Novelty Assessment Report

Paper: MTVCraft: Tokenizing 4D Motion for Arbitrary Character Animation
PDF URL: https://openreview.net/pdf?id=m7AQM9H6wa
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Character image animation has rapidly advanced with the rise of digital humans. However, existing methods rely largely on 2D-rendered pose images for motion guidance, which limits generalization and discards essential 4D information for open-world animation. To address this, we propose MTVCraft (Motion Tokenization Video Crafter), the first framework that directly models raw 3D motion sequences (i.e., 4D motion) for character image animation. Specifically, we introduce 4DMoT (4D motion tokenizer) to quantize 3D motion sequences into 4D motion tokens. Compared to 2D-rendered pose images, 4D motion tokens offer more robust spatial-temporal cues and avoid strict pixel-level alignment between pose images and the character, enabling more flexible and disentangled control. Next, we introduce MV-DiT (Motion-aware Video DiT). By designing unique motion attention with 4D positional encodings, MV-DiT can effectively leverage motion tokens as 4D compact yet expressive context for character image animation in the complex 4D world. We implement MTVCraft on both CogVideoX-5B (small scale) and Wan-2.1-14B (large scale), demonstrating that our framework is easily scalable and can be applied to models of varying sizes. Experiments on the TikTok and Fashion benchmarks demonstrate our state-of-the-art performance. Moreover, powered by robust motion tokens, MTVCraft showcases unparalleled zero-shot generalization. It can animate arbitrary characters in both single and multiple settings, in full-body and half-body forms, and even non-human objects across diverse styles and scenarios. Hence, it marks a significant step forward in this field and opens a new direction for pose-guided video generation.

Core Task Landscape

This paper addresses: **Character Image Animation Using 3D Motion Sequences**

A total of **50 papers** were analyzed and organized into a taxonomy with **27 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Motion Representation and Guidance**
- **Generative Models and Synthesis Frameworks**
- **Motion Capture and Data Acquisition**
- **Animation Editing and Control**
- **Data-Driven Motion Synthesis and Retrieval**
- **Application Domains and Specialized Scenarios**
- **Technical Foundations and Enabling Technologies**
- **Surveys and Comprehensive Reviews**

Complete Taxonomy Tree

- Character Image Animation Using 3D Motion Sequences Survey Taxonomy
- Motion Representation and Guidance
  - 3D Parametric Model-Based Guidance (4 papers)
  - [1] Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance (Shenhao Zhu, 2024) View paper
  - [2] Animate-X: Universal Character Image Animation with Enhanced Motion Representation (Tan Shuai, 2024) View paper
  - [4] CharacterShot: Controllable and Consistent 4D Character Animation (Gao Junyao, 2025) View paper
  - [16] Towards High-Quality 3D Motion Transfer with Realistic Apparel Animation (Rong Wang, 2024) View paper
  - Motion Tokenization and Discrete Representation ★ (2 papers)
  - [0] MTVCraft: Tokenizing 4D Motion for Arbitrary Character Animation (Anon et al., 2026) View paper
  - [44] MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation (Ding Yanbo, 2025) View paper
  - 2D Pose-Based Control (2 papers)
  - [30] RealisDance: Equip controllable character animation with realistic hands (Zhou, 2024) View paper
  - [33] AnyI2V: Animating Any Conditional Image with Motion Control (Li ZiYe, 2025) View paper
  - Mixed and Multi-Modal Motion Dynamics (3 papers)
  - [26] MikuDance: Animating Character Art with Mixed Motion Dynamics (Zhang Jiaxu, 2024) View paper
  - [34] HumanVid: Demystifying Training Data for Camera-controllable Human Image Animation (Wang Zhenzhi, 2024) View paper
  - [35] AnimateAnywhere: Rouse the Background in Human Image Animation (Liu Xiaoyu, 2025) View paper
- Generative Models and Synthesis Frameworks
  - Diffusion-Based Video Generation (2 papers)
  - [8] Hallo3: Highly Dynamic and Realistic Portrait Image Animation with Video Diffusion Transformer (Jiahao Cui, 2024) View paper
  - [17] Animate3D: Animating Any 3D Model with Multi-view Video Diffusion (Chenjie Cao, 2024) View paper
  - Audio-Driven Facial and Portrait Animation (2 papers)
  - [19] JoyVASA: Portrait and Animal Image Animation with Diffusion-Based Audio-Driven Facial Dynamics and Head Motion Generation (Cao Xuyang, 2024) View paper

    - [29] CSTalk: Correlation Supervised Speech-driven 3D Emotional Facial Animation Generation (Xiangyu Liang, 2024) View paper
    - Text-Conditioned Motion Generation (2 papers)
    - [7] Make-an-animation: Large-scale text-conditional 3d human motion generation (Samaneh Azadi, 2023) View paper
    - [13] Multi-Track Timeline Control for Text-Driven 3D Human Motion Generation (Mathis Petrovich, 2024) View paper
    - Other Generative Architectures (1 papers)
    - [5] 3D Character Animation and Asset Generation Using Deep Learning (Vlad-Constantin Lungu-Stan, 2024) View paper
- Motion Capture and Data Acquisition
    - Markerless Video-Based Motion Capture (4 papers)
    - [11] Enhancing 3D Avatar Realism with Indian Clothing and Motion Capture Data (Anushka Jalori, 2024) View paper
    - [24] A Technical Demonstration on Streamlining 3D Motion Production Workflows with a Markerless Motion Capture System (Mengyao Guo, 2024) View paper
    - [40] The use of motion capture technology in 3D animation (Mars Caroline Wibowo, 2024) View paper
    - [45] Markerless Body Motion Capturing for 3D Character Animation based on Multi-view Cameras (Wang Jin-Bao, 2022) View paper
    - Sensor-Based Motion Capture (1 papers)
    - [28] Realtime performance animation using sparse 3D motion sensors (Jongmin Kim, 2012) View paper
    - Motion Capture Data Processing and Retargeting (2 papers)
    - [14] A Study on the Application of Motion Capture Technology and Animation Generation Methods in Film and Television Performances (Zhao, 2025) View paper
    - [31] From motion capture data to character animation (Gaojin Wen, 2006) View paper
    - Video-Based Motion Extraction for Animation Authoring (2 papers)
    - [22] VideoPoseVR: Authoring virtual reality character animations with online videos (Cheng-yao Wang, 2022) View paper
    - [38] Video-based character animation (Starck, 2005) View paper
- Animation Editing and Control
    - Interactive and Immersive Editing Interfaces (2 papers)
    - [6] VidAnimator: User-Guided Stylized 3D Character Animation from Human Videos (Mei Shuhong, 2025) View paper
    - [21] TimeTunnel: Integrating Spatial and Temporal Motion Editing for Character Animation in Virtual Reality (Qian Zhou, 2024) View paper
    - Inverse Kinematics-Based Editing (1 papers)
    - [18] Modification of Skeletal Character Animation Using Inverse Kinematics Controllers (Gazizov, 2024) View paper
    - Physics-Based Motion Synthesis and Transition (2 papers)
    - [20] PIMT: Physics-Based Interactive Motion Transition for Hybrid Character Animation (Yanbin Deng, 2024) View paper
    - [27] Synthesis of complex dynamic character motion from simple animations (C. Karen Liu, 2002) View paper
    - Temporal Consistency and Constraint Enforcement (1 papers)
    - [43] Imposing temporal consistency on deep monocular body shape and pose estimation (Alexandra Zimmer, 2022) View paper
- Data-Driven Motion Synthesis and Retrieval
    - Motion Matching and Graph-Based Retrieval (1 papers)
    - [39] Motion Matching for Character Animation and Virtual Reality Avatars in Unity (Ponton, 2023) View paper
    - Motion Grammars and Structured Synthesis (1 papers)
    - [10] Motion grammars for character animation (Kyunglyul Hyun, 2016) View paper
    - Deep Learning-Based Motion Synthesis (1 papers)
    - [12] A survey on deep learning for skeletonâ□based human animation (Lucas Mourot, 2022) View paper
- Application Domains and Specialized Scenarios
    - Stylized and Non-Human Character Animation (2 papers)
    - [9] Drawingspinup: 3d animation from single character drawings (Jie Zhou, 2024) View paper
    - [25] Character animation from 2D pictures and 3D motion data (Alexander Hornung, 2007) View paper
    - Digital Twins and Metaverse Avatars (1 papers)
    - [23] The Design of a 3D Character Animation System for Digital Twins in the Metaverse (Tanberk, 2024) View paper
    - Film and Game Production Workflows (4 papers)
    - [3] Research on 3D animation character design based on multimedia interaction (Ping Hu, 2025) View paper
    - [37] Making a game character move: Animation and motion capture for video games (Brusi, 2021) View paper
    - [48] Developing Techniques for Rigging and Motion Capture to Simplify 3D Animation (Angela Humphrey, 2022) View paper
    - [49] Designing a Rabbit Character's Motion While Waiting in 3D Animation Short "Yue Bing" (Raynaldo Oscar Tandiary, 2020) View paper
- Technical Foundations and Enabling Technologies
    - 3D Modeling and Rigging Techniques (3 papers)
    - [32] Application of Skeletal Skinned Mesh Algorithm Based on 3D Virtual Human Model in Computer Animation Design (Zhan, 2024) View paper
    - [36] Animation Image Art Design Mode Using 3D Modeling Technology (Tan, 2022) View paper
    - [41] Cutting-edge techniques in 3D modeling and animation: Leveraging mathematical models and advanced software tools (Zhang, 2024) View paper
    - Facial Animation and Expression Control (1 papers)
    - [50] Vision-based control of 3 D facial animation (D Breen, 2003) View paper
    - Multi-Character and Multi-Motion Coordination (1 papers)
    - [42] Semantic-Driven Multi-character Multi-motion 3D Animation Generation (Hui Liang, 2025) View paper
    - Multimedia Interaction and Digital Media Arts (1 papers)
    - [46] The Use of Digital Media Arts in Animation Filmmaking in the Age of Digital Intelligence (Tang, 2024) View paper
- Surveys and Comprehensive Reviews (2 papers)
    - [15] Generative AI for Character Animation: A Comprehensive Survey of Techniques, Applications, and Future Directions (Abootorabi Mohammad Mahdi, 2025) View paper
    - [47] Learning-based 3D human motion capture and animation synthesis (Habibie, 2023) View paper

Narrative

Core task: Character image animation using 3D motion sequences. The field organizes itself around several complementary dimensions. Motion Representation and Guidance explores how to encode and condition animation on 3D pose or skeletal data, including discrete tokenization schemes that compress motion into learnable vocabularies. Generative Models and Synthesis Frameworks focuses on diffusion-based and neural rendering pipelines that translate motion signals into realistic character videos. Motion Capture and Data Acquisition addresses the upstream problem of obtaining high-quality 3D motion, whether through marker-based systems or markerless vision approaches. Animation Editing and Control provides tools for temporal refinement and user-driven adjustments, while Data-Driven Motion Synthesis and Retrieval leverages large motion databases for retrieval and blending. Application Domains span talking heads, full-body dance, and game characters, and Technical Foundations covers enabling components like inverse kinematics and skeletal rigging. Surveys and Comprehensive Reviews tie these threads together, offering periodic snapshots of progress across methods such as Champ[1], Animate-X[2], and CharacterShot[4].

Within Motion Representation, a particularly active line of work examines motion tokenization and discrete representation, seeking compact codes that generative models can consume efficiently. MTVCraft[0] sits squarely in this branch, proposing a tokenization strategy that bridges 3D motion sequences and video synthesis. Its close neighbor MTVCrafter[44] shares a similar emphasis on discrete motion vocabularies, suggesting a small cluster dedicated to learned motion codebooks. By contrast, works like Champ[1] and Animate-X[2] often rely on continuous pose embeddings or direct skeletal conditioning, trading off the compactness of tokens for more direct geometric control. Meanwhile, methods such as VidAnimator[6] and Make-an-animation[7] explore end-to-end diffusion pipelines that may bypass explicit tokenization altogether. The central tension revolves around whether discrete motion codes offer better generalization and editability, or whether continuous representations preserve finer kinematic detail. MTVCraft[0] thus contributes to an emerging conversation on how best to distill motion into a form that balances expressiveness, efficiency, and compatibility with modern generative architectures.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation

Authors: Ding Yanbo, Guo, Zhizhi, Zhang Chi, Wang YaLi | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Human image animation has gained increasing attention and developed rapidly due to its broad applications in digital humans. However, existing methods rely largely on 2D-rendered pose images for motion guidance, which limits generalization and discards essential 3D information for open-world animation. To tackle this problem, we propose MTVCrafter (Motion Tokenization Video Crafter), the first framework that directly models raw 3D motion sequences (i.e., 4D motion) for human image animation. Spe...

⚠ Similarity Notice

These papers share nearly identical titles (MTVCraft vs. MTVCrafter), describe the same core technical approach (4D motion tokenization using 4DMoT and MV-DiT), and present the same experimental results (FID-VID of 6.98). They appear to be the same work or very close variants, likely representing different submission versions of the same research.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: 4D Motion Tokens for character animation

Description: The authors introduce a new motion representation called 4D Motion Tokens that discretizes spatial and temporal motion information. This representation aims to enable more effective character animation by capturing motion dynamics in a tokenized format.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Motionverse: A unified multimodal framework for motion comprehension, generation and editing

URL: View paper

Brief Assessment

Motionverse[56] focuses on multi-stream motion tokenization for motion comprehension and generation tasks using residual vector quantization, not specifically on 4D spatial-temporal tokens for character animation as in the original paper.

---

2. A Unified Framework for Multimodal, Multi-Part Human Motion Synthesis

URL: View paper

Brief Assessment

Unified Motion Synthesis[64] focuses on quantizing motions of diverse body parts into separate codebooks for multimodal control (text, music, speech), while the original paper introduces 4D Motion Tokens that discretize spatial and temporal motion information from SMPL joint coordinates specifically for video-driven character animation. The technical approaches and application domains differ substantially.

---

3. Causal Motion Tokenizer for Streaming Motion Generation

URL: View paper

Brief Assessment

Causal Motion Tokenizer[61] focuses on streaming motion generation from text using causal convolutions for temporal sequences, not on spatial-temporal 4D motion representation for character animation as in the original paper.

---

4. DisCoRD: Discrete Tokens to Continuous Motion via Rectified Flow Decoding

URL: View paper

Brief Assessment

DisCoRD[71] focuses on decoding discrete motion tokens into continuous motion using rectified flow for human motion generation tasks (text-to-motion, gesture, dance). The original paper introduces 4D Motion Tokens that discretize spatial-temporal motion for character image animation. These are fundamentally different applications and technical approaches.

---

5. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts

URL: View paper

Brief Assessment

Tm2t[68] focuses on discrete motion tokens for text-to-motion and motion-to-text generation tasks, not character image animation. The tokenization approach in Tm2t[68] uses vector quantization on 3D pose sequences for bidirectional text-motion translation, which differs

from the original paper's 4D motion tokenization (spatial-temporal joint coordinates) specifically designed for video generation and character animation control.

---

6. Generating human motion from textual descriptions with discrete representations

URL: View paper

Brief Assessment

Discrete Text Motion[69] focuses on text-to-motion generation using VQ-VAE to tokenize motion sequences for GPT-based generation, not on character image animation. The candidate does not address video generation or character animation tasks that the original paper targets.

---

7. A Self-supervised Motion Representation for Portrait Video Generation

URL: View paper

Brief Assessment

Self-supervised Portrait[70] focuses on portrait video generation using semantic latent motion (1D tokens) for audio-driven facial animation, not general character animation with 4D motion tokens encoding spatial-temporal joint coordinates.

---

8. HiTVideo: Hierarchical Tokenizers for Enhancing Text-to-Video Generation with Autoregressive Large Language Models

URL: View paper

Brief Assessment

HiTVideo[73] focuses on hierarchical video tokenization for text-to-video generation using 3D causal VAE with multi-layer discrete tokens for semantic and spatiotemporal encoding. The original paper's 4D Motion Tokens specifically discretize SMPL joint coordinates for character animation control, which is a fundamentally different application and tokenization approach than HiTVideo's general video content encoding.

---

9. Taming Diffusion Probabilistic Models for Character Control

URL: View paper

Brief Assessment

Taming Diffusion Control[72] focuses on real-time character control using diffusion models with autoregressive motion generation, not on discrete token representations encoding spatial and temporal motion information for animation.

---

10. Mogents: Motion generation based on spatial-temporal joint modeling

URL: View paper

Brief Assessment

Mogents[51] focuses on motion generation from text prompts using spatial-temporal joint modeling, not character image animation. The candidate quantizes motion for generation tasks, while the original paper addresses video-based character animation with pose transfer.

Contribution 2: MTVCraft framework for arbitrary character animation

Description: The authors develop MTVCraft, a unified framework that uses the proposed 4D Motion Tokens to perform character animation for arbitrary characters. The framework is designed to handle diverse animation scenarios using the tokenized motion representation.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Motionverse: A unified multimodal framework for motion comprehension, generation and editing

URL: View paper

Brief Assessment

Motionverse[56] presents a unified framework for motion comprehension, generation, and editing using LLMs, not a character animation framework like MTVCraft that transfers poses to arbitrary characters.

---

2. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization

URL: View paper

Brief Assessment

Tokenhsi[62] focuses on physics-based human-scene interaction control using reinforcement learning, not character image animation from videos using diffusion models. The technical approaches are fundamentally different.

---

3. A Unified Framework for Multimodal, Multi-Part Human Motion Synthesis

URL: View paper

Brief Assessment

Unified Motion Synthesis[64] presents a framework for multimodal motion generation using token prediction from specialized codebooks, whereas MTVCraft is specifically designed for video-driven character animation using 4D motion tokens with motion-aware video diffusion transformers. The control modalities and architectural designs are fundamentally different.

---

4. MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation

URL: View paper

Brief Assessment

MTVCrafter[44] focuses on human image animation using 4D motion tokenization, while the original paper's MTVCraft addresses arbitrary character animation including non-human subjects. The candidate's scope is narrower, limited to human characters.

---

5. Causal Motion Tokenizer for Streaming Motion Generation

URL: View paper

Brief Assessment

Causal Motion Tokenizer[61] presents MotionStream for text-to-motion generation in streaming contexts, not a framework for arbitrary character image animation like MTVCraft.

---

6. Versatile multimodal controls for expressive talking human animation

URL: View paper

Brief Assessment

Versatile Multimodal Controls[63] focuses on talking human animation with audio-driven motion and text-controlled gestures, not on arbitrary character animation using tokenized 4D motion representations for diverse character types.

---

7. Moconvq: Unified physics-based motion control via scalable discrete representations

URL: View paper

Brief Assessment

Moconvq[60] focuses on physics-based motion control for simulated characters using reinforcement learning and discrete motion representations, not on arbitrary character image animation from reference images and driving videos as in the original paper.

---

8. VersatileMotion: A Unified Framework for Motion Synthesis and Comprehension

URL: View paper

Brief Assessment

VersatileMotion[67] focuses on human motion generation and understanding across multiple modalities (text, music, speech) using motion tokenization, but does not address character image animation or video generation tasks that MTVCraft targets.

---

9. Dynamic Motion Synthesis: Masked Audio-Text Conditioned Spatio-Temporal Transformers

URL: View paper

Brief Assessment

Dynamic Motion Synthesis[66] focuses on whole-body motion generation from text and audio using VQVAEs and masked language modeling, not on character image animation with 4D motion tokens as in the original paper.

---

10. ParCo: Part-Coordinating Text-to-Motion Synthesis

URL: View paper

Brief Assessment

ParCo[65] focuses on text-to-motion synthesis for generating coordinated human body part motions from text descriptions, not on character image animation or video generation from reference images and driving videos.

Contribution 3: Motion tokenization approach combining spatial and temporal dimensions

Description: The authors present a core methodological insight of tokenizing motion across both spatial and temporal dimensions simultaneously, rather than treating them separately. This approach forms the foundation of their 4D Motion Tokens representation.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Motionverse: A unified multimodal framework for motion comprehension, generation and editing

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

2. How can large language models understand spatial-temporal data?

URL: View paper

Brief Assessment

Spatial-Temporal LLMs[53] focuses on tokenizing spatial-temporal graph data for forecasting tasks using LLMs, not on motion tokenization for character animation. The candidate addresses spatial-temporal dependencies in graph-structured data (traffic, electricity), while the original paper tokenizes 4D human motion (SMPL joint coordinates) for video generation.

---

3. Adversarially-refined vq-gan with dense motion tokenization for spatio-temporal heatmaps

URL: View paper

Brief Assessment

Dense Motion Tokenization[57] focuses on compressing spatio-temporal heatmaps for human pose representation using VQ-GAN, not on general motion tokenization for character animation. The original paper tokenizes 3D joint coordinates across time for arbitrary character animation, while the candidate applies vector quantization to dense heatmap volumes for compression purposes.

---

4. MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation

URL: View paper

Prior Art Analysis

MTVCrafter[44] demonstrates prior work on tokenizing motion across spatial and temporal dimensions simultaneously through their 4D Motion Tokenizer (4DMOT). The candidate paper explicitly describes quantizing 3D motion sequences into 4D motion tokens that capture spatio-temporal information, which directly parallels the original paper's claimed contribution of tokenizing motion across both spatial and temporal dimensions simultaneously. Both papers use the same terminology ('4D motion tokens', '4DMOT') and describe the same fundamental approach of encoding 3D joint coordinates over time into discrete tokens.

Evidence

Evidence 1 - **Rationale**: Both papers describe 4D motion tokens as capturing spatio-temporal information from 3D joint coordinates over time, with the candidate paper demonstrating this concept was already established in prior work. - **Original**: we propose4dmot(4d motiontokenizer) to directly quantize 4d human motion data (i.e., 3d joint coordinates over time). the resulting motion tokens faithfully preserve the information of raw motion - **Candidate**: compared to 2d-rendered pose images, 4d motion tokens offer more robust spatio-temporal cues and avoid strict pixel-level alignment between pose image and character, enabling more flexible and disentangled control.

---

5. LiON-LoRA: Rethinking LoRA Fusion to Unify Controllable Spatial and Temporal Generation for Video Diffusion

URL: View paper

Brief Assessment

LiON-LoRA[52] focuses on LoRA fusion strategies for camera and object motion control in video diffusion models, not on motion tokenization methods. The candidate does not discuss tokenizing motion data into discrete representations combining spatial-temporal dimensions.

---

6. Efficient Long Video Tokenization via Coordinate-based Patch Reconstruction

URL: View paper

Brief Assessment

Coordinate Patch Reconstruction[58] focuses on video tokenization via coordinate-based patch reconstruction from triplane representations, not motion tokenization combining spatial-temporal dimensions for character animation.

---

7. Efficient Temporal Tokenization for Mobility Prediction with Large Language Models

URL: View paper

Brief Assessment

Temporal Mobility Tokenization[54] focuses on mobility prediction by tokenizing daily trajectory segments for capturing temporal patterns in human movement. The ORIGINAL paper tokenizes 4D motion (3D joint coordinates over time) for character animation, which is a fundamentally different domain and representation.

---

8. Diverse Human Motion Prediction Guided by Multi-level Spatial-Temporal Anchors

URL: View paper

Brief Assessment

Multi-level Anchors[59] focuses on human motion prediction using learnable anchors to capture spatial-temporal variation in future motion trajectories, not on tokenizing 4D motion data for character animation as in the original paper.

---

9. The language of motion: Unifying verbal and non-verbal language of 3d human motion

URL: View paper

Brief Assessment

Language of Motion[55] focuses on compositional body motion tokenization (face, hands, upper-body, lower-body) for multimodal language models, not on the simultaneous spatial-temporal tokenization approach described in the original paper's 4D Motion Tokens framework.

---

10. Mogents: Motion generation based on spatial-temporal joint modeling

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

Appendix: Text Similarity Detection

Textual similarity detection checked 24 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation

Detected in: Core Task (sibling), Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] MTVCraft: Tokenizing 4D Motion for Arbitrary Character Animation View paper
- [1] Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance View paper
- [2] Animate-X: Universal Character Image Animation with Enhanced Motion Representation View paper
- [3] Research on 3D animation character design based on multimedia interaction View paper
- [4] CharacterShot: Controllable and Consistent 4D Character Animation View paper
- [5] 3D Character Animation and Asset Generation Using Deep Learning View paper
- [6] VidAnimator: User-Guided Stylized 3D Character Animation from Human Videos View paper
- [7] Make-an-animation: Large-scale text-conditional 3d human motion generation View paper
- [8] Hallo3: Highly Dynamic and Realistic Portrait Image Animation with Video Diffusion Transformer View paper
- [9] Drawingspinup: 3d animation from single character drawings View paper
- [10] Motion grammars for character animation View paper
- [11] Enhancing 3D Avatar Realism with Indian Clothing and Motion Capture Data View paper
- [12] A survey on deep learning for skeletonâ□based human animation View paper
- [13] Multi-Track Timeline Control for Text-Driven 3D Human Motion Generation View paper
- [14] A Study on the Application of Motion Capture Technology and Animation Generation Methods in Film and Television Performances View paper
- [15] Generative AI for Character Animation: A Comprehensive Survey of Techniques, Applications, and Future Directions View paper
- [16] Towards High-Quality 3D Motion Transfer with Realistic Apparel Animation View paper
- [17] Animate3D: Animating Any 3D Model with Multi-view Video Diffusion View paper
- [18] Modification of Skeletal Character Animation Using Inverse Kinematics Controllers View paper
- [19] JoyVASA: Portrait and Animal Image Animation with Diffusion-Based Audio-Driven Facial Dynamics and Head Motion Generation View paper
- [20] PIMT: Physics-Based Interactive Motion Transition for Hybrid Character Animation View paper
- [21] TimeTunnel: Integrating Spatial and Temporal Motion Editing for Character Animation in Virtual Reality View paper
- [22] VideoPoseVR: Authoring virtual reality character animations with online videos View paper
- [23] The Design of a 3D Character Animation System for Digital Twins in the Metaverse View paper
- [24] A Technical Demonstration on Streamlining 3D Motion Production Workflows with a Markerless Motion Capture System View paper
- [25] Character animation from 2D pictures and 3D motion data View paper
- [26] MikuDance: Animating Character Art with Mixed Motion Dynamics View paper

- [27] Synthesis of complex dynamic character motion from simple animations View paper
- [28] Realtime performance animation using sparse 3D motion sensors View paper
- [29] CSTalk: Correlation Supervised Speech-driven 3D Emotional Facial Animation Generation View paper
- [30] RealisDance: Equip controllable character animation with realistic hands View paper
- [31] From motion capture data to character animation View paper
- [32] Application of Skeletal Skinned Mesh Algorithm Based on 3D Virtual Human Model in Computer Animation Design View paper
- [33] AnyI2V: Animating Any Conditional Image with Motion Control View paper
- [34] HumanVid: Demystifying Training Data for Camera-controllable Human Image Animation View paper
- [35] AnimateAnywhere: Rouse the Background in Human Image Animation View paper
- [36] Animation Image Art Design Mode Using 3D Modeling Technology View paper
- [37] Making a game character move: Animation and motion capture for video games View paper
- [38] Video-based character animation View paper
- [39] Motion Matching for Character Animation and Virtual Reality Avatars in Unity View paper
- [40] The use of motion capture technology in 3D animation View paper
- [41] Cutting-edge techniques in 3D modeling and animation: Leveraging mathematical models and advanced software tools View paper
- [42] Semantic-Driven Multi-character Multi-motion 3D Animation Generation View paper
- [43] Imposing temporal consistency on deep monocular body shape and pose estimation View paper
- [44] MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation View paper
- [45] Markerless Body Motion Capturing for 3D Character Animation based on Multi-view Cameras View paper
- [46] The Use of Digital Media Arts in Animation Filmmaking in the Age of Digital Intelligence View paper
- [47] Learning-based 3D human motion capture and animation synthesis View paper
- [48] Developing Techniques for Rigging and Motion Capture to Simplify 3D Animation View paper
- [49] Designing a Rabbit Character's Motion While Waiting in 3D Animation Short "Yue Bing" View paper
- [50] Vision-based control of 3 D facial animation View paper
- [51] Mogents: Motion generation based on spatial-temporal joint modeling View paper
- [52] LiON-LoRA: Rethinking LoRA Fusion to Unify Controllable Spatial and Temporal Generation for Video Diffusion View paper
- [53] How can large language models understand spatial-temporal data? View paper
- [54] Efficient Temporal Tokenization for Mobility Prediction with Large Language Models View paper
- [55] The language of motion: Unifying verbal and non-verbal language of 3d human motion View paper
- [56] Motionverse: A unified multimodal framework for motion comprehension, generation and editing View paper
- [57] Adversarially-refined vq-gan with dense motion tokenization for spatio-temporal heatmaps View paper
- [58] Efficient Long Video Tokenization via Coordinate-based Patch Reconstruction View paper
- [59] Diverse Human Motion Prediction Guided by Multi-level Spatial-Temporal Anchors View paper
- [60] Moconvq: Unified physics-based motion control via scalable discrete representations View paper
- [61] Causal Motion Tokenizer for Streaming Motion Generation View paper
- [62] Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization View paper
- [63] Versatile multimodal controls for expressive talking human animation View paper
- [64] A Unified Framework for Multimodal, Multi-Part Human Motion Synthesis View paper
- [65] ParCo: Part-Coordinating Text-to-Motion Synthesis View paper
- [66] Dynamic Motion Synthesis: Masked Audio-Text Conditioned Spatio-Temporal Transformers View paper
- [67] VersatileMotion: A Unified Framework for Motion Synthesis and Comprehension View paper
- [68] Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts View paper
- [69] Generating human motion from textual descriptions with discrete representations View paper
- [70] A Self-supervised Motion Representation for Portrait Video Generation View paper
- [71] DisCoRD: Discrete Tokens to Continuous Motion via Rectified Flow Decoding View paper
- [72] Taming Diffusion Probabilistic Models for Character Control View paper
- [73] HiTVideo: Hierarchical Tokenizers for Enhancing Text-to-Video Generation with Autoregressive Large Language Models View paper