Novelty Assessment Report

Paper: Instilling an Active Mind in Avatars via Cognitive Simulation
PDF URL: https://openreview.net/pdf?id=80JylHgQn1
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Current video avatar models can generate fluid animations but struggle to capture a character's authentic essence, primarily synchronizing motion with low-level audio cues instead of understanding higher-level semantics like emotion or intent. To bridge this gap, we propose a novel framework for generating character animations that are not only physically plausible but also semantically rich and expressive. Our model is built on two technical innovations. First, we employ Multimodal Large Language Models to generate a structured textual representation from input conditions, providing high-level semantic guidance for creating contextually and emotionally resonant actions. Second, to ensure robust fusion of multimodal signals, we introduce a specialized Multimodal Diffusion Transformer architecture featuring a novel Pseudo Last Frame design. This allows our model to accurately interpret the joint semantics of audio, images and text, generating motions that are deeply coherent with the overall context. Comprehensive experiments validate the superiority of our method, which achieves compelling results in lip-sync accuracy, video quality, motion naturalness, and semantic consistency. The approach also shows strong generalization to challenging scenarios, including multi-person and non-human subjects. Our video results are linked in https://anonymous.4open.science/w/InstillinganActiveMindinAvatars_Anonymous/ .

Core Task Landscape

This paper addresses: **Semantically Coherent Character Animation from Multimodal Inputs**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

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

- **Audio-Driven Facial and Co-Speech Animation**
- **Image-to-Video Character Animation**
- **Text-to-Motion Synthesis and Control**
- **Scene-Aware and Context-Conditioned Motion Synthesis**
- **Long-Form and Narrative-Driven Animation**
- **Motion Style Transfer and Cross-Domain Adaptation**
- **Unified and Multimodal Motion Synthesis Frameworks**
- **Semantic and Conceptual Animation Systems**
- **Surveys, Benchmarks, and Theoretical Foundations**

Complete Taxonomy Tree

- Semantically Coherent Character Animation from Multimodal Inputs Survey Taxonomy
- Audio-Driven Facial and Co-Speech Animation
  - Co-Speech Gesture Synthesis (4 papers)
  - [1] SemTalk: Holistic Co-speech Motion Generation with Frame-level Semantic Emphasis (Zhang xiangyue, 2025) View paper
  - [3] Gesturediffuclip: Gesture diffusion model with clip latents (Tenglong Ao, 2023) View paper
  - [14] Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis (Liu Hai-yang, 2022) View paper
  - [23] MAG: Multi-Modal Aligned Autoregressive Co-Speech Gesture Generation without Vector Quantization (Liu Binjie, 2025) View paper
  - Emotional and Expressive Facial Animation (4 papers)
  - [9] Medtalk: Multimodal controlled 3d facial animation with dynamic emotions by disentangled embedding (Chang Liu, 2025) View paper
  - [29] Moee: Mixture of emotion experts for audio-driven portrait animation (Huaize Liu, 2025) View paper
  - [38] Punchline-Driven Hierarchical Facial Animation via Multimodal Large Language Models (Wang, 2025) View paper
  - [49] Multimodal Speech-Driven Facial Animation with Text Attention Control (Jian Zhang, 2025) View paper
  - Interactive and Real-Time Avatar Synthesis (2 papers)
  - [10] Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation (Chen Ming, 2025) View paper
  - [18] Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis (Ding, 2025) View paper
- Image-to-Video Character Animation
  - Pose-Driven and Controllable Image Animation (3 papers)
  - [2] Animate anyone: Consistent and controllable image-to-video synthesis for character animation (H. Li, 2024) View paper
  - [16] Hallo4: High-Fidelity Dynamic Portrait Animation via Direct Preference Optimization and Temporal Motion Modulation (Cui Jiahao, 2025) View paper
  - [45] CharacterShot: Controllable and Consistent 4D Character Animation (Gao Junyao, 2025) View paper

- Multimodal-Driven Portrait and Character Animation ★ (4 papers)
- [0] Instilling an Active Mind in Avatars via Cognitive Simulation (Anon et al., 2026) View paper
- [8] HunyuanVideo-HOMA: Generic Human-Object Interaction in Multimodal Driven Human Animation (Huang Ziyao, 2025) View paper
- [13] InterActHuman: Multi-Concept Human Animation with Layout-Aligned Audio Conditions (Wang Zhenzhi, 2025) View paper
- [27] Versatile multimodal controls for expressive talking human animation (Zheng Qin, 2025) View paper
- Scene-Aware and Environment-Conditioned Animation (1 papers)
- [39] Animate Anyone 2: High-Fidelity Character Image Animation with Environment Affordance (Hu Li, 2025) View paper

• Text-to-Motion Synthesis and Control
- Diffusion-Based Text-to-Motion Generation (4 papers)
- [7] Mofusion: A framework for denoising-diffusion-based motion synthesis (Rishabh Dabral, 2023) View paper
- [12] Efficient Text-driven Motion Generation via Latent Consistency Training (Zhu, 2024) View paper
- [22] Lead: Latent realignment for human motion diffusion (Nefeli Andreou, 2025) View paper
- [24] Motionclr: Motion generation and training-free editing via understanding attention mechanisms (Ling-Hao Chen, 2024) View paper
- Language-Motion Alignment and Pretraining (1 papers)
- [4] Lamp: Language-motion pretraining for motion generation, retrieval, and captioning (Li Zhe, 2024) View paper
- Trajectory and Multi-Level Control (2 papers)
- [17] Tlcontrol: Trajectory and language control for human motion synthesis (Wan Weilin, 2024) View paper
- [25] Story-to-motion: Synthesizing infinite and controllable character animation from long text (Zhongfei Qing, 2023) View paper
- Style and Attribute Control in Text-to-Motion (2 papers)
- [20] Semantically Consistent Text-to-Motion with Unsupervised Styles (Linjun Wu, 2025) View paper
- [36] Breaking the limits of text-conditioned 3d motion synthesis with elaborative descriptions (Yijun Qian, 2023) View paper

• Scene-Aware and Context-Conditioned Motion Synthesis
- Multimodal Scene-Informed Motion Prediction (2 papers)
- [6] Multimodal sense-informed prediction of 3d human motions (Lou, 2024) View paper
- [11] Multimodal sense-informed forecasting of 3d human motions (Zhenyu Lou, 2024) View paper
- Semantic Scene Representation for Motion Synthesis (2 papers)
- [30] Human Motion Synthesis in 3D Scenes via Unified Scene Semantic Occupancy (Gong Jingyu, 2025) View paper
- [37] Scene-aware generative network for human motion synthesis (Jingbo Wang, 2021) View paper

• Long-Form and Narrative-Driven Animation
- Story-to-Animation and Script-Based Generation (3 papers)
- [5] Moviedreamer: Hierarchical generation for coherent long visual sequence (Liu, 2024) View paper
- [15] A Large Language Model-Based System for Semantic Understanding and Automated Scene Generation in Animation Scripts (Ke Tian, 2025) View paper
- [28] FairyGen: Storied Cartoon Video from a Single Child-Drawn Character (Zheng Jiayi, 2025) View paper
- Agent-Based and Autonomous Animation Systems (1 papers)
- [19] Anim-director: A large multimodal model powered agent for controllable animation video generation (Yunxin Li, 2024) View paper

• Motion Style Transfer and Cross-Domain Adaptation
- Dance and Music-Driven Style Transfer (2 papers)
- [32] Multimodal dance style transfer (Wenjie Yin, 2023) View paper
- [33] Dance style transfer with cross-modal transformer (Wenjie Yin, 2023) View paper
- Cross-Category and Behavioral Motion Transfer (1 papers)
- [21] Behave Your Motion: Habit-preserved Cross-category Animal Motion Transfer (Zhimin Zhang, 2025) View paper

• Unified and Multimodal Motion Synthesis Frameworks
- Unified Task Frameworks for Motion Synthesis (1 papers)
- [31] A unified 3d human motion synthesis model via conditional variational auto-encoder (Yujun Cai, 2021) View paper
- Multimodal Input Integration for Motion Generation (2 papers)
- [46] A Unified Framework for Human Motion Generation with Multimodal Inputs (Isabella M. Cooper, 2025) View paper
- [50] Text-driven Motion Synthesis and Interaction Generation using Masked Deconstructed Diffusion and Multi-task Scene-aware Models (Chen, 2025) View paper

• Semantic and Conceptual Animation Systems
- Semantic Frameworks and Ontology-Based Animation (2 papers)
- [26] Application of animation products via multimodal information and semantic analogy (Keke Chu, 2024) View paper
- [42] Semantic framework for interactive animation generation and its application in virtual shadow play performance (Hui Liang, 2018) View paper
- Multimodal Interface and Natural Language Control (1 papers)
- [40] A multimodal interface for virtual character animation based on live performance and natural language processing (F. Lamberti, 2019) View paper
- Character Generation with Semantic Control (1 papers)
- [48] Character Generation Powered by Multimodal Foundation Models: Exploring Semantic Control and Visual Consistency Mechanisms (Jiangxue Han, 2025) View paper

• Surveys, Benchmarks, and Theoretical Foundations (6 papers)
- [34] Semantic-Driven Multi-character Multi-motion 3D Animation Generation (Hui Liang, 2024) View paper
- [35] AI-driven knowledge-based motion synthesis algorithms for graphics and animation (Chai, 2024) View paper
- [41] A Survey on Human Motion Generation Tasks: Consistency, Diversity, and Customization (Jiang Xinyu, 2024) View paper
- [43] Virtual Reality Generation from Natural Language (Bouali, 2025) View paper
- [44] Multimodal Generative AI with Autoregressive LLMs for Human Motion Understanding and Generation: A Way Forward (Islam Muhammad, 2025) View paper
- [47] Multimodal Expressive Gesturing With Style (Fares, 2023) View paper

Narrative

Core task: Semantically coherent character animation from multimodal inputs. The field has evolved into a rich ecosystem organized around input modalities and synthesis objectives. Audio-Driven Facial and Co-Speech Animation focuses on generating expressive talking heads and gestures synchronized with speech, often leveraging datasets like BEAT Dataset[14] and methods such as SemTalk[1] and MedTalk[9]. Image-to-Video Character Animation emphasizes animating static portraits or full-body characters from reference images, with works like Animate Anyone[2] and HunyuanVideo HOMA[8] demonstrating strong visual fidelity. Text-to-Motion Synthesis and Control explores language-conditioned body motion generation, balancing semantic understanding with physical plausibility through approaches like MoFusion[7] and Efficient Text Motion[12]. Scene-Aware and Context-Conditioned Motion Synthesis integrates environmental constraints, while Long-Form and Narrative-Driven Animation tackles temporal coherence over extended sequences using methods such as MovieDreamer[5] and Story to Motion[25]. Motion Style Transfer and Cross-Domain Adaptation addresses stylistic variation, and Unified and Multimodal Motion Synthesis Frameworks aim to handle diverse input combinations within single architectures, exemplified by Unified Multimodal Motion[46] and Multimodal Autoregressive Motion[44].

Recent efforts reveal a tension between modality-specific depth and cross-modal generalization. Audio-driven methods achieve fine-grained lip-sync and gesture timing but may struggle with broader semantic grounding, whereas text-driven approaches offer flexible high-level control at the cost of temporal precision. Active Mind Avatars[0] sits within the Image-to-Video Character Animation branch, specifically in Multimodal-Driven Portrait and Character Animation, where it shares conceptual ground with InterActHuman[13] and Versatile Multimodal Controls[27]. Unlike purely image-based animators such as Animate Anyone 2[39], Active Mind Avatars[0] emphasizes integrating cognitive or semantic signals to drive character behavior, aligning with the broader push toward semantically aware synthesis seen in works like Semantically Consistent Motion[20] and AI Knowledge Motion[35]. This positioning reflects an emerging interest in bridging low-level visual fidelity with higher-level intentionality, a theme that cuts across multiple branches and remains an active area of exploration.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. HunyuanVideo-HOMA: Generic Human-Object Interaction in Multimodal Driven Human Animation

Authors: Huang Ziyao, Zhou Zixiang, Ziyao Huang, Cao Juan, Zixiang Zhou, et al. (24 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

To address key limitations in human-object interaction (HOI) video generation -- specifically the reliance on curated motion data, limited generalization to novel objects/scenarios, and restricted accessibility -- we introduce HunyuanVideo-HOMA, a weakly conditioned multimodal-driven framework. HunyuanVideo-HOMA enhances controllability and reduces dependency on precise inputs through sparse, decoupled motion guidance. It encodes appearance and motion signals into the dual input space of a multi...

Relationship Analysis

Both papers belong to the Multimodal-Driven Portrait and Character Animation category, focusing on synthesizing expressive character videos from multimodal inputs. While the original paper emphasizes cognitive simulation using MLLM-based agents to generate semantically coherent animations from audio, text, and images with deliberative reasoning (System 1 vs System 2), HunyuanVideo-HOMA specifically targets human-object interaction scenarios with customizable objects and sparse pose sequences as additional control modalities, representing a more specialized application within the broader multimodal animation domain.

---

2. InterActHuman: Multi-Concept Human Animation with Layout-Aligned Audio Conditions

Authors: Wang Zhenzhi, Yang Jiaqi, Zhenzhi Wang, Jiang Jianwen, Jiaqi Yang, et al. (19 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

End-to-end human animation with rich multi-modal conditions, e.g., text, image and audio has achieved remarkable advancements in recent years. However, most existing methods could only animate a single subject and inject conditions in a global manner, ignoring scenarios that multiple concepts could appears in the same video with rich human-human interactions and human-object interactions. Such global assumption prevents precise and per-identity control of multiple concepts including humans and o...

Relationship Analysis

Both papers belong to the Multimodal-Driven Portrait and Character Animation category, focusing on synthesizing expressive character videos from multimodal inputs including audio, images, and text. They overlap in addressing audio-driven animation with semantic control, but differ fundamentally in their core contributions: the original paper introduces a cognitive dual-system framework (System 1/System 2) using MLLM-based reasoning for deliberative motion planning and a pseudo-last-frame conditioning strategy, while the candidate paper focuses on multi-concept human animation with explicit layout-aligned audio conditions through iterative mask prediction for handling multiple identities and human-object interactions.

---

3. Versatile multimodal controls for expressive talking human animation

Authors: Zheng Qin, Ruobing Zheng, Yabing Wang, Tianqi Li, Zixin Zhu, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

In filmmaking, directors typically allow actors to perform freely based on the script before providing specific guidance on how to present key actions. AI-generated content faces similar requirements, where users not only need automatic generation of lip synchronization and basic gestures from audio input but also desire semantically accurate and expressive body movement that can be ''directly guided'' through text descriptions. Therefore, we present VersaAnimator, a versatile framework that syn...

Relationship Analysis

Both papers belong to the Multimodal-Driven Portrait and Character Animation category, focusing on synthesizing expressive character videos from images using combined audio, text, and visual inputs. They overlap in their use of multimodal conditioning (audio + text) to generate semantically coherent animations with lip-sync and body movements. The key difference is that the original paper emphasizes cognitive simulation through MLLM-based reasoning agents (System 1/System 2 framework) and a pseudo-last-frame strategy for identity preservation, while the candidate paper focuses on versatile motion generation using 3D motion tokens with VQ-VAE tokenization and a token-to-pose translator for enhanced body movement realism.

Contributions Analysis

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

Contribution 1: Dual-system cognitive framework for video avatar generation

Description: The authors introduce a novel perspective that frames video avatar generation using dual-process cognitive theory, distinguishing between reactive System 1 processes (low-level audio-to-motion mappings) and deliberative System 2 processes (high-level

semantic reasoning). This framework addresses the limitation that existing methods only simulate reactive behavior without contextual reasoning.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Active Intelligence in Video Avatars via Closed-loop World Modeling

URL: View paper

Brief Assessment

Active Intelligence Avatars[60] focuses on goal-directed planning in interactive environments using a POMDP formulation with closed-loop world modeling, not on distinguishing reactive vs. deliberative processes for avatar generation from audio-to-motion mappings.

Contribution 2: MLLM-based agentic reasoning module with specialized MMDiT architecture

Description: The framework employs Multimodal Large Language Models as agents to generate high-level semantic guidance through multi-step reasoning (Analyzer and Planner). It integrates this with a specialized Multimodal Diffusion Transformer architecture that uses symmetric fusion of text, audio, and video branches, along with a novel pseudo-last-frame conditioning strategy to mitigate modal interference.

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. Multimodal LLM-Guided Semantic Correction in Text-to-Image Diffusion

URL: View paper

Brief Assessment

LLM Semantic Correction[65] focuses on semantic correction in text-to-image diffusion using MLLMs as observers during inference, not on video avatar generation with audio-driven synthesis and temporal motion modeling as in the original paper.

---

2. A survey of multimodal controllable diffusion models

URL: View paper

Brief Assessment

Controllable Diffusion Survey[64] focuses on controllable generation techniques in diffusion models across various modalities and applications, but does not specifically address MLLM-based agentic reasoning modules for semantic guidance in video avatar generation or the specialized MMDiT architecture with pseudo-last-frame conditioning described in the original paper.

---

3. Lavida: A large diffusion language model for multimodal understanding

URL: View paper

Brief Assessment

LaVida[63] focuses on diffusion-based vision-language models for multimodal understanding, not on agentic reasoning for avatar animation. The candidate employs diffusion transformers for image/video generation tasks, whereas the original paper uses MLLMs as deliberative agents to guide avatar motion synthesis with a novel pseudo-last-frame conditioning strategy.

---

4. Next-gpt: Any-to-any multimodal llm

URL: View paper

Brief Assessment

Next GPT[62] focuses on any-to-any multimodal generation using diffusion decoders for output synthesis, not on video avatar animation with MMDiT architectures for motion generation. The candidate employs MLLMs for multimodal understanding and routing to different decoders, whereas the original uses MLLMs specifically for high-level semantic guidance in character animation with a specialized symmetric fusion architecture and pseudo-last-frame conditioning.

---

5. Multimodal llm integrated semantic communications for 6g immersive experiences

URL: View paper

Brief Assessment

LLM Semantic Communications[69] focuses on wireless communication systems for 6G networks, using MLLMs for semantic guidance in bandwidth allocation and compression tasks. The original paper addresses video avatar generation with cognitive simulation, employing MLLMs for character behavior planning and a specialized MMDiT for multimodal fusion with pseudo-last-frame conditioning. These are fundamentally different application domains and technical objectives.

---

6. Dimba: Transformer-mamba diffusion models

URL: View paper

Brief Assessment

Dimba[70] focuses on a hybrid Transformer-Mamba architecture for text-to-image diffusion models, not on MLLM-based agents for semantic guidance or multimodal video generation with audio conditioning.

---

7. Query-kontext: An unified multimodal model for image generation and editing

URL: View paper

Brief Assessment

Query Kontext[68] focuses on image generation and editing tasks using a different architectural approach. While both use MLLMs and diffusion transformers, Query Kontext's design delegates reasoning to VLMs for image synthesis tasks, whereas the original paper addresses video avatar generation with audio-driven motion synthesis and a novel pseudo-last-frame strategy for temporal consistency.

---

8. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms

URL: View paper

Brief Assessment

Mastering Text Image[67] focuses on text-to-image diffusion with MLLMs for recaptioning and spatial planning, not audio-driven video avatar generation with multimodal fusion of audio, text, and video branches.

---

9. Target-aware video diffusion models

URL: View paper

Brief Assessment

Target Aware Diffusion[66] focuses on spatial target specification via segmentation masks for video generation, not on MLLM-based semantic reasoning agents or multimodal diffusion transformer architectures for avatar animation.

---

10. Llmga: Multimodal large language model based generation assistant

URL: View paper

Brief Assessment

LLMGA[61] focuses on using MLLMs to generate detailed language prompts for controlling Stable Diffusion in image generation/editing tasks, not on video avatar generation with multimodal diffusion transformers for audio-visual synthesis.

Contribution 3: Pseudo-last-frame conditioning strategy

Description: A novel conditioning mechanism that discards the reference image during training and instead uses a pseudo-last-frame with shifted positional encoding during inference. This approach eliminates training artifacts where models learn spurious correlations between reference images and generated sequences, enabling better motion dynamics while maintaining identity consistency.

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. Leo: Generative latent image animator for human video synthesis

URL: View paper

Brief Assessment

Leo[51] focuses on human video synthesis using flow maps for motion representation and a warp-and-inpaint approach. The candidate does not discuss discarding reference images during training or using pseudo-last-frames with shifted positional encoding, which are the core elements of the original contribution.

---

2. Proteus-ID: ID-Consistent and Motion-Coherent Video Customization

URL: View paper

Brief Assessment

Proteus-ID[57] appears to focus on identity-consistent video customization rather than avatar animation with reference image conditioning. The candidate's abstract mentions video customization but does not provide sufficient detail about conditioning mechanisms to assess overlap with the pseudo-last-frame strategy.

---

3. Motioncharacter: Identity-preserving and motion controllable human video generation

URL: View paper

Brief Assessment

MotionCharacter[58] focuses on identity preservation in human video generation using face embeddings and ID-preserving modules, not on pseudo-last-frame conditioning strategies for eliminating training artifacts in reference image conditioning.

---

4. Animate anyone: Consistent and controllable image-to-video synthesis for character animation

URL: View paper

Prior Art Analysis

Animate Anyone[2] demonstrates prior work on a similar conditioning mechanism for identity preservation in character animation. Both papers address the challenge of maintaining character identity consistency during video generation while enabling dynamic motion. Animate Anyone[2] introduces ReferenceNet with spatial attention to preserve appearance details from a reference image, which serves a functionally similar purpose to the original paper's pseudo-last-frame strategy—both aim to maintain identity without constraining motion dynamics. The candidate paper explicitly discusses the trade-off between identity preservation and motion dynamics, noting that their approach 'effectively maintains the spatial and temporal consistency of character appearance in videos' while producing 'high-definition videos without issues such as temporal jitter or flickering,' which parallels the original paper's goals of 'eliminating training artifacts' and 'enabling better motion dynamics while maintaining identity consistency.'

Evidence

Evidence 1 - **Rationale**: Both papers address the fundamental challenge of preserving identity from a reference image while enabling dynamic motion. While the technical implementations differ (pseudo-last-frame with shifted positional encoding vs. ReferenceNet with spatial attention), both serve the same functional purpose of maintaining appearance consistency without constraining motion. - **Original**: our solution is to discard the reference image entirely during training and introduce a novel guidance mechanism. as shown in figure 2 (bottomright), we instead probabilistically condition the model on the gt first and last frames of the video clip, both native signals, each with a dropout probabili... - **Candidate**: to address the challenge of maintaining appearance consistency, we introduce referencenet, specifically designed as a symmetrical unet structure to capture spatial details of the reference image. at each corresponding layer of the unet blocks, we integrate features from referencenet into the denoisi...

Evidence 2 - **Rationale**: Both papers identify the same core problem: maintaining character identity consistency from a reference image during video generation. This shows that Animate Anyone[2] was already addressing the challenge that the original paper claims as novel. - **Original**: rethinking reference image conditioning.a critical input in video avatar models is the reference image, which serves two distinct purposes: first, providing an initial frame as a conditioning prefix for the generated sequence, and second, maintaining identity consistency. while the former is a neces... - **Candidate**: character animation aims to generating character videos from still images through driving signals. currently, diffusion models have become the mainstream in visual generation research, owing to their robust generative capabilities. however, challenges persist in the realm of image-to-video, especial...

Evidence 3 - **Rationale**: Both papers claim to achieve the same outcome: maintaining identity consistency while enabling dynamic motion without temporal artifacts. Animate Anyone[2]'s claims of maintaining 'spatial and temporal consistency' and producing videos 'without issues such as temporal jitter or flickering' directly parallel the original paper's goals. - **Original**: this pseudo frame functions as a 'carrot on a stick': it guides the model toward the target identity without ever forcing it to replicate the static image, which is discarded after synthesis. as our experiments show, this approach eliminates training artifacts and mitigates autoregressive error, ach... - **Candidate**: compared to previous methods, our approach presents several notable advantages. firstly, it effectively maintains the spatial and temporal consistency of character appearance in videos. secondly, it produces highdefinition videos without issues such as temporal jitter or flickering. thirdly, it is c...

---

5. DualReal: Adaptive Joint Training for Lossless Identity-Motion Fusion in Video Customization

URL: View paper

Brief Assessment

DualReal[55] focuses on joint identity-motion customization in video generation through adaptive training strategies, not on conditioning mechanisms for reference images. The candidate does not address pseudo-last-frame strategies or shifted positional encoding for eliminating training artifacts in reference-based conditioning.

---

6. DanceTogether! Identity-Preserving Multi-Person Interactive Video Generation

URL: View paper

Brief Assessment

DanceTogether[52] focuses on multi-person video generation with pose-mask fusion for identity preservation, not on reference image conditioning strategies for training artifact elimination. The candidate's maskposeadapter addresses identity-action binding through pose and mask fusion, which is a different technical approach than discarding reference images during training with shifted positional encoding.

---

7. Videomage: Multi-subject and motion customization of text-to-video diffusion models

URL: View paper

Brief Assessment

VideoMage[59] addresses a different problem (multi-subject video customization with identity preservation) using a pseudo-last-frame approach for maintaining subject identity during motion, while the original paper focuses on avatar animation with semantic guidance. The technical contexts and objectives differ substantially.

---

8. Motionbooth: Motion-aware customized text-to-video generation

URL: View paper

Brief Assessment

MotionBooth[53] addresses video generation with customized subjects and motion control, but does not propose a pseudo-last-frame conditioning strategy. The candidate focuses on subject learning and motion control through different mechanisms (subject region loss, video preservation loss, latent shift modules, and cross-attention manipulation), which are technically distinct from the original paper's conditioning approach for identity preservation and motion dynamics.

---

9. Levitor: 3d trajectory oriented image-to-video synthesis

URL: View paper

Brief Assessment

Levitor[54] focuses on 3D trajectory control for image-to-video synthesis using depth information and k-means clustering, not on conditioning mechanisms for identity preservation in avatar generation. The technical domains and objectives differ fundamentally.

---

10. Dreamvideo: Composing your dream videos with customized subject and motion

URL: View paper

Brief Assessment

DreamVideo[56] focuses on customized video generation with subject and motion learning, not on conditioning mechanisms for identity preservation. The paper does not address the specific problem of eliminating training artifacts from reference image conditioning during training.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Instilling an Active Mind in Avatars via Cognitive Simulation View paper
- [1] SemTalk: Holistic Co-speech Motion Generation with Frame-level Semantic Emphasis View paper
- [2] Animate anyone: Consistent and controllable image-to-video synthesis for character animation View paper
- [3] Gesturediffuclip: Gesture diffusion model with clip latents View paper
- [4] Lamp: Language-motion pretraining for motion generation, retrieval, and captioning View paper
- [5] Moviedreamer: Hierarchical generation for coherent long visual sequence View paper
- [6] Multimodal sense-informed prediction of 3d human motions View paper
- [7] Mofusion: A framework for denoising-diffusion-based motion synthesis View paper
- [8] HunyuanVideo-HOMA: Generic Human-Object Interaction in Multimodal Driven Human Animation View paper
- [9] Medtalk: Multimodal controlled 3d facial animation with dynamic emotions by disentangled embedding View paper
- [10] Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation View paper
- [11] Multimodal sense-informed forecasting of 3d human motions View paper
- [12] Efficient Text-driven Motion Generation via Latent Consistency Training View paper
- [13] InterActHuman: Multi-Concept Human Animation with Layout-Aligned Audio Conditions View paper
- [14] Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis View paper
- [15] A Large Language Model-Based System for Semantic Understanding and Automated Scene Generation in Animation Scripts View paper
- [16] Hallo4: High-Fidelity Dynamic Portrait Animation via Direct Preference Optimization and Temporal Motion Modulation View paper
- [17] Tlcontrol: Trajectory and language control for human motion synthesis View paper
- [18] Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis View paper
- [19] Anim-director: A large multimodal model powered agent for controllable animation video generation View paper
- [20] Semantically Consistent Text-to-Motion with Unsupervised Styles View paper
- [21] Behave Your Motion: Habit-preserved Cross-category Animal Motion Transfer View paper
- [22] Lead: Latent realignment for human motion diffusion View paper
- [23] MAG: Multi-Modal Aligned Autoregressive Co-Speech Gesture Generation without Vector Quantization View paper
- [24] Motionclr: Motion generation and training-free editing via understanding attention mechanisms View paper
- [25] Story-to-motion: Synthesizing infinite and controllable character animation from long text View paper
- [26] Application of animation products via multimodal information and semantic analogy View paper
- [27] Versatile multimodal controls for expressive talking human animation View paper
- [28] FairyGen: Storied Cartoon Video from a Single Child-Drawn Character View paper
- [29] Moee: Mixture of emotion experts for audio-driven portrait animation View paper
- [30] Human Motion Synthesis in 3D Scenes via Unified Scene Semantic Occupancy View paper
- [31] A unified 3d human motion synthesis model via conditional variational auto-encoder View paper
- [32] Multimodal dance style transfer View paper

- [33] Dance style transfer with cross-modal transformer View paper
- [34] Semantic-Driven Multi-character Multi-motion 3D Animation Generation View paper
- [35] AI-driven knowledge-based motion synthesis algorithms for graphics and animation View paper
- [36] Breaking the limits of text-conditioned 3d motion synthesis with elaborative descriptions View paper
- [37] Scene-aware generative network for human motion synthesis View paper
- [38] Punchline-Driven Hierarchical Facial Animation via Multimodal Large Language Models View paper
- [39] Animate Anyone 2: High-Fidelity Character Image Animation with Environment Affordance View paper
- [40] A multimodal interface for virtual character animation based on live performance and natural language processing View paper
- [41] A Survey on Human Motion Generation Tasks: Consistency, Diversity, and Customization View paper
- [42] Semantic framework for interactive animation generation and its application in virtual shadow play performance View paper
- [43] Virtual Reality Generation from Natural Language View paper
- [44] Multimodal Generative AI with Autoregressive LLMs for Human Motion Understanding and Generation: A Way Forward View paper
- [45] CharacterShot: Controllable and Consistent 4D Character Animation View paper
- [46] A Unified Framework for Human Motion Generation with Multimodal Inputs View paper
- [47] Multimodal Expressive Gesturing With Style View paper
- [48] Character Generation Powered by Multimodal Foundation Models: Exploring Semantic Control and Visual Consistency Mechanisms View paper
- [49] Multimodal Speech-Driven Facial Animation with Text Attention Control View paper
- [50] Text-driven Motion Synthesis and Interaction Generation using Masked Deconstructed Diffusion and Multi-task Scene-aware Models View paper
- [51] Leo: Generative latent image animator for human video synthesis View paper
- [52] DanceTogether! Identity-Preserving Multi-Person Interactive Video Generation View paper
- [53] Motionbooth: Motion-aware customized text-to-video generation View paper
- [54] Levitor: 3d trajectory oriented image-to-video synthesis View paper
- [55] DualReal: Adaptive Joint Training for Lossless Identity-Motion Fusion in Video Customization View paper
- [56] Dreamvideo: Composing your dream videos with customized subject and motion View paper
- [57] Proteus-ID: ID-Consistent and Motion-Coherent Video Customization View paper
- [58] Motioncharacter: Identity-preserving and motion controllable human video generation View paper
- [59] Videomage: Multi-subject and motion customization of text-to-video diffusion models View paper
- [60] Active Intelligence in Video Avatars via Closed-loop World Modeling View paper
- [61] Llmga: Multimodal large language model based generation assistant View paper
- [62] Next-gpt: Any-to-any multimodal llm View paper
- [63] Lavida: A large diffusion language model for multimodal understanding View paper
- [64] A survey of multimodal controllable diffusion models View paper
- [65] Multimodal LLM-Guided Semantic Correction in Text-to-Image Diffusion View paper
- [66] Target-aware video diffusion models View paper
- [67] Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms View paper
- [68] Query-kontext: An unified multimodal model for image generation and editing View paper
- [69] Multimodal llm integrated semantic communications for 6g immersive experiences View paper
- [70] Dimba: Transformer-mamba diffusion models View paper