CLIP-Conditioned Diffusion For Text-to-Pose Generation

Joel Markapudi
Northeastern University
markapudi.j@northeastern.edu

Tanay Grover
Northeastern University
grover.t@northeastern.edu

Abstract

Generating 3D human poses from natural language is a challenging problem at the intersection of generative modeling and embodied semantics. The core research objective is to synthesize anatomically valid poses that align semantically with open-ended text. We propose a conditional diffusion framework that maps language to pose using a CLIP-guided UNet architecture with multi-scale residual blocks and cross-attention layers. Our training pipeline incorporates a cluster-aware sampling strategy to ensure pose diversity and applies first-action syntactic segmentation to isolate the primary intent in caption text. Built on a joint-pruned representation of HumanML3D, our system operates on 66-dimensional static pose representations and leverages an optimized noise scheduler for stable denoising. The model achieves strong structural consistency and semantic fidelity, highlighting the potential of diffusion-based generative models for language-conditioned human pose synthesis.

1 Introduction

Human motion is one of the richest expressions of intent, yet generating anatomically coherent poses from language remains a non-trivial task. While full-scale libraries like MDM, T2M-GPT, and MotionDiffuse offer powerful pipelines for text-to-motion generation, they often involve complex orchestration, sequence modeling, and pretrained components. In contrast, this project takes on the fundamental challenge: can a clean, tractable diffusion model generate a realistic pose from a short natural language prompt — without full motion sequence modeling?

We present a research implementation of a conditional diffusion pipeline for static pose generation, built on a dimension-reduced representation of HumanML3D. Our system learns a mapping from caption to pose using a CLIP-guided UNet architecture with cross-attention layers, residual blocks, and an optimized noise scheduler. To reduce ambiguity, we train on the first meaningful action described in each caption, paired with the first frame of motion as its representative pose.

This work aims to probe the design space of grounded diffusion models, offering a simplified framework that bridges language and embodiment without the overhead of full-body motion synthesis. We demonstrate a strong capacity to internalize spatial priors and linguistic intent.

1.1 Dataset

We utilize a modification of the HumanML3D dataset—a large-scale benchmark designed specifically for text-to-motion generation tasks. The dataset is sourced via the HuggingFace datasets library (TeoGchx/HumanML3D) and includes three standard splits: 23,384 training samples, 1,460 validation samples, and 4,384 test samples. Each sample in the dataset is composed of a motion sequence paired with natural language descriptions, making it highly suitable for learning cross-modal representations between text and 3D human motion.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

1.1.1 Dataset Information

The motion data is provided in the form of sequences of frames, with each frame capturing 3D joint positions. Each frame contains 263 values. HumanML3D adopts a 22-joint skeleton, resulting in 66 float values per frame (i.e., 22 joints × 3 coordinates). The remaining dimensions encode joint rotations, joint velocities, and root-level motion information.

On average, each motion sequence consists of approximately 141 frames. While the motion vectors are stored as flattened arrays, they internally represent structured 3D pose dynamics. The dataset is already normalized and centered for direct use in learning pipelines.

The textual annotations paired with each motion sample provide strong linguistic variation. Each sample includes multiple captions, typically both in raw sentence form and with part-of-speech (POS) tagging to support syntactic analysis.

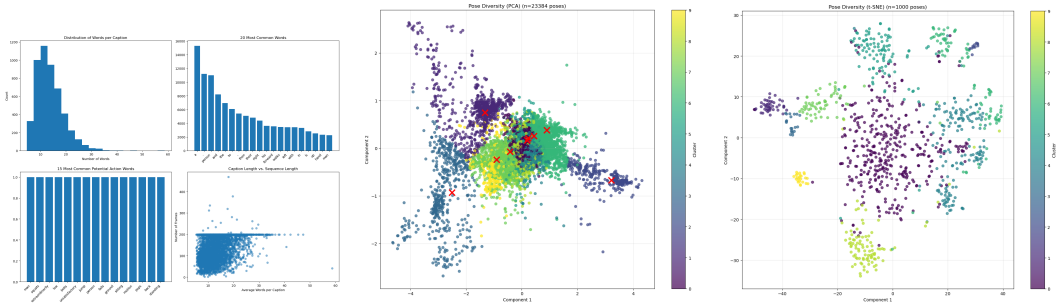
1.1.2 Exploratory Data Analysis

Summary statistics—mean, standard deviation, minimum, and maximum—were computed across all dimensions to understand the numeric range and signal dynamics. This helped identify near-constant dimensions and revealed where the dataset concentrated most of its variance, indicating **which dimensions might dominate learning** or introduce redundancy.

Caption data analysis was done for linguistic study and semantic density. Each motion sample is paired with multiple natural language descriptions, often containing **compound actions or multi-part expressions**. We tokenized the captions, measured word frequencies, and computed average caption lengths. We further analyzed caption content to quantify the frequency of body-part terms (e.g., "hand", "leg", "shoulder") and action verbs. These distributions confirmed that the dataset consistently references core limbs and motions, making it viable for grounded spatial learning.

For motion analysis, we examined intra-frame structure by treating each frame as a 263-dimensional vector and **visualizing dimension-wise statistics**. A detailed correlation analysis across dimensions revealed strong patterns in the first 200 dimensions. A heatmap of inter-dimensional correlations helped confirm the presence of spatially and temporally coupled features. We also detected inactive or padding-like signals.

A significant component of the EDA focused on **pose clustering and diversity analysis**. We applied PCA and t-SNE to reduce the 66-dimensional static pose vectors into 2D latent spaces, enabling visual inspection of pose distribution and structural patterns. K-means clustering was then performed on the reduced embeddings, and each pose was assigned a corresponding cluster ID, which was stored alongside the data. Representative captions were extracted for each cluster to assess semantic consistency within clusters. The resulting dataset package includes raw poses, dimensionality-reduced projections, cluster labels, and the fitted K-means model—forming a foundation for downstream control. This enabled the design of a **non-uniform, cluster-aware sampling strategy** to guide training with structured pose diversity.

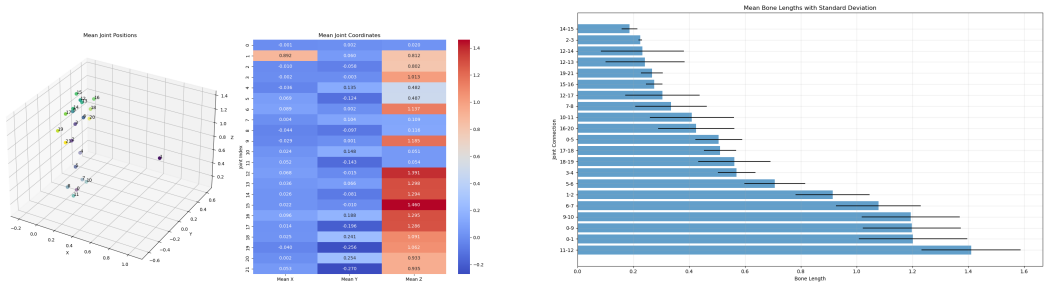


(a) Distributions & Frequency (b) Cluster Plot - PCA (c) Cluster Plot - TSNE

Figure 1: Exploratory Data Analysis Result Plots.

In the final stage of our EDA, we examined skeletal structure and spatial configuration through direct analysis of joint positions. We computed mean lengths, and inter-joint distances within the skeleton hierarchy. When visualized as 3D skeletal graphs, they revealed consistent anatomical proportions.

This skeletal-level EDA was motivated by unexpected visual anomalies observed during early pose sampling—such as **extreme left hip offsets** and spine joints appearing unnaturally close to the floor. Through analysis, we confirmed that HumanML3D adopts a **canonicalization scheme where all bodies face the +Y axis and the left side aligns with +X**, resulting in large static offsets in root-relative coordinates. Poses are expressed in a body-centric frame rather than world space, emphasizing local joint relationships over absolute placement—a critical detail when interpreting spatial configurations or generating 3D visualizations.



(a) Mean Joint Positions

(b) Mean Bone Lengths

Figure 2: Skeletal analysis plots showing average joint positions and inter-joint bone lengths across the dataset.

2 Related Work

In developing our text-to-pose diffusion framework, we drew inspiration from several pivotal works that integrate cross-attention mechanisms within UNet architectures to enhance semantic alignment in generative models. For instance, StableMoFusion[2] employs a Conv1D UNet combined with linear cross-attention to condition motion generation. Latent Diffusion Model (LDM)[3] framework, utilized in Stable Diffusion, incorporates cross-attention within its UNet backbone to condition image generation.

MotionDiffuse [7] was a core study, it integrates text features into a diffusion-based motion generation pipeline via cross-attention, facilitating the synthesis of human motions. Cross-attention is valued high in bridging textual semantics with motion generation, informing our design choices.

T2P (Text-to-Pose) [1] explores the isolated generation of static human poses from free-form text. One core insight from this work is the emphasis on anatomical realism even in non-sequential pose synthesis. They employ **classifier-free guidance** and **sampling techniques** from the literature.

Motion Diffusion Model (MDM) [5] is quite popular, and has a fully transformer-based architecture for human motion generation using a denoising diffusion probabilistic framework. A key contribution of MDM is its unified formulation that removes the need for explicit recurrent or convolutional pose encoders, instead relying entirely on transformer blocks with masked denoising during training.

Recent works in text-to-motion generation, such as T2M-GPT[4] and MotionCLIP[6], showcase other techniques in bridging the gap between text understanding and pose generation.

3 Methodology

3.1 Overview

The generation process begins by encoding a textual description into a fixed-length semantic embedding using CLIP, which serves as the conditioning signal throughout the diffusion process. A random noise vector in the pose space is iteratively denoised by a UNet-based model, guided at each timestep by the text embedding. To integrate linguistic context, the model introduces cross-attention layers that fuse pose features with the text representation at multiple spatial scales. Rather than relying on full token sequences, the conditioning operates over CLIP's global text embedding, projected into the attention layers to steer the generative trajectory.

Training Flow

1. Encode input text using CLIP $\rightarrow$ obtain text embedding.
2. Sample clean pose x_0 from dataset.
3. Add noise to x_0 using diffusion scheduler $\rightarrow x_t$.
4. Feed x_t, timestep t, and projected text embedding into UNet.
5. UNet predicts noise $\hat{\epsilon}$ via residual blocks + cross-attn, with tanh-based clamping.
6. Compute MSE between $\hat{\epsilon}$ and true noise; add anatomy loss on denormalized poses.
7. Backpropagate total loss and update parameters.

Generation Flow

1. Encode user prompt using CLIP $\rightarrow$ project text embedding.
2. Initialize x_T as pure Gaussian noise.
3. For $t = T \rightarrow 0$:
   a. Predict noise $\hat{\epsilon}$ via UNet.
   b. Apply classifier-free guidance: $\epsilon_{cond} = \epsilon_{uncond} + s \cdot (\epsilon_{text} - \epsilon_{uncond})$
   c. Use scheduler to compute x_{t-1} from x_t and $\hat{\epsilon}$.
4. Apply FK-based anatomy constraints with reference bone lengths.
5. Output: semantically aligned, anatomically plausible pose x_0.

Figure 3: Side-by-side overview of the training and generation flows in the text-to-pose diffusion model. Abbreviations: cross-attn (cross-attention), FK (forward kinematics).

3.2 Intuition Behind the Methods Used

Our approach aims to tackle a fundamental challenge in text-to-pose synthesis: translating abstract language into physically valid human configurations.

Unlike systems that access token-level text features, we use CLIP's global 512-dimensional text embedding—a design decision balancing semantic richness against computational efficiency. While this **single-vector representation** limits fine-grained word-to-joint correspondences, our multi-head cross-attention mechanism still enables the model to extract different aspects of the same global semantic embedding. Each attention head specializes in interpreting distinct pose attributes (body orientation, limb positions, action dynamics) from this unified semantic representation. We complement this learning-based bridge with explicit anatomical constraints derived from statistical bone-length analysis.

3.3 Algorithms

3.3.1 Dataset Preprocessing and Strategic Sampling

Our implementation begins with preprocessing of the HumanML3D dataset to optimize for pose diversity and semantic clarity. Rather than uniform random sampling, we employ a **cluster-based selection strategy** that ensures balanced representation across the pose distribution space. This approach addresses an inherent challenge: the dominance of common standing poses that would otherwise bias generation toward neutral stances.

A critical pre-processing step involves a **first-action syntactic segmentation** algorithm. Since the dataset captions describe motion sequences with multiple actions, we leverage part-of-speech tags to identify conjunction markers ("and", "then", "while") and extract only the first meaningful action.

3.3.2 Visualization Algorithms

For 3D pose visualization, we developed a specialized rendering pipeline. The most significant obstacle was correctly interpreting the dataset's orientation scheme, **where bodies face the +Y axis with the left side aligned with +X**. Our solution used coordinate transformation matrices: a 90° rotation about the X-axis (Rx) for skeletal alignment, with 180° yaw rotation (Rz) to orient avatars toward the camera.

We also maintained canonical skeleton mapping functions and standardized joint labeling. To understand unexpected visual anomalies, we implemented **root pelvis centering** (explicit subtraction of the root joint coordinates), anchoring the entire skeleton in a consistent reference frame. We further enhanced the system with a directional orientation arrow anchored at the spine, clearly indicating the avatar's forward direction. These visualization algorithms proved critical for diagnosing early issues.

3.3.3 Pipeline Overview

Resuming the modular pipeline, the architecture consists of four key components: (1) a CLIP-based text encoder that converts natural language into fixed-length semantic embeddings, (2) a UNet-based conditional diffusion model with cross-attention mechanisms, (3) a specialized noise scheduler optimized for pose space, and (4) an anatomical constraint system that enforces physical plausibility.

- Encode the input text using **CLIP** to obtain a semantic embedding (fixed-length).
- Initialize a random noise vector, pass it through a **UNet-based denoising model**.
- Perform denoising over multiple timesteps, each guided by the **CLIP embedding**.
- Encode **timestep information** via sinusoidal embeddings and inject it into UNet blocks.
- Apply **cross-attention** at selected layers to fuse pose features with the **text representation**.
- CLIP's embedding is the conditioning vector $\rightarrow$ projected to match attention dimensions.
- After denoising, apply **anatomy enforcement via forward kinematics**
- Ensure structural plausibility.

3.3.4 Pose Embedding via CLIP and Projection Layer

The text encoding system uses a frozen CLIP model (ViT-B / 32) to transform descriptions into dense semantic representations. The **freezing of CLIP weights is intentional** - preserving the model's understanding while isolating the learning task from the diffusion model and projection layer. The process is straightforward: first cleaning to remove any POS-tagging markers from the dataset, then encoding through CLIP's transformer, and finally L2-normalizing the output embeddings.

A **learned projection layer** then transforms this embedding from CLIP's 512-dimensional space to the model's 256-dimensional working space. This is domain adaptation, bridging the gap between CLIP's text-image space and our pose-specific modeling space.

3.4 Core Diffusion Model Architecture

3.4.1 Phase 1: Baseline Diffusion Implementation

Our initial implementation was experiments, for unconditioned pose generation with a vanilla UNet architecture. Despite achieving relatively low diffusion loss values, generated poses displayed catastrophic anatomical implausibility. We identified fundamental flaws: bone length consistency calculations issues, normalized versus de-normalized space issues, and distorted skeletal geometry.

3.4.2 Phase 2: Anatomical Awareness and Structural Optimization

The second phase introduced architectural improvements focusing on **anatomical realism**:

1. **Normalization-Aware Loss Calculation**: Rewrote the bone length consistency loss to operate on de-normalized poses, correctly preserving skeletal proportions.
2. **Reference Anatomy Repository**: Computed median bone lengths from the entire dataset in normalized space.
3. **Hierarchical Constraint System**: Implemented a kinematic chain-based constraint mechanism that enforced bone lengths through parent-to-child propagation.
4. **Output Stabilization**: Added tanh-bounded output projection with scaled activation to constrain generated values within reasonable physical ranges.
5. **Enhanced Time Embedding**: Improved the sinusoidal time embedding by increasing its dimensionality (from dim/4 to dim/2), providing richer temporal signal.
6. **Balanced Loss Weighting**: Implemented a careful weighting system for the anatomy loss, balancing anatomical correctness against the primary diffusion objective.
7. **Comprehensive Diagnostic Monitoring**: Diagnostic tracking system to monitor critical metrics during training: per-joint position ranges, bone length variance, anatomical losses across different body parts, gradient norms, and noise prediction accuracy.

These modifications established anatomical plausibility as a core objective, improving pose quality by a lot, even without semantic conditioning.

3.4.3 Phase 3: Text Conditioning via Cross-Attention

The final phase transformed the architecture into a text-conditioned generation system:

1. **CLIP Integration**: Dimensional text embeddings as a global conditioning signal.
2. **Skip Connection Enhancement**: Redesigned skip connections to correctly handle both feature and conditioning information.
3. **Balanced Multi-Resolution Architecture**: Designed the UNet with progressive channel multiplication across levels.
4. **Cross-Attention Mechanism**: Implemented multi-head cross-attention modules throughout the UNet architecture, allowing pose features to selectively attend to different aspects of the text embedding.
5. **Conditioned Residual Blocks**: Redesigned residual blocks to incorporate both time and text conditioning, with a sequential integration pattern: normalization, time embedding, cross-attention, and second normalization.
6. **Architectural Bridging**: Replaced conventional GroupNorm with LayerNorm throughout the network for consistent tensor shape handling, addressing dimensional mismatches between CNN and transformer paradigms.
7. **Dual-Path Classifier-Free Guidance**: Implemented a classifier-free guidance mechanism that performs two parallel forward passes: one with the text embedding and another with a null embedding. We also temporarily disable attention tracking during the unconditional pass, then stores attention weights only during the conditional pass.

Conditioned UNet Flow

1. Receive input: noised pose x_t, timestep t, and CLIP text embedding.
2. Encode timestep t using sinusoidal encoding $\rightarrow$ MLP (dim/2) $\rightarrow$ time embedding.
3. Project CLIP embedding $\rightarrow$ reused across UNet levels.
4. UNet forward pass through downsampling blocks:
   a. Each residual block receives time + text conditioning.
   b. Cross-attn applied to queries from pose features.
5. Mid-block applies additional residuals and cross-attn.
6. Upsample symmetrically, injecting skip connections.
7. Final projection: LNorm $\rightarrow$ SiLU $\rightarrow$ Linear $\rightarrow$ Tanh $\rightarrow$ 1.5× scaling.

Conditioned Residual Block Flow

1. Input: hidden state + time embedding + text embedding.
2. First LNorm $\rightarrow$ SiLU $\rightarrow$ Linear projection.
3. Add time embedding via learned projection and SiLU activation.
4. Apply LNorm $\rightarrow$ Cross-attn with text embedding as context.
5. Second LNorm $\rightarrow$ SiLU $\rightarrow$ Linear projection with dropout.
6. Add residual connection (with projection if dimensions change).
7. Output: conditioned hidden representation for next layer.

Figure 4: Architectural-level flows of the modified UNet and Residual Blocks used in the diffusion model. Abbreviations: LNorm (LayerNorm), cross-attn (cross-attention), MLP (multi-layer perceptron).

The final architecture bridges text understanding with structural generation, maintaining physical plausibility. Cross-attention selectively injects semantic guidance while anatomical constraints ensure physical realism.

4 Experiments and Results

The anatomy, diffusion, and final losses across all epochs can be observed in Table 1 for all phase-wise results.

Table 1: Comparison of Model phases for Text-to-Single-Pose Generation

Metric	**phase 1**	**phase 2**	**phase 3**
Final Loss	1.52×10^{15}	1.17	0.69
Diffusion Loss	0.96	1.10	0.63
Anatomy Loss	4.37×10^{16}	0.16	0.07
Input Pose Range	$[-0.88, 1.00]$	$[-0.65, 1.00]$	$[-0.57, 1.00]$
Noisy Pose Range	$[-3.72, 3.45]$	$[-4.49, 5.04]$	$[-4.64, 5.39]$
Predicted Noise Range	$[-1.47, 1.03]$	$[-1.32, 0.99]$	$[-2.74, 2.96]$
Estimated Pose Range	$[-149.06, 196.20]$	$[-4.63, 4.53]$	$[-4.68, 5.35]$
Text Conditioning	No	No	Yes
Anatomy Enforcement	Basic	Multi-stage	Multi-stage + FK
Guidance Scale	N/A	N/A	7.00

4.1 Phase 1: Baseline Diffusion Implementation

- Severe convergence problems, Anatomy loss exceeded 1e+16, extreme instability.
- Noise prediction showed moderate success.
- Despite reasonable noise prediction, total loss remained unworkable.
- Produced poses had disconnected limbs, lacked structural coherence, anatomical plausibility, and had extreme outliers.

4.2 Phase 2: Anatomical Awareness and Structural Optimization

We see the results of upgraded elements with anatomical awareness working. The focus on preserving skeletal proportions, usage of reference lengths, and other methods were key to ensuring that the generated poses were realistic, better than before.

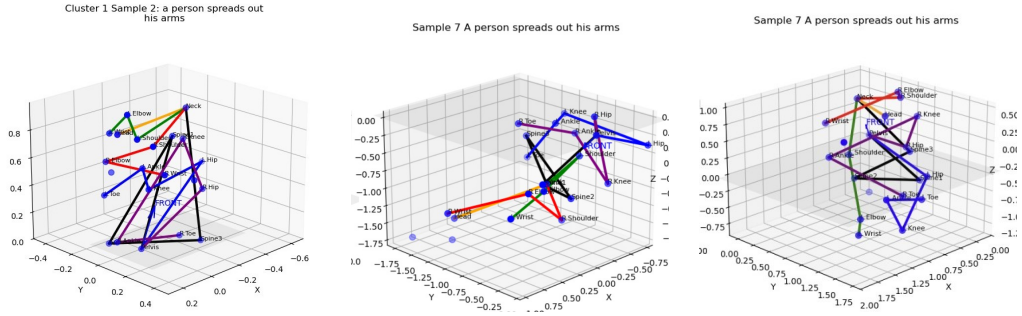
- Diffusion Loss Stability: The model achieved consistent diffusion loss values between 1.04-1.07, indicating successful noise prediction convergence.
- Loss Composition: Total loss was appropriately dominated by diffusion loss, as expected in unconditional denoising.
- Visual Observations: Significant improvements, but several limitations. Anatomically plausible but semantically random poses.
- Model does produce poses with asymmetric limb lengths and positioning between left and right sides. Some generated poses tend toward common dataset positions rather than capturing the full diversity (biased learning).

4.3 Phase 3: Text Conditioning via Cross-Attention

Following the aforementioned techniques in methodology, we saw notable improvements in generating text-relevant poses.

- Poses had stable structural integrity, no degenerate quality, no implausible configurations.
- Maintained contiguous skeletons, had realistic joint lengths and relationships.
- Text conditioning mechanism demonstrated semantic understanding, producing distinct postures for different descriptions.
- Nearly all samples (especially "arms raised," "golf swing," and "crouching"), the predicted poses reflect clear semantic alignment with the associated text.

- A good range of postural diversity across the examples—from static upright poses to dynamic bent-knee, crouch, and limb-extended postures. This indicates the model is not collapsing to mean poses, and is sampling from meaningful distributions.
- The joint connectivity structure remains intact in most cases. Torso-limb anchoring, bilateral leg formations, and spine linkage are preserved. That suggests the model retains the underlying kinematic skeleton. It is **topologically consistent.**



(a) Ground truth pose (b) Generated pose (early) (c) Generated pose (final)

4.3.1 Limitations:

The model exhibits clear limitations. We observed **misaligned limbs** in several samples (e.g., 2, 4), with femur and forearm distortions—highlighting that **forward kinematics alone is insufficient**.

Instances of **semantic ambiguity** suggest the model struggles with precise limb gestures, such as hand placement or fine articulation. Many outputs also show unstable Z-axis positioning, pointing to a **3D spatial reasoning gap**, likely due to CLIP's 2D-projected semantics rather than true 3D-aware conditioning.

These issues reflect a broader problem of **limited expressiveness**, requiring deeper architectural refinement to achieve precise anatomical and semantic alignment.

5 Conclusion

Our work bridges the gap between language and physical embodiment through a focused exploration of text-to-pose generation. Beyond technical achievements, this research demonstrates that even simplified models can grasp the fundamental relationship between descriptive language and human posture. The progression from chaotic poses to anatomically sound, semantically aligned figures reveals how critical structural understanding is when translating abstract concepts into physical form. While perfect semantic alignment remains distant, our system provides a foundation for intuitive human pose creation accessible to non-experts.

This work strips away unnecessary complexity to address a fundamental question: can machines understand how language translates to physical stance? Our results suggest they can, offering a pathway to more natural interactions between humans and embodied AI systems.

6 Future Work

Future development should focus on four key directions. First, we propose investigating **finer-grained conditioning mechanisms** that incorporate token-level attention, allowing body parts to selectively attend to relevant words rather than relying on global embeddings. Second, going **beyond first-action** extraction method for frames, would be a promising approach. Third, adapting the model to diverse conditioning signals, including visual references, skeletal keypoints, and partial pose completions as some advanced models do. Finally, **extending beyond static frames** to multi-pose synthesis with coherent transitions would be a truly powerful goal. However, this may require lots of memory, and computation.

References

[1] Clément Bonnet, Ariel N. Lee, Franck Wertel, Antoine Tamano, Tanguy Cizain, and Pablo Ducru. From text to pose to image: Improving diffusion model control and quality, 2024.

[2] Yiheng Huang, Hui Yang, Chuanchen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework, 2024.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[4] Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022.

[5] Guy Tevet, Sigal Raab, Brian Gordon, Yaron Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022.

[6] Jianhui Zhang, Yixin Liu, Xuehan Li, Stella X. Yu, and Zhiyong Luo. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14730–14740, 2023.

[7] Mingyuan Zhang, Ran He, Haibo Li, and Zhenan Sun. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024.