, Zhiyang Dou, Xin Chen, Jingbo Wang, Taku Komura, and Lei Zhang. 2025b. Motion2Motion: Cross-topology Motion Transfer

- with Sparse Correspondence. In *SIGGRAPH Asia 2025 Conference Papers*. Association for Computing Machinery, 1–11. doi:10.1145/3757377.3763811
- Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. 2024. Taming Diffusion Probabilistic Models for Character Control. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 1–10. doi:10.1145/3641519.3657440
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53.
- Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023a. Objaverse-XL: A Universe of 10M+ 3D Objects. In *Advances in Neural Information Processing Systems*. 35799–35813.
- Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023b. Objaverse: A universe of annotated 3D objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13142–13153.
- Zhiyang Dou, Chen Peng, Xinyu Lu, Xiaohan Ye, Lixing Fang, Yuan Liu, Wenping Wang, Chuang Gan, Lingjie Liu, and Taku Komura. 2025. CFC: Simulating Character-Fluid Coupling using a Two-Level World Model. *ACM Transactions on Graphics (TOG)* 44, 6 (2025), 1–17.
- Inbar Gat, Sigal Raab, Guy Tevet, Yuval Reshef, Amit Haim Bermano, and Daniel Cohen-Or. 2025. AnyTop: Character Animation Diffusion with Any Topology. In *SIGGRAPH 2025 Conference Papers*. Association for Computing Machinery, 1–10. doi:10.1145/3721238.3730621
- Kehong Gong, Zhengyu Wen, Weixia He, Mingxi Xu, Qi Wang, Ning Zhang, Zhengyu Li, Dongze Lian, Wei Zhao, Xiaoyu He, et al. 2025. MoCapAnything: Unified 3D Motion Capture for Arbitrary Skeletons from Monocular Videos. arXiv preprint arXiv:2512.10881.
- Niv Granot, Ben Feinstein, Assaf Shocher, Shai Bagon, and Michal Irani. 2022. Drop the GAN: In Defense of Patches Nearest Neighbors as Single Image Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13460–13469.
- Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions from Text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5152–5161.
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598.
- Zehuan Huang, Haoran Feng, Yang-Tian Sun, Yuan-Chen Guo, Yan-Pei Cao, and Lu Sheng. 2025. AnimaX: Animating the Inanimate in 3D with Joint Video-Pose Diffusion Models. In *SIGGRAPH Asia 2025 Conference Papers*. Association for Computing Machinery, 1–13. doi:10.1145/3757377.3763885
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.
- Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. 2024. Animate3D: Animating Any 3D Model with Multi-view Video Diffusion. In *Advances in Neural Information Processing Systems*, Vol. 37. 125879–125906. doi:10.52202/079017-3999
- Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 2023. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2151–2162.
- Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O’Sullivan. 2007. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*. 39–46.
- Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*. PMLR, 3744–3753.
- Sunmin Lee, Taeho Kang, Jungnam Park, Jehee Lee, and Jungdam Won. 2023. Same: Skeleton-agnostic motion embedding for character animation. In *SIGGRAPH Asia 2023 Conference Papers*. Association for Computing Machinery, 1–11. doi:10.1145/3610548.3618206
- Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. 2024c. Instant3D: Fast Text-to-3D with Sparse-View Generation and Large Reconstruction Model. In *The Twelfth International Conference on Learning Representations*.
- Peizhuo Li, Kfir Aberman, Rana Hanocka, Libin Liu, Olga Sorkine-Hornung, and Baoquan Chen. 2021. Learning skeletal articulations with neural blend shapes. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–15. doi:10.1145/3450626.3459852
- Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. 2022. Ganimator: Neural motion synthesis from a single sequence. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–12. doi:10.1145/3528223.3530157
- Peizhuo Li, Sebastian Starke, Yuting Ye, and Olga Sorkine-Hornung. 2024b. Walkthedog: Cross-morphology motion alignment via phase manifolds. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 1–10. doi:10.1145/3641519.3657508
- Tianyu Li, Calvin Qiao, Guanqiao Ren, KangKang Yin, and Sehoon Ha. 2024a. AAMD: accelerated auto-regressive motion diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1813–1823.
- Weiye Li, Xuelin Chen, Peizhuo Li, Olga Sorkine-Hornung, and Baoquan Chen. 2023. Example-based motion synthesis via generative motion matching. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–12. doi:10.1145/3592395
- Xuan Li, Qianli Ma, Tsung-Yi Lin, Yongxin Chen, Chenfanfu Jiang, Ming-Yu Liu, and Donglai Xiang. 2025. Articulated kinematics distillation from video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 17571–17581.
- Hanwen Liang, Yuyang Yin, Dejie Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. 2024. Diffusion4D: Fast Spatial-temporal Consistent 4D Generation via Video Diffusion Models. In *Advances in Neural Information Processing Systems*, Vol. 37. 110854–110875. doi:10.52202/079017-3519
- Derek Lim, Joshua David Robinson, Lingxiao Zhao, Tess Smidt, Suvrit Sra, Haggai Maron, and Stefanie Jegelka. 2023. Sign and basis invariant networks for spectral graph representation learning. In *International Conference on Learning Representations*.
- Isabella Liu, Zhan Xu, Wang Yifan, Hao Tan, Zexiang Xu, Xiaolong Wang, Hao Su, and Zifan Shi. 2025. Riganything: Template-free autoregressive rigging for diverse 3d assets. *ACM Transactions on Graphics (TOG)* 44, 4 (2025), 1–12. doi:10.1145/3731149
- Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varna T, Zexiang Xu, and Hao Su. 2023b. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. In *Advances in Neural Information Processing Systems*, Vol. 36. 22226–22246.
- Ruoshi Liu, Rundui Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023a. Zero-1-to-3: Zero-shot One Image to 3D Object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9298–9309.
- Siqi Liu, Maoyu Wang, Bo Dai, and Cewu Lu. 2026. PALUM: Part-based Attention Learning for Unified Motion Retargeting. arXiv preprint arXiv:2601.07272.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003.
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.
- Yanzhe Lyu, Chen Geng, Karthik Dharmarajan, Yunzhi Zhang, Hadi Alzayer, Shangzhe Wu, and Jiajun Wu. 2026. Choreographing a World of Dynamic Objects. arXiv preprint arXiv:2601.04194.
- Nadia Magnenat-Thalmann, Richard Laperrière, and Daniel Thalmann. 1988. Joint-dependent local deformations for hand animation and object grasping. In *Proceedings of Graphics Interface*. 26–33.
- Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5442–5451.
- Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. 2021. Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning. arXiv preprint arXiv:2108.10470.
- Zichong Meng, Zeyu Han, Xiaogang Peng, Yiming Xie, and Huaizu Jiang. 2025. Absolute coordinates make motion generation easy. arXiv preprint arXiv:2505.19377.
- Linshan Mou, Jiahui Lei, Chen Wang, Lingjie Liu, and Kostas Daniilidis. 2025. DIMO: Diverse 3D Motion Generation for Arbitrary Objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14357–14368.
- Alex Nichol, Heewoo Jun, Pratul Dharwal, Pamela Mishkin, and Mark Chen. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. arXiv preprint arXiv:2212.08751.
- Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10975–10985.
- William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.
- Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10985–10995.
- Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The kit motion-language dataset. *Big data* 4, 4 (2016), 236–252.
- Sigal Raab, Inbar Gat, Nathan Sala, Guy Tevet, Rotem Shalev-Arkushin, Ohad Fried, Amit Haim Bermano, and Daniel Cohen-Or. 2024a. Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer. In *SIGGRAPH Asia 2024 Conference Papers*. Association for Computing Machinery, 1–13. doi:10.1145/3680528.3687579

- Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. 2023. Modi: Unconditional motion synthesis from diverse data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13873–13883.
- Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. 2024b. Single motion diffusion. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=DrhZneqz4n>
- Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, et al. 2026. Kimodo: Scaling Controllable Human Motion Generation. arXiv preprint arXiv:2603.15546.
- Remy Sabathier, David Novotny, Niloy J Mitra, and Tom Monnier. 2026. Action-Mesh: Animated 3D Mesh Generation with Temporal 3D Diffusion. arXiv preprint arXiv:2601.16148.
- Haim Sawdayee, Chuan Guo, Guy Tevet, Bing Zhou, Jian Wang, and Amit H Bermano. 2026. Dance like a chicken: Low-rank stylization for human motion diffusion. *Computer Graphics Forum* (2026), e70365. doi:10.1111/cgf.70365
- Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. 2024. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=dTpbEdN9kr>
- Yahao Shi, Yang Liu, Yanmin Wu, Xing Liu, Chen Zhao, Jie Luo, and Bin Zhou. 2025. Drive any mesh: 4d latent diffusion for mesh deformation from video. arXiv preprint arXiv:2506.07489.
- Aaditya Singh et al. 2025. OpenAI GPT-5 System Card. arXiv preprint arXiv:2601.03267. doi:10.48550/arXiv.2601.03267
- Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11050–11059.
- Chaoyue Song, Xiu Li, Fan Yang, Zhongcong Xu, Jiacheng Wei, Fayao Liu, Jiashi Feng, Guosheng Lin, and Jianfeng Zhang. 2025a. Puppeteer: Rig and Animate Your 3D Models. arXiv preprint arXiv:2508.10898.
- Chaoyue Song, Jianfeng Zhang, Xiu Li, Fan Yang, Yiwen Chen, Zhongcong Xu, Jun Hao Liew, Xiaoyang Guo, Fayao Liu, Jiashi Feng, et al. 2025b. Magicarticulate: Make your 3d models articulation-ready. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 15998–16007.
- Truebones Motions Animation Studios. 2022. Truebones. <https://truebones.gumroad.com/>.
- Jianlin Sun, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.
- Keqiang Sun, Dor Litvak, Yunzhi Zhang, Hongsheng Li, Jiajun Wu, and Shangzhe Wu. 2024. Ponymation: Learning Articulated 3D Animal Motions from Unlabeled Online Videos. In *European Conference on Computer Vision*. Springer, 100–119.
- Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H Bermano, and Michiel van de Panne. 2024. Cload: Closing the loop between simulation and diffusion for multi-task character control. arXiv preprint arXiv:2410.03441.
- Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2023. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=SJ1kSyO2jwu>
- Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 5026–5033.
- Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. 2025. MotionDreamer: Exploring Semantic Video Diffusion features for Zero-Shot 3D Mesh Animation. In *International Conference on 3D Vision 2025*.
- Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. 2024. Tlcontrol: Trajectory and language control for human motion synthesis. In *European Conference on Computer Vision*. Springer, 37–54.
- Wan Team, Alibaba Group. 2025. Wan: Open and Advanced Large-Scale Video Generation Models. arXiv preprint arXiv:2503.20314.
- Miaowei Wang, Qingxuan Yan, Zhi Cao, Yayuan Li, Oisín Mac Aodha, Jason J Corso, and Amir Vaxman. 2026b. BiMotion: B-spline Motion for Text-guided Dynamic 3D Character Generation. arXiv preprint arXiv:2602.18873.
- Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. 2023. RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4563–4573.
- Xuan Wang, Kai Ruan, Liyang Qian, Zhizhi Guo, Chang Su, and Gaoang Wang. 2026a. X-MoGen: Unified Motion Generation Across Humans and Animals. *Proceedings of the AAAI Conference on Artificial Intelligence* 40, 12 (2026), 10234–10242. doi:10.1609/aaai.v40i12.37992
- Xuan Wang, Kai Ruan, Xing Zhang, and Gaoang Wang. 2025. Animo: Species-aware model for text-driven animal motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1929–1939.
- Yuxin Wen, Qing Shuai, Di Kang, Jing Li, Cheng Wen, Yue Qian, Ningxin Jiao, Changhai Chen, Weijie Chen, Yiran Wang, et al. 2025. HY-Motion 1.0: Scaling Flow Matching Models for Text-To-Motion Generation. arXiv preprint arXiv:2512.23464.
- Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. 2023. MagicPony: Learning Articulated 3D Animals in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8792–8802.
- Zijie Wu, Chaohui Yu, Fan Wang, and Xiang Bai. 2025. AnimateAnyMesh: A Feed-Forward 4D Foundation Model for Text-Driven Universal Mesh Animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13557–13568.
- Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2024. Structured 3D Latents for Scalable and Versatile 3D Generation. arXiv preprint arXiv:2412.01506.
- Tianyi Xie, Yunuo Chen, Yaowei Guo, Yin Yang, Bolei Zhou, Demetri Terzopoulos, Ying Jiang, and Chenfanfu Jiang. 2025. AnimaMimic: Imitating 3D Animation from Video Priors. arXiv preprint arXiv:2512.14133.
- Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. 2020. Rignet: Neural rigging for articulated characters. *ACM Transactions on Graphics* 39, 4 (2020), 1–14. doi:10.1145/3386569.3392379
- Zhangsihao Yang, Mingyuan Zhou, Mengyi Shan, Bingbing Wen, Ziwei Xuan, Mitch Hill, Junjie Bai, Guo-Jun Qi, and Yalin Wang. 2024. Omnimotiongpt: Animal motion generation with limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1249–1259.
- Jiraphon Yenphraphai, Ashkan Mirzaei, Jianqi Chen, Jiaxu Zou, Sergey Tulyakov, Raymond A Yeh, Peter Wonka, and Chaoyang Wang. 2025. Shapegen4d: Towards high quality 4d shape generation from videos. arXiv preprint arXiv:2510.06208.
- Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems* 34 (2021), 28877–28888.
- Kwan Yun, Seokhyeon Hong, Chaelin Kim, and Junyong Noh. 2025. AnyMoLe: Any Character Motion In-betweening Leveraging Video Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 27838–27848.
- Bowen Zhang, Sicheng Xu, Chuxin Wang, Jiaolong Yang, Feng Zhao, Dong Chen, and Baining Guo. 2025c. Gaussian variation field diffusion for high-fidelity video-to-4d synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12502–12513.
- Jia-Peng Zhang, Cheng-Feng Pu, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. 2025b. One model to rig them all: Diverse skeleton rigging with unirig. *ACM Transactions on Graphics (TOG)* 44, 4 (2025), 1–18. doi:10.1145/3730930
- Xinyi Zhang, Naiqi Li, and Angela Dai. 2025a. DNF: Unconditional 4D Generation with Dictionary-based Neural Fields. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26047–26056.
- Kaifeng Zhao, Gen Li, and Siyu Tang. 2024a. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. arXiv preprint arXiv:2410.05260.
- Qingqing Zhao, Peizhuo Li, Yifan Wang, Olga Sorkine-Hornung, and Gordon Wetstein. 2024b. Pose-to-Motion: Cross-Domain Motion Retargeting with Pose Prior. *Computer Graphics Forum* 43, 8 (2024), e15170. doi:10.1111/cgf.15170
- Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. 2025. Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation. arXiv preprint arXiv:2501.12202.
- Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. 2024. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*. Springer, 18–38.
- Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5745–5753.
- Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 2017. 3D Menagerie: Modeling the 3D shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6365–6373.