Crowd Motion Generation: Report

Daniel Bar-Lev, Anton Shchukin, Kfir Barzilay

Tel-Aviv University

Submition Date: 28.2.2024

1 Problem Statement And Background

1.1 Theoretical Background

AI-generated 3D human model animation has evolved into an established field, transforming digital content creation in industries such as gaming, film, virtual reality, and social media. By leveraging advanced machine learning algorithms, including deep neural networks, generative adversarial networks (GANs), and diffusion models, this field has achieved remarkable realism in character movement, facial expressions, and body dynamics. Diffusion models, a class of generative models that learn to progressively refine noise into structured data, have recently shown great promise in motion synthesis. These models generate diverse and high-quality motion sequences by iteratively denoising random samples, allowing for greater control and variety in character movements. Unlike traditional animation methods that rely heavily on manual keyframing or expensive motion capture systems, AI-driven approaches, particularly those using diffusion-based motion generation, automate motion synthesis and behavioral modeling, significantly reducing production time and cost while maintaining high fidelity and fluidity in animation.

1.2 Problem Statement

In digital content creation for industries such as film, gaming, virtual reality, and simulation training, generating large quantities of 3D human models moving in semantically meaningful ways presents significant challenges. Traditional animation techniques require extensive manual effort and expensive motion capture systems to create diverse, realistic movements, especially when depicting complex scenarios such as crowd dynamics, group choreography, or interactive narrative scenes. Additionally, achieving semantic coherence—where characters' movements align with contextual intentions specified by a text prompt (e.g., running away in fear, dancing joyfully, or executing coordinated tactical maneuvers)—is time-consuming and often requires skilled animators.

Key Industry Needs:

- Scalable Crowd Simulation: Animations of large crowds with unique, contextually appropriate movements.
- Semantic Motion Control: Ensuring that character movements align with

narrative intent, such as expressing specific emotions or following complex instructions.

- Interactivity and Real-Time Adaptation: In interactive applications such as virtual reality or gaming, characters need to respond dynamically to user input or environmental changes.
- Diverse Movement Styles and Personalization: Creating varied movement patterns to represent different personalities or cultural expressions.

2 Solution Proposal

2.1 Models Used

2.1.1 TRACE

TRACE [1] is a deep learning-based trajectory prediction model designed to generate realistic and semantically meaningful human movement paths in complex environments. Unlike traditional pathfinding algorithms, TRACE learns from real-world motion patterns to create trajectories that respect spatial constraints, social groups, and scene layouts. This enables the generation of large-scale, non-overlapping, and dynamically coherent crowd movements that adapt to environmental factors.

2.1.2 PriorMDM

PriorMDM [2] is a state-of-the-art motion diffusion model that generates detailed and diverse human motions based on natural language prompts. It refines human animations by synthesizing expressive and contextually accurate movements, capturing subtle variations in motion style and intent. By conditioning on textual descriptions (e.g., "running in fear" or "dancing gracefully"), PriorMDM provides a high degree of control over character animations, making it a powerful tool for fine-tuning motion sequences in AI-driven animation workflows.

2.2 Solution Proposal

To address the challenges of generating large quantities of 3D human models moving in semantically meaningful ways, we propose a pipeline that combines TRACE and PriorMDM, leveraging their complementary strengths to achieve scalable, coherent, and fine-tuned motion generation. Our approach ensures that animated crowds not only follow semantically cohesive trajectories (thanks to TRACE) but also exhibit diverse and contextually appropriate movements in response to text prompts (thanks to PriorMDM).

2.3 Methodology

2.3.1 Trajectory Generation with TRACE

PriorMDM is a state-of-the-art motion diffusion model that generates detailed and diverse human motions based on natural language prompts. It refines human animations by synthesizing expressive and contextually accurate movements, capturing subtle variations in motion style and intent. By conditioning on textual descriptions (e.g., "running in fear" or "dancing gracefully"), PriorMDM provides a high degree of control over character animations, making it a powerful tool for fine-tuning motion sequences in AI-driven animation workflows.

2.3.2 Motion Fine-Tuning with PriorMDM

Once TRACE provides structured trajectories, we apply PriorMDM to generate fine-grained and contextually relevant motions for each individual character. PriorMDM, a text-conditioned motion diffusion model, refines animations by generating detailed human movements that align with specific prompts (e.g., "dancing energetically," "limping in pain," or "performing parkour jumps"). This step adds natural motion diversity while preserving trajectory coherence, allowing for a variety of expressive animations across large-scale scenes.

2.4 Advantages of Our Approach

Our TRACE-PriorMDM pipeline effectively addresses the industry needs outlined in (1.2) while offering standalone advantages that enhance scalability, coherence, customization, and efficiency. By integrating TRACE for trajectory planning and PriorMDM for detailed motion synthesis, our solution provides a robust framework for AI-driven 3D human crowd animation, paving the way for more immersive and dynamic digital experiences.

- **Scalability:** TRACE generates large-scale, semantically meaningful trajectories for many characters, ensuring coherent movement across entire crowds. This allows for efficient, high-volume animation generation without manual intervention.
- **Semantic Coherence:** After TRACE generates trajectories, PriorMDM adds detailed, prompt-driven movements to each individual, ensuring that motions align with high-level narrative intentions (e.g., running away in fear, dancing joyfully). By first defining structured trajectories and then refining individual motions, we ensure that movements remain meaningful at both macro (crowd) and micro (individual) levels.
- **Customization and Diversity:** PriorMDM's text-based control enables highly customizable motions, allowing for different movement styles, emotional expressions, and character-specific behaviors. This flexibility supports

personalized animations across film, VR, and gaming, whether it's age-based walking variations, culturally distinct gestures, or stylized artistic movements.

- **Interactivity and Real-Time Adaptation:** While TRACE and PriorMDM primarily generate precomputed animations, future work can integrate them into game engines (Unreal Engine, Unity) to enable real-time adaptation, allowing AI-driven characters to respond dynamically to user interactions and environmental changes.
- **Efficiency:** The automation of both trajectory generation (TRACE) and detailed motion synthesis (PriorMDM) eliminates the need for extensive manual keyframing or motion capture, significantly reducing production time and costs, making it viable for large-scale applications while maintaining high-quality, realistic animations.

3 Solution Implementation and Results

3.1 Solution Implementation

Our pipeline integrates TRACE for trajectory generation and PriorMDM for detailed motion synthesis, ensuring large-scale, semantically coherent 3D human animation.

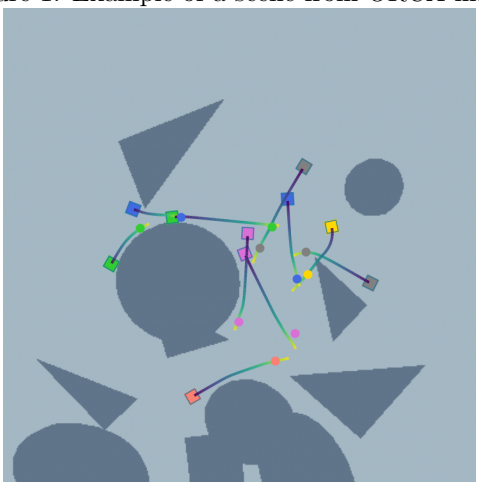
We start by running TRACE's ORCA-trained model on the ORCA-Maps test set, simulating all agents within each test scene. This generates local coordinate trajectories for each agent, which we then convert into world coordinates, adding the missing z-dimension to create fully defined 3D movement paths. Each agent's initial position is saved for later use in scene reconstruction.

Since TRACE produces trajectories at 10 FPS, while PriorMDM requires 20 FPS, we interpolate positions to maintain temporal consistency. Additionally, because HumanML3D represents the ground plane as x-z, while TRACE uses x-y, we swap the y and z axes to properly align the coordinate systems.

To efficiently process multiple motion sequences, we broadcast root trajectory data across a batch, enabling PriorMDM to generate synthetic motion data at scale. Each trajectory is then passed into PriorMDM, which applies text-conditioned motion generation, creating realistic and semantically meaningful human movements. To refine control over motion details, we sample from a pre-trained motion diffusion model, ensuring smoother transitions and greater variability in movement expression.

Finally, we reconstruct the full animated scene by placing multiple agents in their correct initial positions, maintaining their relative spatial arrangements. This ensures that the resulting crowd motion is both cohesive and contextually accurate, meeting the needs of applications in film, gaming, virtual environments, and large-scale simulations.

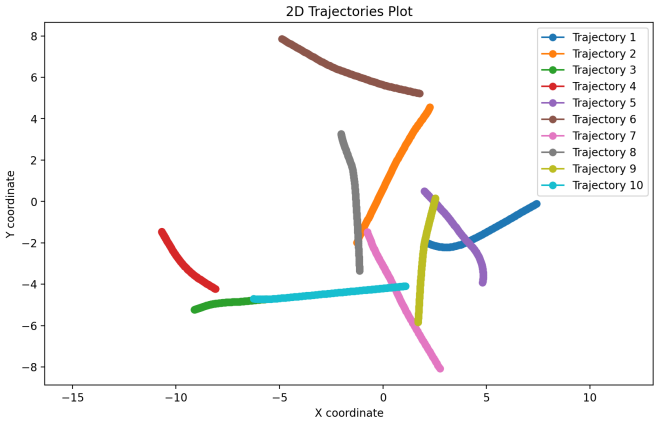
Figure 1: Example of a scene from ORCA maps.



3.2 Results

The implemented model successfully generated crowd animations according to different prompts, demonstrating its ability to create varied and dynamic movements. The generated animations displayed realistic crowd behaviors, with agents responding appropriately to given conditions.
Overall, the model performed as intended, with clear differentiation in movement patterns based on the input prompts. The animations exhibited fluid motion, and the individuals within the crowd displayed appropriate spacing and interactions, aligning with expected emergent behaviors in crowd dynamics.
However, a notable issue was observed: the generated crowd movements consistently moved backward instead of forward. This unexpected behavior suggests a potential inversion in directional encoding within the simulation parameters. Further analysis is required to determine whether this issue stems from input misinterpretation, incorrect vector assignments, or an inherent bias in the model's movement generation algorithm.
Despite this issue, the results confirm the model's capability in handling crowd simulation effectively. Future iterations will focus on resolving the directional movement error while maintaining the diversity and responsiveness achieved in the current implementation.
The results are shown in figures 1,2,3 and 4.

Figure 2: Example of a trajectories plot.



4 Ideas for Future Work and Continuation

By advancing the following areas, future work can push our solution (and AI-generated human animation in general) toward greater realism, interactivity, and adaptability, ensuring that this technology continues to shape and expand the creative and practical possibilities across industries such as gaming, virtual reality, film production, and digital art.

4.1 Correction of Movement Direction in Crowd Simulation

An important aspect of future improvements involves addressing the issue where all agents in the simulation moved backward instead of forward. This behavior suggests a potential misalignment in the animation direction, which could be caused by incorrect movement vectors, coordinate system inconsistencies, or an error in how the motion data was applied. Future work should focus on diagnosing this issue by refining the movement logic, ensuring that the directional vectors align correctly with the intended crowd behavior.

4.2 Real-Time Motion Adaptation

Implement real-time trajectory and motion adaptation to enhance interactivity in AI-generated animations. This would allow characters to dynamically adjust their movements based on changes in their environment, such as obstacles, hazards, or direct user input. Additionally, the system could incorporate more fluid adjustments to motions, such as the ability to switch from walking to running or alter posture in response to situational cues. Integration with game engines like Unreal Engine or Unity would be crucial for interactive applications, enabling

Figure 3: A gif example of the resulting trajectories with prompt "A person is running".

users to seamlessly manipulate characters within virtual environments in gaming or virtual reality (VR) experiences. Real-time adaptation would also support more immersive, responsive NPCs in games, simulations, and training environments.

4.3 Enhanced Text-to-Motion Understanding

Enhance the TRACE model by adding text conditioning directly into the trajectory generation process, ensuring that the generated trajectories align with high-level instructions or narrative prompts from the start. This would enable the system to not only create motion paths but also embed the specific intent of the movement. For example, the prompt "a crowd evacuating from a burning building" would not only dictate the direction of movement but also influence the speed, urgency, and clustering of the individuals. By integrating more detailed text-to-motion understanding, the system can create both group dynamics and individual movements that are contextually relevant and responsive to the narrative or environment. This improvement will allow users to provide a simple prompt and automatically generate both trajectories and highly detailed motions based on that input.

Figure 4: A gif example of the resulting animation with prompt "A person is running".

4.4 Personalized Motion Styles and Character-Specific Behaviors

To expand the flexibility of AI-generated human animation, training the models on diverse datasets will enable the generation of personalized motion styles based on unique character traits. This customization could incorporate factors such as age-based gait variations, emotion-based movement adjustments, or even personality-driven behavior (e.g., a nervous individual walking with fidgety steps or a confident person walking with deliberate strides). Additionally, integrating user-controlled parameters would allow fine-tuning of motion realism, exaggeration, or stylistic preferences for specific contexts, such as animation and art. For instance, users could specify a more "cartoonish" style or a "hyper-realistic" movement, expanding the potential applications for this model in entertainment, art, and marketing.

4.5 Benchmarking and Evaluation Metrics

To establish a standard for evaluating the quality and authenticity of AI-generated motions, we propose the development of both quantitative and qualitative metrics. These metrics would assess the realism, diversity, and semantic accuracy of the motions, ensuring that the system not only produces varied movements but that these movements align correctly with the intended context. Benchmarking frameworks could include measures of motion smoothness, naturalness of interaction, and the adherence to semantic prompts. Additionally, the creation of

standardized datasets for benchmarking will provide a more structured approach to comparing different models in the field. This effort will support the broader research community in refining models, enhancing AI-generated motion fidelity, and ensuring consistency in evaluating motion generation systems.

4.6 Generating Databases for Future Models

One avenue for future work is using the model to generate large-scale databases of 3D human crowd movements. These databases could be used for training future models, contributing to a more expansive and diverse corpus of human motion data. By capturing a broad range of activities, such as walking, running, dancing, interacting, and even reacting to environmental stimuli, these databases would help improve the robustness of motion generation models. This could significantly enhance the diversity of movements in future AI-driven animation systems, making them more adaptable to varied contexts and applications, from large-scale simulations to personalized experiences.

4.7 Adding Non-Human Objects

Integrating non-human objects into the animation system would enhance diversity, interactivity, and realism, allowing AI-generated characters to engage with their environment in more dynamic ways. Characters could pick up, manipulate, and react to objects, adding depth to crowd behaviors and individual actions. This would improve narrative flexibility, enabling characters to interact with props or tools, respond to changes in their surroundings, and drive more complex storylines. For the entertainment industry, this could create more immersive, engaging experiences in gaming, VR, and animation, offering rich, context-driven interactions.

References

[1] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, Or Litany, Trace and pace: Controllable pedestrian animation via guided trajectory diffusion, in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[2] Yoni Shafir, Guy Tevet, Roy Kapon, Amit Haim Bermano, Human motion diffusion as a generative prior, in *The Twelfth International Conference on Learning Representations*, 2024.