Università di Genova
DIBRIS DIPARTIMENTO
DI INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

How Credible Are Movements in Synthetically Generated Humans?

Erfan Fathi

Master Thesis

Università di Genova, DIBRIS Via Opera Pia, 13 16145 Genova, Italy
https://www.dibris.unige.it/

MSc Computer Science
Data Science and Engineering Curriculum

How Credible Are Movements in Synthetically Generated Humans?

Erfan Fathi

Advisor: Nicoletta Noceti, Francesca Odone Examiner: Manuela Chessa

March, 2026

Table of Contents

Abstract

Evaluating physical plausibility in synthetically generated human motion remains a critical challenge, as standard evaluation pipelines typically rely on visual judgments and high-level feature distances. This thesis investigates the credibility of synthetic walking motions by proposing a quantitative framework grounded in classical gait analysis. Using a state-of-the-art text-to-motion model, synthetic sequences were generated and compared directly against authentic human motion capture recordings. The methodology extracts a comprehensive suite of physical markers, including step length, cadence, center-of-mass acceleration, and ankle kinematics. A Logistic Regression classifier trained solely on these features achieved over 98% accuracy in distinguishing real from synthetic walking. This demonstrates that despite visual realism and semantic alignment, generated motions retain statistically significant deviations from authentic human locomotion. Additionally, experiments reveal that generative models respond more effectively to qualitative descriptive prompts than strict numerical constraints. Ultimately, this work provides a reproducible methodology for benchmarking the true physical credibility of synthetic human motion.

Chapter 1

Introduction

The synthesis of human motion from natural language descriptions is a complex, high-dimensional challenge that has seen rapid advancement due to the rise of deep generative models. State-of-the-art architectures, including Variational Autoencoders (VAEs) [KW13], Generative Adversarial Networks (GANs) [GPAM$^+$14], Diffusion Models [HJA20], and Autoregressive Transformers [VSP$^+$17], have significantly improved the ability to produce highly realistic and diverse motion sequences from discrete text tokens. Modern text-to-motion models can now successfully capture the semantic essence of an input prompt while stochastically generating temporal pose trajectories [GZZ$^+$22a, TRG$^+$22], enabling the controlled generation of fundamental behaviors such as walking.

However, evaluating the true quality and physical plausibility of these synthetically generated motions remains a critical bottleneck in the field. Standard evaluation pipelines typically rely on quantitative distance metrics such as R-Precision for semantic consistency, or Fréchet Inception Distance (FID) for distribution-based realism [GZZ$^+$22a, HRU$^+$17], alongside qualitative human observation. While a generated motion may successfully satisfy basic geometric constraints and visually align with its textual description, it can still harbor imperceptible but statistically significant physical anomalies. Humans are sensitive to obvious artifacts like foot-skating or severe balance loss, but subtle violations in temporal rhythm or center-of-mass acceleration often go undetected in standard benchmarks. As highlighted by recent comprehensive reviews [ZMR$^+$23], there is a pressing need to incorporate precise physical and kinematic evaluations to truly assess the naturalness of generated movements.

This thesis addresses this gap by posing a fundamental question: How credible are the movements of synthetically generated humans? Our objective is to move beyond subjective visual judgments and high-level feature distances, substituting them with a framework that quantifies spatial, temporal, and dynamic descriptors of human locomotion.

To achieve this, this thesis employs a state-of-the-art text-to-motion generative model [GZZ+22b] to synthesize a broad dataset of walking sequences simulating different conditions. These synthetic motions are then directly compared against authentic human motion capture recordings sourced from a large-scale full-body gait dataset [VCST+23]. We extract a comprehensive suite of classical biomechanical markers, including geometric foot placement (step length and width), temporal pacing (cadence), and dynamic indicators (instantaneous sacrum velocity, center-of-mass acceleration, and foot-level kinematics).

Our empirical analysis reveals that while current generative models produce visually plausible and semantically diverse motions, they retain detectable statistical signatures that separate them from real data. Specifically, a simple linear classifier can differentiate real from synthetic walking with high accuracy based solely on fundamental gait statistics. Beyond these classification results, this work systematically investigates how different textual conditioning strategies influence the biomechanical fidelity of the output. Ultimately, this thesis provides a structured, reproducible methodology for diagnosing artifacts and benchmarking the physical credibility of synthetic human motion, offering broader insights into the control mechanisms and limitations of current generative architectures.

The chapter 2 of this thesis provides the theoretical background and state of the art, reviewing the primary families of deep generative models used for motion synthesis, standard evaluation metrics, and relevant motion datasets. Chapter 3 details the core methodology, including the text-to-motion generative framework, the extraction of specific biomechanical gait features, and the statistical classification protocol designed to objectively evaluate motion realism. Finally, Chapter 4 presents the experimental evaluation and analysis, assessing the generative variability, semantic consistency, and temporal regularity of the synthesized sequences, while also comparing the efficacy of numerical versus descriptive prompting strategies.

Chapter 2

Background and State of the Art

This chapter provides the conceptual and practical background required for applying text-conditioned human motion synthesis. We first review the main families of deep generative models that underpin modern motion generation systems, emphasizing the modeling choices that matter for sequential, high-dimensional, and physically constrained data such as human movement. Because motion generation is inherently one-to-many, multiple distinct motions can legitimately satisfy the same textual description, we focus on how different generative paradigms represent uncertainty, promote diversity, and enforce realism.

We then summarize evaluation methodology for text-to-motion models. Since no single metric captures semantic alignment, physical plausibility, temporal coherence, and perceptual naturalness at once, the chapter highlights commonly used quantitative metrics (retrieval-based and distance-based), learned perceptual scores, and human evaluation protocols, along with their typical failure modes.

Finally, we overview the motion datasets relevant to this thesis, clarifying the type and granularity of supervision they provide (action labels vs. free-form captions; sequence-level vs. frame-level annotations) and how these choices affect both training and evaluation. This structure sets up the methodological decisions used in the remainder of the thesis: selecting appropriate generative backbones, and assessing generated motion credibility via reproducible motion, and gait-based analyses in addition to qualitative inspection.

2.1 General Descriptions

This section introduces the main families of deep generative models used in practice. We start from autoencoders (representation learning), then introduce structured latent-

variable generation via VAEs, realism-driven adversarial training via GANs, diffusion models, and finally sequence-to-sequence generation via autoregressive Transformers.

2.1.1 Introduction to Generative Models

The trajectory of modern generative models is rooted in the early development of Autoencoders, which established a fundamental architecture for learning efficient data representations. While representation learning focuses on compressing and reconstructing observed samples, modern generative modelling additionally aims to produce novel, plausible samples by enabling meaningful sampling in a learned latent space.

Recent advancements in deep learning have led to powerful deep generative models capable of synthesizing highly realistic and complex data, such as human motion. The state-of-the-art in motion synthesis is now underpinned by models that explicitly regularize the latent space, or model sequential probability directly, including:

1. **Variational Autoencoders (VAEs):** Introduce a probabilistic approach to the latent space, forcing the latent distribution to conform to a simple prior (e.g., a standard normal distribution), thereby enabling meaningful sampling.
2. **Generative Adversarial Networks (GANs):** Employ a two-player, adversarial training mechanism between a generator and a discriminator to produce highly realistic samples.
3. **Diffusion Models:** A class of models that generate data by gradually reversing a process of adding noise, achieving unprecedented quality in synthesis tasks.
4. **Autoregressive Models (Transformers):** Formulate data generation as a sequential next-step prediction task. By leveraging self-attention mechanisms, these models excel at capturing long-term temporal dependencies and generating highly coherent sequences from discrete tokens.

2.1.2 Autoencoders

Architecture

An Autoencoder is a specialized type of neural network designed for unsupervised learning of efficient data codings. Its core function is to compress an input data instance, $\mathbf{x}$, into a lower-dimensional representation, $\mathbf{z}$ (the latent space), and subsequently reconstruct the original input, $\hat{\mathbf{x}}$, from this compressed form.

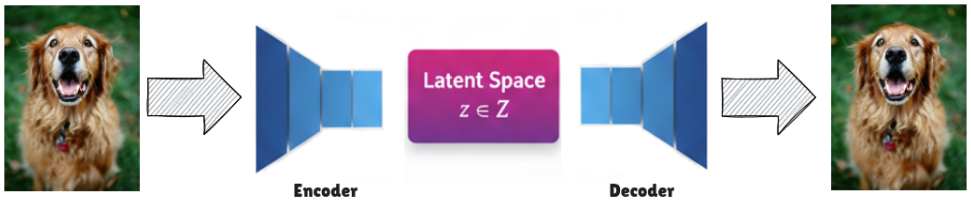

Figure 2.1: Autoencoder architecture showing the encoding of input data through the latent space and subsequent decoding for reconstruction. The encoder compresses the input into a lower-dimensional latent representation $\mathbf{z} \in \mathcal{Z}$, while the decoder reconstructs the original input from this compressed form.

The architecture comprises three key components:

- **Encoder (E):** A function that maps the input $\mathbf{x} \in \mathcal{X}$ to the latent representation $\mathbf{z} \in \mathcal{Z}$, where $\dim(\mathcal{Z}) < \dim(\mathcal{X})$. This achieves dimensionality reduction:

$$\mathbf{z} = \mathbf{E}(\mathbf{x}) \tag{2.1}$$

- **Latent Space ($\mathcal{Z}$):** The bottleneck layer containing the compressed, feature-rich representation $\mathbf{z}$.
- **Decoder (D):** A function that attempts to map the latent vector $\mathbf{z}$ back to the original input space $\mathcal{X}$, producing the reconstruction $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}} = \mathbf{D}(\mathbf{z}) \tag{2.2}$$

The primary objective of training an Autoencoder is to minimize the difference between the input $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$. This is quantified by the reconstruction loss, $\mathcal{L}_{\text{rec}}$, often implemented using the Mean Squared Error (MSE) or binary cross-entropy, depending on the data type.

Training Objective

For continuous data, the reconstruction loss is commonly defined as:

$$\mathcal{L}_{\text{AE}}(\mathbf{x}, \hat{\mathbf{x}}) = \mathcal{L}_{\text{rec}}(\mathbf{x}, \mathbf{D}(\mathbf{E}(\mathbf{x}))) \tag{2.3}$$

Using the MSE for instance:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2 \tag{2.4}$$

Autoencoders are instrumental in various tasks due to their ability to learn meaningful, compact data representations:

Applications

- **Dimensionality Reduction:** Simplifying high-dimensional data for visualization or subsequent processing.
- **Denoising:** Training the AE to reconstruct a clean input from a noisy version (e.g., Denoising Autoencoders).
- **Feature Learning:** Extracting salient features for downstream tasks like classification or clustering.

Limitations for Generation

However, the major limitation of the standard Autoencoder, particularly in the context of generation, lies in the nature of its latent space ($\mathcal{Z}$). The latent space is generally unstructured and deterministic. While the AE learns to map existing data points to a compact region in $\mathcal{Z}$, there is no explicit mechanism to ensure that points sampled randomly within $\mathcal{Z}$ correspond to semantically valid or plausible data instances upon decoding. Generating novel samples thus requires prior knowledge of where meaningful features are encoded, hindering its use as a direct generative model.

This inherent gap, the inability to guarantee the generation of valid data from random latent space samples, catalyzed the transition to models that could impose structure and regularity on the latent distribution.

Autoencoders are excellent for representation learning (encoder–latent–decoder), but their unstructured latent spaces make straightforward sampling unreliable for generation, a key motivation for VAEs.

2.1.3 Variational Autoencoders

Probabilistic Formulation

Variational Autoencoders (VAEs) [KW13] constitute one of the core foundations of modern generative modelling. In contrast to conventional autoencoders, which learn deterministic mappings that compress and reconstruct input data, VAEs adopt a probabilistic formulation that aims to learn the underlying distribution of the data itself. This probabilistic perspective enables the generation of novel samples that are not exact replicas of the training instances but instead reflect the broader structure of the learned data distribution.

At their core, VAEs address a fundamental problem in generative modelling: how to represent complex, high-dimensional data (e.g., images or motion sequences) within a continuous, low-dimensional latent space in such a way that meaningful and coherent new samples can be generated by sampling from this space. As illustrated in Figure 2.2, the VAE framework maps an input datum (x) to a latent representation (z) through an encoder, and subsequently reconstructs an approximation ($\hat{x}$) via a decoder. The key innovation lies in the stochastic "bottleneck" between the encoder and the decoder.

Encoder, Latent Space, Decoder

Encoder Network ($q_\phi(z|x)$)**:** The left portion of the architecture consists of the input data and the encoder network (depicted by the grey and blue blocks in the diagram). The encoder, parameterized by (ϕ), does not output a single deterministic latent vector, but instead predicts the parameters of a probability distribution over latent variables. This is typically a diagonal Gaussian distribution, allowing the model to quantify uncertainty and represent variability in the learned latent space.

Probabilistic Latent Representation: The encoder outputs two key components: the mean vector (μ) (green block) and the log-variance vector ($\log(\sigma^2)$) (purple block). These jointly define the approximate posterior distribution

$$q_\phi(z|x) = \mathcal{N}(z; \mu, \sigma^2 I). \tag{2.5}$$

Estimating the log-variance rather than the variance directly is standard practice, as it ensures numerical stability and preserves positivity of the variance.

Sampling and the Reparameterization: The sampling step (yellow block) constitutes the stochastic bottleneck of the architecture. Direct sampling from $z \sim \mathcal{N}(\mu, \sigma^2 I)$ is not

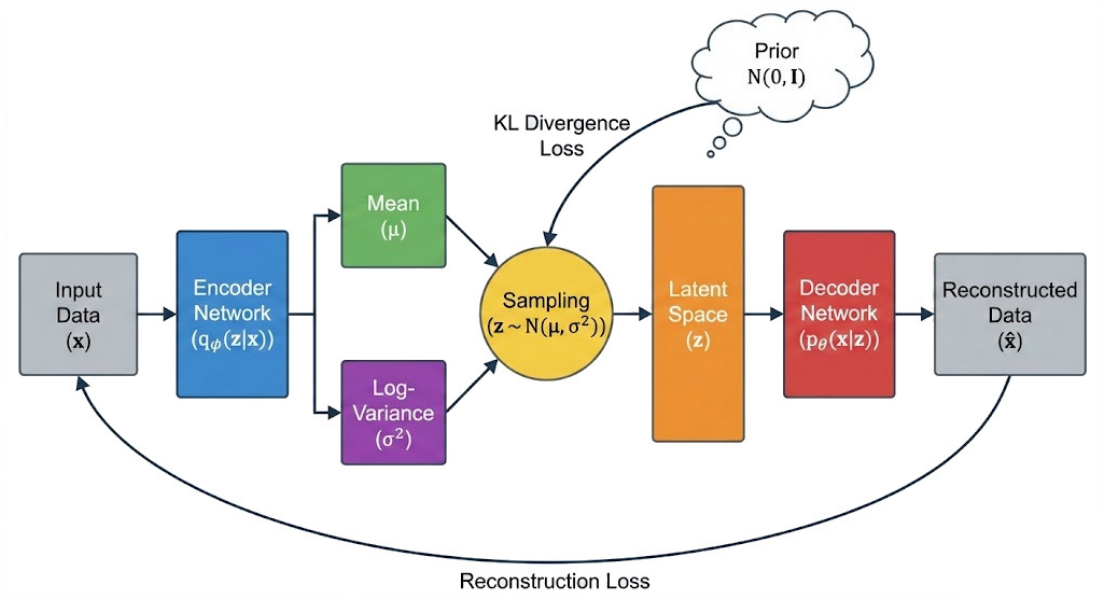

Figure 2.2: VAE Architecture: The diagram illustrates the complete VAE framework, showing the flow from input data through the encoder network to the probabilistic latent space, followed by sampling and reconstruction through the decoder network. The two loss components (reconstruction loss and KL divergence) are highlighted.

differentiable, and would obstruct gradient flow during training. VAEs circumvent this limitation through the reparameterization trick, which expresses the sampling process as

$$z = \mu + \sigma \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{2.6}$$

This formulation isolates randomness in the noise term (ϵ), allowing gradients to propagate through (μ) and (σ) during optimization.

Latent Space (z)**:** The sampled latent vector (z) (orange block) resides in a continuous, structured latent space. A central property of VAEs is that this space becomes smooth and semantically meaningful: interpolations between two latent vectors typically produce valid and coherent outputs, making VAEs particularly valuable for generative tasks requiring controlled variation.

Decoder Network ($p_\theta(x|z)$)**:** The decoder (red block), parameterized by (θ), maps the latent representation back to the data space, producing the reconstructed output ($\hat{x}$). Its objective is to model the conditional likelihood ($p_\theta(x|z)$), effectively learning how to generate realistic data conditioned on latent variables sampled from the learned distribution.

Training Objective (ELBO)

The VAE is trained by maximizing the Evidence Lower Bound (ELBO), which balances reconstruction fidelity with latent space regularization. The total loss is typically expressed as

$$L_{\text{total}} = L_{\text{reconstruction}} + L_{\text{KL}}. \tag{2.7}$$

Reconstruction Loss: The reconstruction loss (illustrated along the bottom curved arrow in the diagram) measures how accurately the decoder reproduces the input from its latent encoding. It is commonly defined as

$$L_{\text{recon}} = -\mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z) \right], \tag{2.8}$$

and instantiated using Mean Squared Error (for continuous data) or Binary Cross-Entropy (for binary data).

KL Divergence Loss: The KL divergence (top curved arrow) enforces regularization on the latent distribution. Specifically, it penalizes deviations of the approximate posterior

($q_\phi(z|x)$) from the prior distribution, typically a standard normal ($\mathcal{N}(0, I)$). The closed-form KL divergence for two Gaussians is

$$D_{\mathrm{KL}}\big(\mathcal{N}(\mu, \sigma^2)\,||\,\mathcal{N}(0,1)\big) = -\frac{1}{2}\sum\left(1 + \log(\sigma^2) - \mu^2 - \sigma^2\right). \tag{2.9}$$

This regularization prevents the model from memorizing data and encourages the formation of a coherent, structured latent space.

2.1.4 Generative Adversarial Networks

Adversarial Learning Principle

A Generative Adversarial Network (GAN) [GPAM+14] consists of two neural networks trained simultaneously in a competitive framework: the Generator and the Discriminator. Conceptually, the generator learns to synthesize samples that resemble the training data, while the discriminator learns to distinguish real samples from synthetic ones. Training proceeds as an adversarial game, where progress in one network creates pressure for improvement in the other.

Architecture and Objective

As illustrated in Figure 2.3, these networks form a coupled adversarial system in which each model pursues an opposing objective.

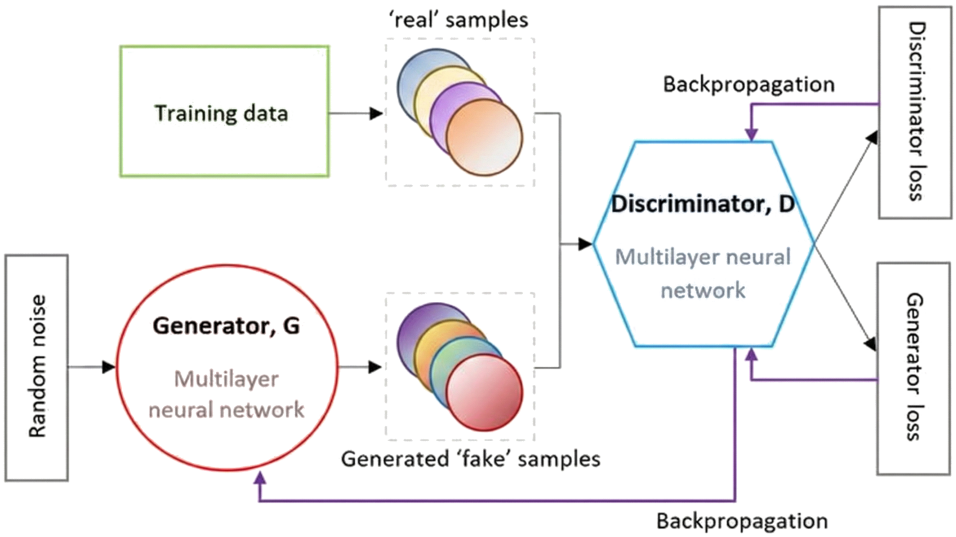

Figure 2.3: GAN Architecture: The Generator (red circle) transforms random noise into synthetic samples, while the Discriminator (blue hexagon) classifies real vs. fake samples. Backpropagation paths (purple arrows) enable adversarial training between the two networks.

The Generator (G): Visually represented by the red circular block in Figure 2.3, the Generator functions as a synthesis network whose purpose is to produce samples that resemble data drawn from the true data distribution. It receives as input a random latent vector $z \sim p_z(z)$, typically sampled from a simple prior such as a Gaussian or uniform distribution. This latent vector is transformed through a multilayer neural network to generate an output sample. Initially, these generated samples bear little resemblance to real data; however, through iterative training, they evolve toward high-fidelity outputs that approximate the structure of the training data.

The Discriminator (D): Depicted as the blue hexagonal block, the Discriminator serves as a binary classifier trained to distinguish between genuine data samples and synthetic samples created by the Generator. It processes two types of inputs:

1. Real samples drawn from the dataset (corresponding to the green box in the figure), and
2. Fake samples produced by the Generator.

Through a multilayer neural network, the Discriminator outputs a probability $D(x) \in [0, 1]$,

representing its confidence that the input sample is real. A value near 1 indicates the sample is classified as real, and a value near 0 indicates it is classified as generated.

Training Dynamics: The training of a GAN proceeds through a coupled optimization cycle in which both networks improve simultaneously by exploiting each other's weaknesses. Figure 2.3 illustrates the data flow and backpropagation pathways.

1. **Forward Pass:** The Generator synthesizes a batch of fake samples, which are combined with real samples and fed to the Discriminator. The Discriminator evaluates each sample and predicts a binary label indicating whether it believes the sample to be real or fake.
2. **Loss Computation:** Two complementary loss functions guide the training:
   - Discriminator Loss: Measures the ability of the Discriminator to correctly distinguish real samples from generated ones.
   - Generator Loss: Measures the Generator's success in fooling the Discriminator, i.e., the degree to which generated samples are misclassified as real.
3. **Backpropagation:** As shown by the purple arrows in the diagram, gradients are propagated back through the networks. Importantly, during the Generator update step, the parameters of the Discriminator are held fixed. This prevents rapid shifts in the decision boundary and stabilizes training by allowing the Generator to optimize against a consistent critic.

Mathematical Formulation: GAN training is formally expressed as a two-player minimax optimization problem. The Generator and Discriminator jointly optimize the value function $V(D, G)$:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \tag{2.10}$$

In this formulation:

- The Discriminator seeks to maximize the likelihood of correctly classifying real samples and rejecting generated ones.
- The Generator seeks to minimize the same objective by producing outputs that cause the Discriminator to assign high probability to synthetic samples.

In practice, the Generator loss is often modified to avoid vanishing gradients early in training. Instead of minimizing $\log(1 - D(G(z)))$, the Generator maximizes $\log(D(G(z)))$, providing stronger gradients that accelerate convergence.

Training Flow Summary: The full system depicted in Figure 2.3 can be summarized as follows:

1. A latent noise vector enters the Generator (red circle).
2. The Generator outputs a synthetic "fake" sample.
3. Real samples are drawn from the dataset (green box).
4. The Discriminator (blue hexagon) processes both real and fake samples.
5. Discriminator and Generator losses are computed (right-hand side).
6. Backpropagation updates each model: the Discriminator improves its discriminative ability, and the Generator learns to reduce the discrepancies that the Discriminator detects.

This adversarial loop continues until the Generator produces samples sufficiently realistic that the Discriminator can no longer reliably distinguish them, converging toward a 50% classification probability for real vs. fake inputs.

2.1.5 Diffusion Models

Forward and Reverse Processes

Diffusion Models [HJA20] have emerged as one of the most influential classes of generative models in contemporary machine learning, forming the foundation of state-of-the-art systems such as DALL-E 3, Stable Diffusion, and Midjourney. In contrast to Generative Adversarial Networks (GANs), which rely on an adversarial game between a generator and discriminator, and Variational Autoencoders (VAEs), which learn latent probabilistic representations, diffusion models are inspired by thermodynamic diffusion processes. Their central idea is to gradually transform complex data distributions into simple noise distributions and to learn an inverse transformation capable of reconstructing high-fidelity samples.

Overview of the Diffusion Trajectory: The diffusion framework consists of two complementary stochastic processes, illustrated in Figure 2.4: a forward diffusion process that incrementally corrupts data with Gaussian noise, and a reverse denoising process that attempts to invert this corruption.

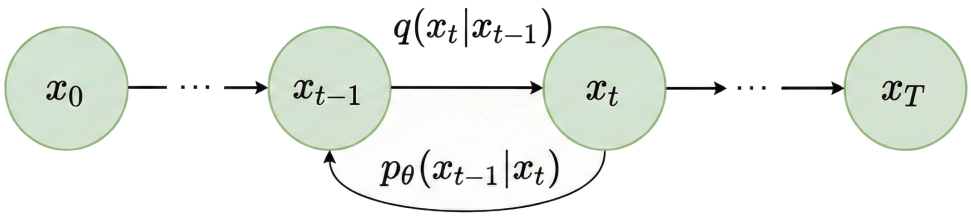

Figure 2.4: The diffusion process showing the forward trajectory (left to right) that gradually adds noise to transform data x_0 into noise x_T, and the reverse denoising process (right to left) that learns to reconstruct clean data from noise.

The Forward Process (Left to Right): The forward process defines a Markov chain

$$x_0 \to x_1 \to \cdots \to x_T, \tag{2.11}$$

where:

- x_0 denotes the clean data sample,
- x_{t-1}, x_t represent intermediate noisy states,
- x_T approximates standard Gaussian noise.

Each transition is governed by a fixed conditional distribution:

$$q(x_t \mid x_{t-1}), \tag{2.12}$$

which gradually adds noise according to a predefined variance schedule. This process is not learned; it is analytically defined.

The Reverse Process (Right to Left): Generation is performed by reversing the diffusion trajectory:

$$x_T \to x_{T-1} \to \cdots \to x_0. \tag{2.13}$$

Since the true reverse distribution $q(x_{t-1} \mid x_t)$ is intractable for real data, it is approximated using a neural network:

$$p_\theta(x_{t-1} \mid x_t), \tag{2.14}$$

parameterized by θ. The network learns to reconstruct a slightly denoised state from its noisier predecessor, effectively serving as a learned denoising operator.

The Forward Diffusion Process: The forward process gradually converts samples from the data distribution into Gaussian noise over T steps. At each step, additive Gaussian noise is applied according to a small variance parameter β_t:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right). \tag{2.15}$$

Here:

- β_t controls the noise magnitude,
- $\sqrt{1-\beta_t}$ ensures that the signal energy decreases smoothly.

Closed-Form Reparameterization: Rather than simulating all intermediate states sequentially, one can directly sample x_t from x_0 using the closed-form expression:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{2.16}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

This formulation is essential for efficient training, as it allows the model to sample noisy inputs at arbitrary timesteps without performing the full diffusion trajectory.

The Reverse Denoising Process: The reverse process is defined as a learned Markov chain with transitions of the form:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right), \tag{2.17}$$

where the neural network estimates both the mean and the variance of the reverse distribution.

A key insight is that training becomes substantially more stable when the network is trained to predict the noise component ϵ rather than the denoised sample directly. Given an estimate of the noise, the model can reconstruct the clean sample through the reparameterization equation.

Training and Sampling

The training objective is derived from the variational lower bound (ELBO) but is typically simplified to a noise-prediction loss:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}\left[\left|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon; t\right)\right|^2\right]. \tag{2.18}$$

This objective trains the model to infer the noise that was added at timestep t, enabling high-quality denoising during generation.

Algorithm Summary

Training Phase:

1. Sample a clean data point x_0.
2. Sample a timestep t uniformly.
3. Sample random noise ϵ.
4. Generate x_t using the closed-form formula.
5. Predict $\epsilon_\theta(x_t, t)$ using the denoising network.
6. Compute the MSE loss between predicted and true noise.
7. Update network parameters.

Sampling Phase:

1. Initialize $x_T \sim \mathcal{N}(0, \mathbf{I})$.
2. Iteratively denoise:
$$x_{t-1} \sim p_\theta(x_{t-1} \mid x_t). \tag{2.19}$$
3. Continue until x_0 is produced.

2.1.6 Transformers and Autoregressive Models

Sequence Modeling and the Attention Mechanism

While VAEs, GANs, and diffusion models excel at mapping continuous spatial or temporal distributions, Transformers have revolutionized the generation of sequential data by treating it as a sequence of discrete elements. At the core of the Transformer architecture is the self-attention mechanism [VSP+17], which overcomes the bottleneck of earlier recurrent neural networks (RNNs) by allowing the model to process sequences in parallel and explicitly model long-range dependencies.

The self-attention mechanism maps a query and a set of key-value pairs to an output. For a given input sequence, the model computes Query (Q), Key (K), and Value (V) matrices through learned linear transformations. The attention weights are calculated by taking the dot product of the queries with all keys, scaled by the square root of the key dimension (d_k), and passed through a softmax function. This yields a probability distribution used to compute a weighted sum of the values:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.20}$$

This formulation allows the network to dynamically weigh the importance of different temporal elements in the sequence, regardless of their distance from one another.

Autoregressive Generation

Generative Pre-trained Transformers and similar autoregressive architectures leverage the decoder block of the Transformer to generate data sequentially. To apply this paradigm to continuous domains like human motion, the data is commonly mapped into a discrete latent space. This is typically achieved using Vector Quantized-Variational Autoencoders, which compress continuous sequences into a finite vocabulary of discrete "tokens."

Once the data is represented as a sequence of discrete tokens, $x = (x_1, x_2, \ldots, x_n)$, generation is framed as a next-token prediction task. The model learns the joint probability of the sequence by decomposing it into a product of conditional probabilities, predicting the next token in the sequence given the context of all preceding tokens.

Mathematical Formulation

The objective of an autoregressive model is to maximize the likelihood of the data sequence under the model parameters θ:

$$p_\theta(x) = \prod_{i=1}^{n} p_\theta(x_i \mid x_{<i}) \tag{2.21}$$

where $x_{<i}$ represents all tokens preceding position i. During training, this is optimized by minimizing the negative log-likelihood (NLL) loss over the sequence:

$$\mathcal{L}_{\text{AR}} = -\sum_{i=1}^{n} \log p_\theta(x_i \mid x_{<i}) \tag{2.22}$$

This simple objective, combined with scalable self-attention, enables autoregressive models to generate highly coherent, contextually rich sequences. In the context of motion synthesis, it effectively turns text-to-motion into a translation task: translating a sequence of text tokens into a sequence of motion tokens.

2.2 Existing Generation Approaches for Human Motion

A particularly challenging and prominent task is text-to-motion generation, where the goal is to synthesize a plausible, natural, and diverse motion sequence given a natural language description (e.g., "walking briskly then waving"). This problem is complex because human motion is high-dimensional and simultaneously governed by stringent physical and stylistic constraints. A successful generative model must effectively capture the semantic essence of the text while ensuring the output motion is kinematically natural and physically valid. This section summarizes representative model families and design choices that have been applied specifically to human motion generation, with an emphasis on text-conditioned synthesis.

2.2.1 VAEs for Human Motion

VAEs have played a significant role in the advancement of generative models for human motion. Early text-conditioned methods such as Language2Pose [AM19] adopted deterministic encoder-decoder structures, mapping textual descriptions into a joint embedding

space shared with motion data. While pioneering, these approaches typically produced a single deterministic motion per input text, limiting output diversity.

The introduction of VAE-based models brought stochasticity and diversity to motion synthesis. For instance, Action2Motion [GZW+20] employs a conditional VAE to generate 3D human motion from action labels. By incorporating Lie algebra representations for human skeleton kinematics and GRU-based temporal decoders, the model captures natural motion variations and produces diverse outputs for the same action category.

Transformer-based VAE models further enhanced motion generation quality. The ACTOR model [PBV21] integrates transformer architectures with a VAE latent structure to synthesize non-autoregressive motions conditioned on action classes. This demonstrated that transformer-based sequence encoders paired with probabilistic latent sampling yield high-quality, diverse motion sequences.

Petrovich et al. [PBV22] extended this line of work to natural language conditioning with TEMOS, which incorporates a text encoder into the VAE framework. The text encoder output parameterizes the Gaussian latent distribution, enabling the generation of multiple plausible motion sequences from a single textual description through stochastic sampling. This substantially improves diversity and expressiveness in text-to-motion synthesis.

In summary, VAEs introduce a powerful probabilistic mechanism that enables generative models to produce diverse and realistic outputs by sampling from learned latent distributions. Their incorporation into motion generation, particularly when combined with transformer architectures, has proven highly effective for capturing the stochastic and multimodal nature of human motion.

Compared to deterministic encoder–decoder pipelines, VAEs add controlled stochasticity, which is crucial for producing multiple plausible motions for the same condition.

2.2.2 GANs for Motion Generation

While GANs are widely recognized for producing high-quality, sharp outputs in image domains, their adoption in motion generation has been more nuanced. Early work leveraged adversarial objectives to improve the realism of generated pose sequences. For instance, Text2Action [AHC+17] incorporated a GAN objective atop a recurrent motion generator to enhance the authenticity of short action sequences.

However, applying GANs to high-dimensional sequential data such as human motion remains challenging. Issues such as mode collapse, instability, and difficulty modeling long-term temporal dependencies have limited the prevalence of pure GAN architectures in text-conditioned motion generation. Surveys in the field note that GANs have contributed more substantially to unconditional motion tasks, such as dance or gesture synthesis, than

to text-to-motion frameworks.

A notable example is MoCoGAN [TLYK17], which decomposed motion and content for video synthesis using an adversarial approach. Similar ideas have inspired subsequent work on applying GANs to latent representations of motion.

Overall, GAN-based discriminators can effectively encourage realism by penalizing unnatural or implausible motion patterns. However, they are typically integrated alongside other generative mechanisms, such as VAEs or diffusion models, to ensure that the generated motion not only appears realistic but also aligns semantically with textual input. Conceptually, adversarial training functions as an automatic "Turing test" for motion, with the Discriminator acting as a learned critic of movement plausibility.

In long, high-dimensional sequences (e.g., motion), pure GAN training can be unstable; in practice, adversarial losses are often used as an auxiliary realism term rather than the sole generative mechanism.

2.2.3 Diffusion for Human Motion

Diffusion probabilistic models have recently become the dominant approach for generative motion synthesis, owing to their strong mode coverage, iterative refinement capability, and ability to model complex temporal dynamics. Following the success of diffusion models in vision and language domains, several works in late 2022 successfully adapted diffusion to the domain of human motion generation.

A notable example is the Human Motion Diffusion Model (MDM) introduced by Tevet et al. [TRG+22]. MDM is a classifier-free diffusion framework that operates directly on pose sequences. It employs a transformer-based architecture to learn the reverse denoising process over temporal motion trajectories. Unlike conventional implementations, where the network predicts the noise residual, MDM predicts the pose itself at each diffusion step. This design choice enables the application of physically meaningful constraints directly on human pose parameters during training.

One such constraint is the foot-contact loss, which penalizes implausible foot motion and substantially reduces foot-sliding artifacts, improving physical realism. MDM achieved state-of-the-art performance in text-to-motion and action-conditioned motion generation, surpassing earlier VAE and GAN-based approaches.

Parallel research introduced additional diffusion-based motion generators, including MotionDiffuse [ZCP+22] and several subsequent variants focusing on efficiency, controllability, or conditioning strategies.

Overall, diffusion models yield highly natural and temporally coherent motion sequences,

though they can be computationally expensive and less directly controllable than deterministic models. Recent work addresses these challenges through guided sampling, constraint-based control, and integration with learned motion embeddings.

Because diffusion models provide strong mode coverage and iterative refinement, they are well-suited for temporally coherent and physically plausible motion. In the remainder of this thesis, we use generated motion sequences and assess their credibility using quantitative gait- and motion-based features, complementing qualitative inspection with systematic, reproducible evaluation.

2.2.4 GPT-Based and Autoregressive Models for Motion

While diffusion models and continuous VAEs operate on real-valued representations of human motion, autoregressive models have recently demonstrated that treating motion generation as a discrete sequence prediction task can yield highly semantic and diverse results. This paradigm shift relies heavily on the success of Large Language Models (LLMs), adapting their architecture to the kinesthetic domain.

Motion Tokenization via VQ-VAEs

The fundamental challenge in applying Generative Pre-trained Transformers to human motion is the continuous and high-dimensional nature of kinematic data. To bridge this gap, modern autoregressive motion models typically employ a two-stage pipeline. In the first stage, a Vector Quantized-Variational Autoencoder [vdOVK17] is trained to compress continuous motion sequences into a sequence of discrete latent codes. This effectively creates a finite "vocabulary" of human movement, where each discrete token represents a localized spatio-temporal motion snippet.

Text-to-Motion as Next-Token Prediction

Once the motion is tokenized, the generation process is mathematically reframed as a machine translation task. During the second stage, a Transformer-based autoregressive model (such as a GPT architecture) is trained to predict the next motion token given a textual description and all previously generated motion tokens. By maximizing the likelihood of the token sequence, the model learns the underlying grammar and temporal dependencies of human movement conditioned on natural language.

State-of-the-Art Implementations

Several recent works have successfully implemented this paradigm. T2M-GPT [ZZC+23] demonstrated that a standard GPT architecture paired with a high-quality VQ-VAE can achieve state-of-the-art performance on text-to-motion benchmarks like HumanML3D [GZZ+22a]. By mapping language features directly to discrete motion indices, T2M-GPT generates highly diverse and semantically aligned sequences. Similarly, models like MotionGPT [ZHL+24] treat human motion as a specific "foreign language," allowing them to unify multiple motion-related tasks (such as text-to-motion synthesis, motion captioning, and motion prediction) within a single generative framework.

Autoregressive models provide excellent semantic alignment and sequence diversity by leveraging the powerful sequence-to-sequence capabilities of Transformers. However, because they rely on a discretized latent space, they can occasionally suffer from quantization artifacts or a loss of fine-grained kinematic detail compared to continuous approaches like diffusion models.

2.3 Evaluation Metrics

Quantitative and qualitative evaluation remains a core challenge in text-to-motion generation: a motion can satisfy basic physical constraints yet look unnatural, or align semantically with text while exhibiting perceptual artifacts. Consequently, recent works commonly report multiple complementary metrics, combining automatic quantitative scores, learned perceptual evaluators, and human studies.

2.3.1 Quantitative Metrics and Benchmarks

Quantitative metrics rely on explicit numerical computations, typically involving physical distances, classification accuracy, or task-specific success criteria. These metrics are objective, reproducible, and easy to compute, but often struggle to reflect perceptual realism or semantic adequacy.

Semantic Consistency Metrics

R-Precision: R-Precision is widely used in text-to-motion and action-to-motion tasks to evaluate semantic correspondence between generated motions and their textual descriptions. Rather than comparing joint trajectories directly, it operates in a shared feature space learned by text and motion encoders.

Evaluation logic: for each generated motion, its ground-truth description is grouped with $k-1$ randomly sampled distractor descriptions. Distances (or similarities) are computed between the motion feature and all text features, and candidates are ranked accordingly.

Metric definition: R-Precision reports the Top-k retrieval accuracy: if the correct description appears among the Top-k closest descriptions, the sample is a hit.

Significance: this directly measures semantic alignment, making it more suitable than pure geometric distance for language-conditioned generation.

Multimodal Distance (MM-Dist): MM-Dist evaluates feature-level alignment between motion and text modalities by measuring distances in a learned joint embedding space:

$$\text{MM-Dist} = \frac{1}{N} \sum_{i=1}^{N} \left\| f_{m,i} - f_{t,i} \right\|_2 , \tag{2.23}$$

where $f_{m,i}$ and $f_{t,i}$ denote the motion and text features for sample i.

Significance: lower MM-Dist indicates generated motions are semantically closer to their conditioning text; it is often reported alongside R-Precision as a complementary distance-based view.

Task-Specific Accuracy Metrics

Action recognition accuracy: In action-to-motion settings, generated motions are evaluated using a pretrained action recognition model, which predicts the action label from the generated motion alone:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}}, \tag{2.24}$$

where N_{correct} is the number of motions classified as the intended action.

Significance: provides high-level semantic validation, but is limited by biases and capacity of the pretrained recognizer.

Distance-Based Motion Errors

Distance metrics (MSE, ADE, FDE): Distance-based metrics measure physical deviation between generated joint positions and the ground-truth motion sequence.

- **MSE** penalizes large deviations but is sensitive to temporal misalignment.

- **ADE** measures average joint-wise L_2 error across all frames.
- **FDE** evaluates the error at the final frame only.

A common ADE definition is:

$$\text{ADE} = \frac{1}{T\,J} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| x_{t,j} - \hat{x}_{t,j} \right\|_2 , \tag{2.25}$$

where T is the number of frames, J is the number of joints, $x_{t,j}$ is the ground-truth joint position, and $\hat{x}_{t,j}$ is the generated joint position.

Limitations: while simple and interpretable, these metrics penalize valid alternative motions and are therefore insufficient in isolation for one-to-many generation.

2.3.2 Learned Perceptual Metrics

Learned perceptual metrics leverage deep neural networks to evaluate motion quality in a feature space aligned with human perception. Instead of comparing individual samples in raw joint space, they often compare distributions of real and generated motions in an embedding space, or measure cross-modal consistency using learned representations.

Distribution-Based Realism Metrics

Fréchet Inception Distance (FID): Originally developed for image generation, FID [HRU+17] has been adapted to human motion by computing statistics over learned motion embeddings. Assuming Gaussian distributions for real and generated motion features, the FID is:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\Big). \tag{2.26}$$

Significance: lower FID indicates generated motions are closer to real motions in terms of feature statistics (often interpreted as realism and diversity). Results depend strongly on the choice of feature extractor and evaluation protocol.

Cross-Modal Alignment Metrics

Motion CLIP Score (mCLIP): The Motion CLIP Score evaluates text–motion alignment in a CLIP-style embedding space [RKH+21]:

$$\text{mCLIP} = \frac{f_t \cdot f_m}{\|f_t\|\,\|f_m\|}, \tag{2.27}$$

where f_t and f_m are text and motion embeddings.

Significance: higher cosine similarity indicates stronger semantic alignment between generated motion and input text.

Mutual Information Divergence (MID): MID [KKL+22] measures mutual information between modalities (e.g., text and motion), quantifying how much information from the conditioning signal is preserved in the generated motion.

Significance: unlike pointwise distance metrics, MID captures global multimodal dependence, and can serve as a robust indicator of semantic consistency across modalities.

2.3.3 Human Evaluation

Despite advances in automatic metrics, human evaluation remains indispensable: humans are sensitive to subtle artifacts such as unnatural timing, foot skating, balance issues, or culturally inappropriate gestures.

Pairwise Preference Studies

Protocol: Participants are shown two motion clips (e.g., generated by different models) and asked comparative questions such as:

- "Which motion better matches the text?"
- "Which motion looks more realistic?"

Metric: Results are summarized as Win Rate, i.e., the percentage of trials in which a method is preferred over a baseline.

Elo Rating

Protocol: Pairwise preferences are aggregated across comparisons using an Elo rating system, producing a single scalar rating per model.

Significance: Elo yields a global ranking over multiple models on a continuous scale, reducing inconsistencies that can arise from isolated pairwise comparisons.

Explicit Scoring (Likert Scale)

Protocol: Participants rate generated motions on a fixed scale (commonly 1–5) across multiple dimensions, such as:

- **Quality:** naturalness and realism
- **Diversity:** variation among generated samples
- **Consistency:** alignment with the conditioning signal

Significance: provides fine-grained diagnostic feedback, but is more subjective and sensitive to evaluator bias and experimental setup.

2.4 Available Datasets

This section summarizes the motion datasets used or considered in this work, spanning classic indoor motion-capture benchmarks, large-scale aggregated corpora, and recent motion-language resources. We emphasize, for each dataset, (i) the scale and diversity of subjects and actions, (ii) the underlying motion representation (e.g., skeletal joints, SMPL/SMPL-X parameters, or multi-modal sensor streams), and (iii) the form and granularity of supervision available for semantic conditioning, ranging from fixed action-class names to free-form natural-language captions at sequence level and, in some cases, dense frame-level labels. This overview clarifies which datasets are suitable for purely kinematic training, which enable text-conditioned generation, and which are most appropriate for evaluation under realistic and clinically relevant motion variability.

2.4.1 Human3.6M

Human3.6M [IPOS14] is a large indoor mocap dataset of 3.6 million frames (approximately 3.6 million 3D poses) collected from 11 professional actors (6 male, 5 female) performing 15 everyday activities. The motions include typical actions like walking, sitting, eating, taking photos, smoking, and talking (the original data uses camera and motion-capture synchrony). The poses are provided in SMPL-like skeletal joint format for each frame, and each sequence is labeled by one of 15 action categories (e.g. “Directions”, “Discussion”, “Waiting”, “WalkingDog”, etc.). Thus Human3.6M carries only category labels (action names) rather than free-form language descriptions, but these textual class labels allow it to be used for action-conditioned modeling. Its sheer size (3.6M frames at 50Hz from four viewpoints) and variety of human subjects make it a common benchmark for 3D pose and action modeling.

2.4.2 CMU Motion Capture

The CMU Graphics Lab Motion Capture Database [Car03] is a long-standing free archive of full-body motion trials from over 140 subjects. It contains on the order of 2,600+ trials organized into broad categories (Human Interaction, Interaction with Environment, Locomotion, Physical Activities & Sports, Situations & Scenarios). Motions span diverse activities, e.g. locomotion (walking, running), sports (basketball, soccer), gestures (waving, jumping, pushing), and object interactions, with dozens of sub-category labels for each trial. Each clip comes with only categorical tags or filenames (e.g. "basketball", "washing window"), but no natural-language sentences; nonetheless, these action labels (as indexed in the CMU database) serve as text annotations linking each motion to its semantic type. The data are provided as raw C3D files and skeleton poses (often 38–41 joints), making CMU Mocap a broad, general-purpose motion collection in many motion generation projects.

2.4.3 AMASS

AMASS [MGFT+19] (Archive of Motion Capture As Surface Shapes) aggregates many existing marker-based mocap datasets into a unified SMPL body-parameter representation. In total it comprises over 40 hours of motion data from more than 300 subjects and roughly 11,000 motion sequences. The source motions come from a variety of activities (walking, running, gestures, exercises, dance, etc.), but AMASS itself primarily provides body-shape and pose parameters for each frame. Importantly, AMASS does not include any explicit language descriptions or action labels by default; each sequence is referenced by its original dataset and clip name. (Recent works often overlay annotations on AMASS data, but the core AMASS release supplies only skeletal motion data.) In summary, AMASS is a massive, diverse motion corpus (40h+, 300+ people) in SMPL form, used for training many motion models, albeit without built-in text labels.

2.4.4 HuMMan

HuMMan [CRZ+22] is a very large-scale multi-modal human motion dataset. It contains motion and sensor data from 1,000 distinct human subjects, amounting to about 400,000 action clips (60 million frames total). The subjects performed 500 predefined actions designed to cover fundamental movements (everyday activities, exercise moves, etc.), so each clip is annotated with one of these 500 action category names. The data are highly multi-modal, e.g. synchronized color images, LiDAR point clouds, 3D keypoints, SMPL parameters, and even textured meshes, but for motion modeling the key labels are the action categories. In short, HuMMan's text labels consist of a large vocabulary of 500

action names, one per sequence. The dataset's novelty lies more in its scale and sensor modalities than natural language annotation: it provides many millions of frames of full-body motion across hundreds of people, but only at the level of action names (not free-form descriptions).

2.4.5 Motion-X

Motion-X [LZL+23] is a newly constructed large-scale whole-body motion dataset with rich text labels. It consists of about 81,100 motion sequences (15.6 million SMPL-X frames) collected from online videos, spanning diverse indoor/outdoor scenes. Unlike many prior datasets, Motion-X includes fine details: it models full SMPL-X poses (body + face + hand articulation) for each frame. Crucially, the creators automatically annotated every sequence with semantic text labels. Specifically, each sequence has a short descriptive label (a few words describing the action), and each frame has a fine-grained textual description of the whole-body pose. In other words, Motion-X provides two tiers of language: sequence-level captions (words summarizing the action) and per-frame sentences detailing the posture. These annotations were generated by an automatic pipeline and cover an open vocabulary of human actions and expressions. Thus Motion-X is both vast (15.6M frames) and explicitly language-annotated, enabling supervision at both sequence and frame granularity.

2.4.6 KIT Motion-Language Dataset (KIT-ML)

The KIT Motion-Language (KIT-ML) dataset [PMA16] specifically pairs whole-body motions with natural-language descriptions. It contains 3,911 motion clips (about 11.23 hours total, 111 actors) recorded by a marker-based system, each labeled via crowdsourcing with multiple English sentences. In total there are 6,278 sentence annotations (≈53,000 words) across the clips. Each entry's labels are provided in JSON: every motion ID has an array of free-form textual descriptions (one-to-many). The texts describe what is being done in the clip (e.g. "a person steps forward and picks up a box," etc.), allowing full-sentence conditioning. Thus, in KIT-ML each recorded motion is explicitly linked to one or more English sentences. (Notably, many entries have only one sentence, but up to a few per clip.) This dataset was designed as a benchmark for text-to-motion, providing authentic natural-language annotations for about 4K motion sequences.

2.4.7 NTU RGB+D

The NTU RGB+D benchmarks, including NTU RGB+D 120 (NTU-120) [LSP+19], are very large datasets of video-recorded human actions, captured with Kinect v2. The original NTU-60 set has 56,880 video samples of 60 action classes (40 subjects, 80 viewpoints), and NTU-120 extends this to 114,480 samples over 120 classes (106 subjects, 155 viewpoints). Actions range from daily activities (e.g. drinking, eating, reading) to gestures (e.g. clapping, typing on keyboard) and interactions (e.g. touching head, throwing). The primary annotations are action class names, canonical short phrases like "brushing hair," "kicking with left leg," "putting on shoes," etc. assigned to each clip. These textual labels are not descriptive paragraphs but fixed category phrases. In addition to RGB video and depth, NTU provides 3D skeleton (joint) data. NTU is relevant because its class names serve as semantic targets for text-conditioned models (e.g. generating a "drinking water" motion) and its huge diversity (subjects, views, speeds) tests generalization.

2.4.8 HumanAct12

HumanAct12 [GZW+20] is a curated mocap dataset of human actions derived from the PHSPD collection. It contains 1,191 motion clips (≈90,100 frames total) organized into 12 action categories. These 12 categories are hierarchical action types (e.g. various warm-up exercises, jumping, kicking, etc.), each covering a set of similar motions. Every clip is annotated with exactly one of these 12 action labels. The names of the categories are fairly coarse (e.g. "Warm_up_wristankle", "Warm_up_pectoral"), reflecting groups of related movements. No free-text descriptions are provided; instead the dataset uses these fixed labels so that one can condition generation on a single verb-like label. HumanAct12 is fairly small by modern standards, but was constructed to have a balanced number of examples per action type and to complement NTU and CMU data for model training.

2.4.9 BABEL

BABEL [PCA+21] adds English language labels to motion-capture data in AMASS. It comprises about 43 hours of motion from AMASS, fully annotated with action labels. Importantly, BABEL provides labels at two granularities: sequence-level and frame-level. There are over 28,000 overall sequence labels (one label per motion clip) and about 63,000 fine-grained frame labels (labeling each time segment in the clip, possibly multiple overlapping actions). In total BABEL uses over 250 unique English action categories (e.g. "jump forward", "pick up something", "sit down"), and each frame label is aligned precisely with the action's duration. Thus BABEL offers natural-language descriptions that cover every portion of the motion, a detailed annotation scheme much richer than simple category

tags. Because it is built on AMASS, BABEL's motions are SMPL body poses. BABEL's strength is that it provides dense human-written labels for a large, diverse motion collection (43h of data, 250+ categories).

2.4.10 HumanML3D

HumanML3D [GZZ+22a] is a modern 3D motion-language benchmark that merges AMASS and HumanAct12 motions with crowdsourced descriptions. It has 14,616 distinct motion clips (∼28.6 hours of data) paired with 44,970 English sentence descriptions. Each clip has 3–4 textual annotations collected via Mechanical Turk, so in aggregate the dataset uses a vocabulary of ∼5,371 words. The actions covered are very broad (daily tasks like walking, jumping; sports like swimming, golf; also dance and acrobatics). Thus HumanML3D provides free-form sentence-level descriptions for each motion. For example, one motion might have descriptions like "a man kicks something or someone with his left leg" and "the standing person kicks with their left foot". All motions use a 22-joint SMPL skeleton; the annotations enable training and evaluation of text-to-motion models with rich natural language targets.

2.4.11 Full-Body Gait Dataset

A full-body gait dataset of 138 able-bodied adults and 50 stroke survivors [VCST+23] is a specialized biomechanical reference for human walking. It contains synchronized motion capture, force plate, and EMG recordings of 188 individuals walking at their natural speed. The cohort is very diverse, 138 healthy subjects aged 21–86 (including many over 70) and 50 post-stroke patients (aged 19–85). Full-body kinematics (Plug-in-Gait model), kinetics (ground forces), and 14-channel muscle activity were recorded in a Vicon lab. The data are provided both as raw C3D files and as post-processed stride-normalized MAT/Excel files. We include this dataset because it offers real human motion with high fidelity and demographic variety, ideal for validating generative models on realistic gait. Its annotations are highly detailed (e.g. gait cycle events, full-body joint angles, clinical metadata) and consistent across subjects. In particular, the inclusion of aged and pathological gait makes it valuable as a diverse validation set: generated walking motions can be compared against these ground-truth biomechanical profiles to assess realism across different populations.

Chapter 3

Methodologies

This chapter describes the methodological framework adopted to evaluate the credibility of synthetically generated human walking motions. Our objective is not merely to generate plausible-looking animations, but to rigorously assess whether these motions exhibit biomechanical characteristics consistent with real human gait. To achieve this, we design a structured pipeline that integrates motion generation, feature extraction, dataset harmonization, and statistical evaluation within a unified analytical framework.

The methodology is organized around three core components. First, we employ a state-of-the-art text-to-motion generative model [GZZ+22b] to synthesize walking sequences from natural language prompts. This model produces high-dimensional temporal pose trajectories conditioned on semantic input, enabling controlled generation of forward walking behaviors. The generated motions serve as the synthetic counterpart to real motion capture recordings.

Second, we construct a biomechanically grounded feature extraction framework to characterize gait dynamics at multiple levels of abstraction. Rather than relying solely on perceptual or qualitative evaluation, we quantify spatial, temporal, and dynamic descriptors of locomotion. These include geometric foot placement measures (step length and step width), temporal pacing metrics (cadence), global kinematic indicators (sacrum velocity and stride velocity), acceleration-based smoothness measures (center-of-mass acceleration), and limb-level contact dynamics (ankle velocity and acceleration). Together, these descriptors provide a multi-scale representation of gait behavior, capturing both global forward progression and fine-grained contact mechanics.

Third, we formulate realism assessment as a supervised statistical discrimination task. Real motion capture data are obtained from a large-scale full-body gait dataset [VCST+23], restricted to healthy adult subjects to avoid pathological confounds. Synthetic and real sequences are preprocessed using an identical normalization and feature extraction pipeline

to ensure comparability. Each motion sequence is represented as a fixed-dimensional vector summarizing its biomechanical statistics. A Logistic Regression classifier is then trained to determine whether these features alone are sufficient to distinguish real from generated gait.

This methodological design enables us to move beyond subjective realism judgments and instead quantify credibility in measurable biomechanical terms. If synthetic walking motions are statistically indistinguishable from real human gait in feature space, they can be considered biomechanically credible. Conversely, systematic separability would indicate that, despite visual plausibility, generative models retain detectable deviations from authentic locomotion patterns.

The following sections detail each component of this pipeline, beginning with the generative walking model, followed by the definition of gait features, and concluding with the statistical evaluation protocol.

3.1 Walking Generation Methods

Generating realistic human walking motion synthetically requires a generative framework capable of modeling both the stochasticity and the structured regularities of human gait. In this thesis, we adopt a state-of-the-art text-to-motion synthesis model originally introduced by Guo et al. [GZZ+22b], which produces human motion sequences directly from natural-language descriptions. The central challenge lies in capturing the high-dimensional, time-varying nature of movement while enforcing semantic consistency with the textual prompt.

To address this, the model employs a two-stage probabilistic architecture designed to decouple the prediction of motion duration from the generation of spatial pose trajectories. This separation enables fine-grained control over motion type, such as walking, while ensuring that the synthesized sequence remains faithful to the semantic content of the prompt.

3.1.1 The Two-Stage Probabilistic Framework

Let the input text description be denoted as T, and the resulting motion sequence be represented as $M = \{p_1, p_2, \ldots, p_L\}$, where p_t denotes the pose vector at frame t. The generative process is modeled as a joint probability distribution conditioned on the textual input:

$$P(M, L \mid T) = P(M \mid L, T) \cdot P(L \mid T). \tag{3.1}$$

This formulation decomposes the problem into two sequential modules:

1. **Text-to-Length (Text2Length):** Estimating the motion duration distribution $P(L \mid T)$.
2. **Text-to-Motion (Text2Motion):** Generating the pose sequence from the conditional distribution $P(M \mid L, T)$.

Stage 1: Text-to-Length Prediction

The Text2Length module predicts a plausible motion duration based on the semantics of the prompt. It samples L from the learned distribution $P(L \mid T)$. This mechanism is essential for preserving semantic fidelity, for instance, a description such as "a person takes two steps forward" implies substantially fewer frames than "a person walks in a circle". As a result, the model produces variable-length motion sequences that reflect the intended action, rather than enforcing a fixed-duration output.

Stage 2: Text-to-Motion Generation via VAE

Once L is sampled, the Text2Motion module generates the motion sequence. This component is implemented as a temporal Variational Autoencoder (VAE), which introduces a stochastic latent embedding that enables diversity of motion outcomes.

The latent variable z is sampled from a learned posterior conditioned on the text:

$$\hat{M} = Dec(z, T, L), \quad z \sim \mathcal{N}(\mu(T), \sigma(T)). \tag{3.2}$$

This probabilistic sampling allows the model to produce multiple valid motion trajectories from the same textual prompt, capturing, for example, different gait styles, stride characteristics, or path geometries.

Figure 3.1 illustrates this one-to-many mapping effect. A single prompt (e.g., "a person walks forward") is encoded into the latent space, from which multiple distinct realizations emerge, such as a slow walk, a long-stride walk, or a curved-path walk.

3.1.2 Segment-Level Motion Encoding

To enhance biomechanical realism and temporal coherence, the model does not generate joint configurations frame-by-frame. Instead, it operates on an internal motion snippet code representation, wherein the motion sequence is decomposed into short, overlapping temporal segments.

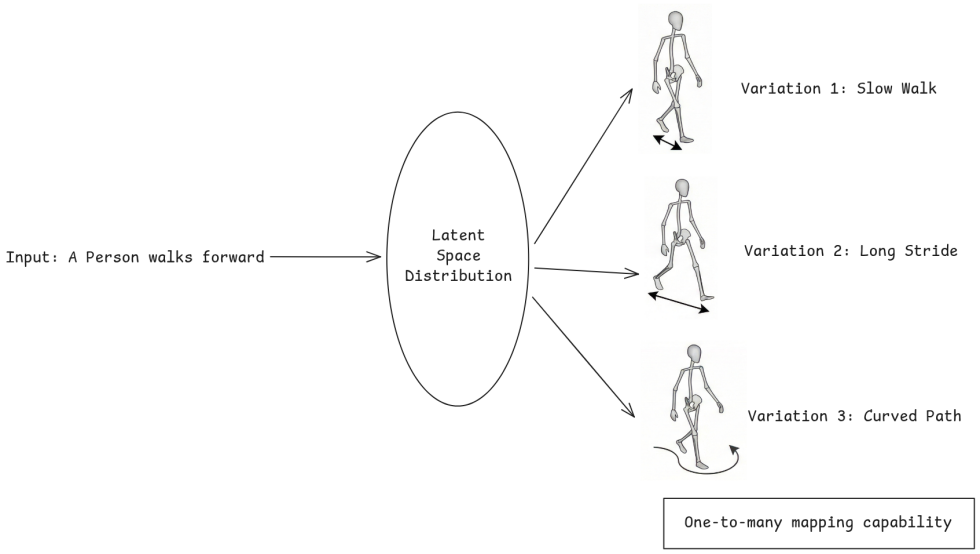

Figure 3.1: Stochastic Diversity in Walking Generation. The latent distribution allows for one-to-many mapping, producing variations like slow walk, long stride, or curved path from a single input prompt.

Let $S = \{s_1, s_2, \ldots, s_N\}$ represent the collection of snippets, each capturing a brief interval of locally coherent movement. The encoder maps each snippet into a compact latent code:

$$c_i = E_{\text{snippet}}(s_i). \tag{3.3}$$

These snippet-level codes serve as the generation targets for the VAE. This representation yields two primary benefits:

1. **Temporal Consistency:** Generating coherent snippets rather than independent frames reduces jitter and enforces smooth transitions.
2. **Local Semantic Fidelity:** Snippet codes encapsulate local motion patterns, such as stride timing and limb coordination, crucial for natural gait generation.

3.1.3 Implementation and Usage

We employ a pre-trained text-to-motion model[GZZ+22b] to synthesize the walking dataset used in our experiments. The system is conditioned on textual prompts describing specific

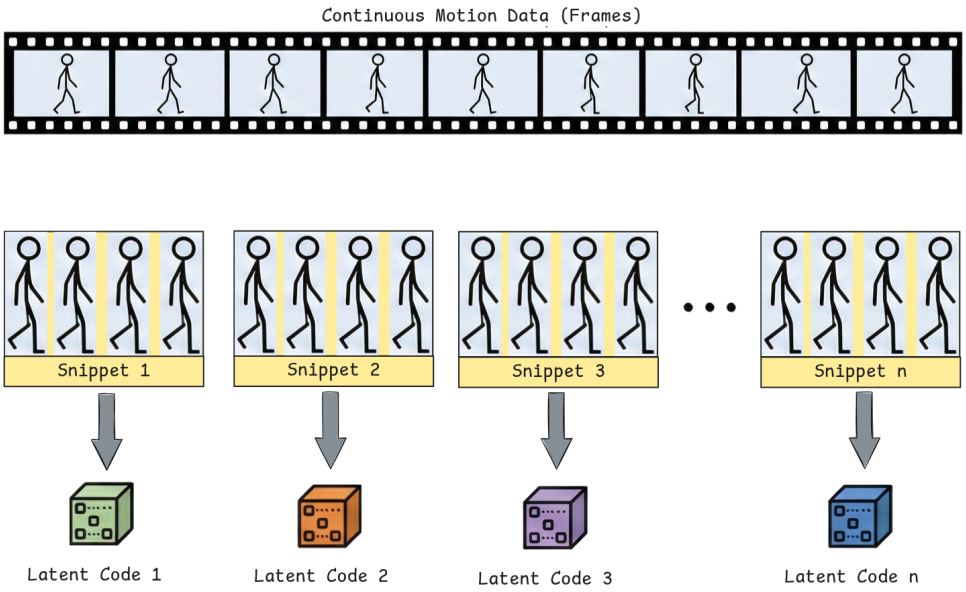

Figure 3.2: Motion Snippet Decomposition. Continuous motion data is divided into snippets, which are then encoded into latent codes.

gait characteristics (e.g., "a person walking forward at a normal pace," "the subject walks with short, quick steps").

The model is trained on HumanML3D, a large-scale dataset containing more than 14,600 motion sequences paired with approximately 45,000 textual descriptions. Leveraging this resource, the model learns statistical regularities of human gait and produces diverse, semantically faithful walking motions. These generated sequences form the foundation for the quantitative analyses of realism and biomechanical plausibility presented in later sections.

3.2 Our Gait Features

In this section, our approach to feature extraction was informed by recent comprehensive reviews on human motion generation[ZMR+23], which highlight the necessity of precise physical and kinematic evaluations. To assess the naturalness and physical plausibility of the walking sequences , we adopted metrics inspired by this domain, specifically focusing on the velocity and acceleration of the sacrum and ankles. Because standard generative evaluations often rely on foot-ground contact mechanics and joint derivatives, we further expanded our feature set to include classical biomechanical markers such asstep width, step length, and cadence to ensure a robust and holistic analysis of the gait cycle. These

features span spatial, temporal, and dynamic domains, enabling a multi-scale characterization of human locomotion. Spatial descriptors capture geometric properties of foot placement, while temporal measures quantify the rhythm and pacing of gait. Dynamic indicators such as sacrum velocity, center-of-mass acceleration, and foot-level kinematics provide insight into smoothness, stability, and compliance with biomechanical constraints. By jointly analyzing these complementary metrics, we obtain a robust and interpretable representation of gait quality, one that allows us to diagnose artifacts, assess statistical realism, and benchmark the generative model's ability to reproduce natural human walking patterns.

3.2.1 Step Length and Step Width

Step length represents the forward (anteroposterior) distance between the feet at the moment of foot contact, whereas step width captures the mediolateral separation of the feet. Together, these geometric relationships define both the progression of the gait cycle and the base of support during walking.

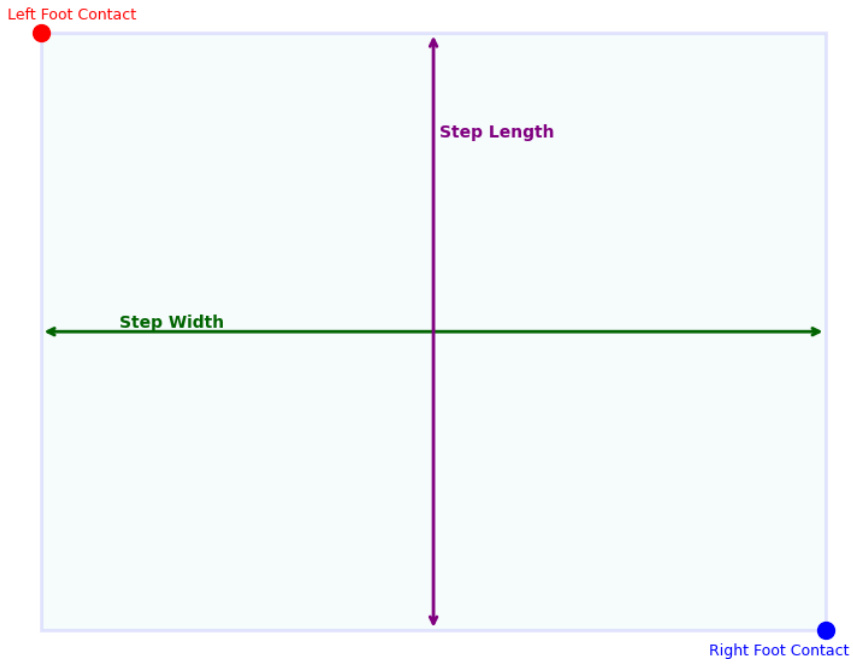

Figure 3.3: Schematic illustration of step length and step width. Step length (vertical arrow) represents the anteroposterior distance between the left and right foot contact points, while step width (horizontal arrow) captures the mediolateral separation of the feet during walking.

Figure 3.3 illustrates these two primary spatial gait descriptors. In our framework, step

metrics are computed directly from the 3D positional data of the left ankle ($\mathbf{P}_L$) and right ankle ($\mathbf{P}_R$) at frames where a foot-contact (heel-strike) event is detected. Let

$$T = \{t_1, t_2, \ldots, t_n\} \tag{3.4}$$

denote the ordered set of all such contact frames.

Based on the adopted foot-contact algorithm, the step length for a given event at frame $t_k \in T$ is defined as the Euclidean distance between the left and right ankle positions at that moment:

$$SL_k = \|\mathbf{P}_L(t_k) - \mathbf{P}_R(t_k)\|_2. \tag{3.5}$$

Expanding this expression over 3D coordinates yields:

$$SL_k = \sqrt{(x_L(t_k) - x_R(t_k))^2 + (y_L(t_k) - y_R(t_k))^2 + (z_L(t_k) - z_R(t_k))^2}. \tag{3.6}$$

Conceptually, the step length corresponds to the vertical arrow shown in Figure 3.3, representing the forward separation of the two feet during heel-strike.

Conversely, step width corresponds to the horizontal arrow in Figure 3.3, representing the lateral spacing that contributes to the subject's base of support. It is defined as the absolute difference between the left and right ankle positions along the mediolateral axis (the axis perpendicular to the direction of motion). Assuming the mediolateral axis aligns with the y-axis, the step width for frame t_k is computed as:

$$SW_k = |y_L(t_k) - y_R(t_k)|. \tag{3.7}$$

For both spatial measures, we extract the minimum, maximum, mean, and standard deviation across all steps in a sequence (i.e., $\{SL_1, \ldots, SL_n\}$ and $\{SW_1, \ldots, SW_n\}$). This comprehensive extraction enables the characterization of typical gait progression, base of support, and stride-to-stride variability.

Rationale and Relevance to Gait Naturalness

Human walking tends to maintain moderate and relatively consistent step geometry. Large deviations or irregular fluctuations in either step length or width can signal instability, compensatory motion, or non-biomechanical patterns. Notably, the variability of step width is a well-established indicator of gait stability: highly erratic lateral spacing often reflects poor balance or unnatural locomotion.

For generative motion models, these metrics serve as important indicators of realism. Extremely short or long steps, abnormally wide or narrow stances, or isolated outlier values (captured via min/max statistics) may reflect artifacts or implausible kinematic configurations. Meanwhile, the mean and standard deviation summarize the overall consistency of the synthetic gait. By jointly analyzing these spatial measures, we obtain a robust assessment of whether a motion sequence adheres to natural human walking patterns.

3.2.2 Step Frequency (Cadence)

Step frequency, or cadence, quantifies how frequently steps occur during locomotion and is typically expressed in steps per minute (SPM). As a temporal characteristic of gait, cadence, together with step length, largely determines walking speed: increasing either the step rate or the step length results in a higher forward velocity. Consequently, analyzing step frequency allows us to verify whether the generated walking motions exhibit realistic pacing patterns. Unnatural gait sequences often manifest implausible cadence–step-length combinations (e.g., extremely rapid "shuffling" steps with very small step lengths, or very slow stepping paired with unrealistically large strides). Irregular timing between successive steps may also reflect poor temporal coherence in the generated motion.

To formally compute this metric, let C_L and C_R denote the sets of detected foot-contact events for the left and right foot, respectively, obtained using the foot-contact detection algorithm. The total number of steps in a sequence is:

$$N_{\text{steps}} = |C_L| + |C_R|. \tag{3.8}$$

Let the duration of the motion sequence in minutes be T_{min}, derived from the total number of frames N_{frames} and the sampling frame rate f_{fps} (frames per second):

$$T_{\text{min}} = \frac{N_{\text{frames}}}{f_{\text{fps}} \cdot 60}. \tag{3.9}$$

The Step Frequency (Cadence), measured in steps per minute, is therefore computed as:

$$SF = \frac{N_{\text{steps}}}{T_{\text{min}}} = \frac{(|C_L| + |C_R|) \cdot f_{\text{fps}} \cdot 60}{N_{\text{frames}}}. \tag{3.10}$$

By tracking cadence, we can assess whether the generated motion sequences maintain realistic temporal dynamics and identify inconsistencies in gait rhythm or pace.

3.2.3 Gait Speed

Gait speed is a fundamental descriptor of locomotion and offers insight into both global movement characteristics and fine-grained temporal dynamics. In this thesis, we quantify walking speed using two complementary measures: instantaneous sacrum velocity, which provides a continuous estimate of the center-of-mass motion, and stride-level velocity, which captures per-cycle forward progression.

To formally define these metrics, let $\mathbf{P}_j(t) \in \mathbb{R}^3$ denote the 3D position of joint j at frame t, where $j \in \{\text{Sacrum}, \text{Ankle}_L, \text{Ankle}_R\}$, and let $\Delta t = 1/\text{FPS}$ be the time interval between frames.

Instantaneous Sacrum Velocity

The instantaneous velocity vector of joint j is computed using a first-order finite difference:

$$\mathbf{v}_j(t) = \frac{\mathbf{P}_j(t+1) - \mathbf{P}_j(t)}{\Delta t}. \tag{3.11}$$

The scalar speed is obtained as the Euclidean norm:

$$S_j(t) = \|\mathbf{v}_j(t)\|_2. \tag{3.12}$$

For the sacrum joint, which best approximates the pelvis or center-of-mass motion, we compute the mean, standard deviation, and range (minimum and maximum) of $S_j(t)$ over the full sequence duration T. This continuous velocity profile reflects fine-scale variations in walking smoothness and stability. In natural steady walking, these fluctuations remain low; large variability, abrupt changes, or oscillatory patterns may indicate unrealistic or unstable motion generated by the model.

Stride Velocity

To capture forward progression at the timescale of full gait cycles, we use stride velocity. Let $E = \{t_1, t_2, \ldots, t_N\}$ denote the ordered set of heel-strike events. For each pair of consecutive contacts t_k and t_{k+1}, the stride velocity is computed from the forward displacement of the sacrum, typically its x-coordinate, P^x_{sacrum}:

$$V_{\text{stride},k} = \frac{P^x_{\text{sacrum}}(t_{k+1}) - P^x_{\text{sacrum}}(t_k)}{(t_{k+1} - t_k) \cdot \Delta t}. \tag{3.13}$$

We then aggregate these values to obtain the mean stride velocity and its variability:

$$\mu_{\text{stride}} = \frac{1}{N-1} \sum_{k=1}^{N-1} V_{\text{stride},k}. \tag{3.14}$$

Stride velocity provides one value per gait cycle and allows us to analyze consistency across consecutive steps. A human walking at a constant pace exhibits nearly uniform stride velocities; large deviations may reflect unnatural pacing, irregular gait generation, or discontinuities in the synthesized motion.

Together, these two measures, continuous sacrum velocity and stride-level velocity, capture both moment-to-moment and cycle-level characteristics of gait speed. The mean values indicate whether the generated motion corresponds to a realistic walking speed, while the

variability metrics (standard deviation and range) reflect smoothness and stability. Consistent stride velocities indicate coherent forward progression, whereas large stride-to-stride differences may signal unnatural or jerky motion. Thus, these speed-related features allow us to assess whether the model produces a steady, human-like pace or displays irregularities that deviate from natural gait behavior.

3.2.4 Center-of-Mass Acceleration (Sacrum Acceleration)

In addition to evaluating velocity, we compute the acceleration of the sacrum (pelvis) marker, which approximates the acceleration of the body's center of mass (CoM). This quantity reflects how the forward motion of the body changes over time, capturing both speed-ups and slow-downs during walking. Formally, let $P_{\text{com}}(t)$ denote the 3D position of the sacrum marker at frame t, and let $\Delta t = 1/\text{FPS}$ be the fixed time interval between frames. The instantaneous CoM velocity is estimated using a first-order finite difference:

$$v_{\text{com}}(t) = \frac{P_{\text{com}}(t+1) - P_{\text{com}}(t)}{\Delta t}. \tag{3.15}$$

The instantaneous acceleration vector is then computed as the finite difference of the velocity:

$$a_{\text{com}}(t) = \frac{v_{\text{com}}(t+1) - v_{\text{com}}(t)}{\Delta t}. \tag{3.16}$$

We further compute the scalar acceleration magnitude using the Euclidean norm:

$$A_{\text{com}}(t) = \|a_{\text{com}}(t)\|_2 = \sqrt{a_x(t)^2 + a_y(t)^2 + a_z(t)^2}, \tag{3.17}$$

and extract its minimum, maximum, mean (μ_a), and standard deviation (σ_a) over the full sequence.

Acceleration provides a direct measure of gait dynamics, specifically, how smoothly or abruptly the walking speed varies over time. During steady, natural walking, the average forward acceleration is close to zero, as the body moves at approximately constant speed. Small rhythmic fluctuations occur within each gait cycle, such as a brief increase in speed during push-off and a slight deceleration during double support, but these patterns are subtle and well-regulated. A stable gait minimizes unnecessary acceleration and deceleration, since abrupt changes in momentum increase energetic cost and produce perceptible jerkiness.

Therefore, abnormally large peaks or high variability in the sacrum acceleration profile indicate irregularities in motion. Excessive spikes may reflect unrealistic dynamics (e.g., sudden stops or starts), while unusually dampened signals may suggest under-energized

or overly smoothed motion. By analyzing the acceleration magnitude distribution, we quantify the smoothness and physical realism of the generated gait. When considered together with velocity-based metrics, this measure provides a clear indicator of whether the model exhibits a fluid, natural cadence or deviates from expected human locomotion patterns.

3.2.5 Foot (Ankle) Trajectory Velocity

To characterize the dynamics of foot motion, we analyze the kinematic trajectories of the left and right ankle markers across each sequence. For each foot, we compute the instantaneous velocity and acceleration over time, and summarize these signals using their minimum, maximum, mean, and standard deviation. These features describe limb-level gait kinematics and are particularly relevant for identifying artifacts and assessing physical realism.

Let $P_{\text{foot}}(t) \in \mathbb{R}^3$ denote the 3D position of a foot marker, either left (L) or right (R), at frame t, and let $\Delta t = 1/\text{FPS}$ be the temporal resolution.

The instantaneous velocity vector $v_{\text{foot}}(t)$ is computed using a first-order finite difference:

$$v_{\text{foot}}(t) = \frac{P_{\text{foot}}(t+1) - P_{\text{foot}}(t)}{\Delta t}. \tag{3.18}$$

The scalar foot speed is defined as the Euclidean norm of this vector:

$$S_{\text{foot}}(t) = \|v_{\text{foot}}(t)\|_2. \tag{3.19}$$

3.2.6 Foot (Ankle) Trajectory Acceleration

The instantaneous acceleration vector $a_{\text{foot}}(t)$ is obtained as the discrete derivative of the velocity:

$$a_{\text{foot}}(t) = \frac{v_{\text{foot}}(t+1) - v_{\text{foot}}(t)}{\Delta t}, \tag{3.20}$$

with corresponding magnitude:

$$A_{\text{foot}}(t) = \|a_{\text{foot}}(t)\|_2. \tag{3.21}$$

This formulation allows us to quantify how the feet move during the gait cycle. In normal human walking, each foot follows a periodic pattern: it remains nearly stationary during the stance phase, with velocity approaching zero when in contact with the ground, and then accelerates forward during swing, reaching peak velocity mid-air before decelerating

again. By analyzing the minimum foot speed, we can detect whether the model correctly produces stationary foot contacts, violations of this (i.e., the foot moving when it should be planted) indicate foot sliding, a well-known artifact in poorly generated motion.

Acceleration statistics provide complementary information: realistic gait exhibits smooth, continuous transitions between stance and swing, whereas unnaturally high acceleration spikes suggest jitter, snapping, or teleportation-like artifacts. Moreover, evaluating both left (L) and right (R) foot trajectories enables us to detect asymmetries or irregularities between the legs, since healthy gait is typically symmetric in timing and amplitude.

Overall, ankle velocity and acceleration metrics serve as key indicators of physical plausibility at the contact level, ensuring both grounded foot behavior and realistic swing-phase dynamics.

3.2.7 Summary of Feature Use

Taken together, the extracted gait features provide a comprehensive, multi-dimensional evaluation of the generated motions. Spatial metrics (step length and width) assess foot placement geometry and stability; temporal metrics (cadence) characterize the pacing and rhythm of the walk; dynamic measures (sacrum velocity, stride velocity, and center-of-mass acceleration) capture smoothness, forward progression, and global motion coherence; and foot-level kinematics reveal local artifacts such as foot sliding, jitter, or asymmetries between legs. Because each feature set emphasizes a different aspect of locomotor behavior, their combined analysis allows us to detect deviations that may not be apparent from any individual metric alone. This integrated feature framework therefore, serves as the foundation for evaluating both the physical plausibility and semantic consistency of the synthesized gait sequences throughout the experimental analyses that follow.

3.3 Metrics

This section describes the quantitative evaluation protocol adopted to assess whether the extracted gait features are sufficiently discriminative to distinguish real from synthetically generated walking motions. The evaluation is formulated as a supervised binary classification problem, where the objective is to predict the origin of each motion sequence based solely on the biomechanical feature set defined in Section 3.2.

3.3.1 Datasets

Real Data

Real motion sequences were obtained from the Full-Body Gait Dataset [VCST+23], a large-scale motion capture collection comprising 138 able-bodied adults across the lifespan and 50 stroke survivors. In this study, only the 138 able-bodied subjects were considered to avoid pathological gait patterns that could confound the realism analysis.

Each participant performed multiple walking trials recorded in C3D format. Trials that were incomplete or contained missing joint trajectories were excluded during preprocessing. After filtering, a total of:

$$N_{\text{real}} = 223 \text{ trials}$$

were retained for analysis.

To ensure compatibility with the text-to-motion representation, we mapped the original motion capture skeleton to a joint configuration as close as possible to the HumanML3D-style[GZZ+22b] format used by the generative model. All joint positions were normalized independently along each spatial axis using min–max normalization, defined as:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

applied per sequence and per coordinate axis. This normalization ensures scale consistency between real and generated data while preserving intra-sequence kinematic structure.

Generated Data

Synthetic walking sequences were produced using the text-to-motion model[GZZ+22b] described in section 3.1. Since the real dataset consists of subjects walking approximately along a straight trajectory, prompting strategies were designed to encourage forward straight-line walking in the generated samples.

To ensure statistical comparability, we generated:

$$N_{\text{gen}} = 223 \text{ samples},$$

matching the number of real trials exactly. The same preprocessing pipeline (joint selection, normalization, and feature extraction) was applied to generated sequences.

3.3.2 Feature-Based Classification Protocol

Each motion sequence was represented by the full set of extracted gait descriptors, including:

- Spatial metrics: step length and step width,
- Temporal metric: cadence,
- Velocity metrics: sacrum mean/variance and stride velocity,
- Acceleration metrics: sacrum acceleration,
- Foot kinematics: ankle velocity and acceleration.

For each feature, we summarize its distribution over the sequence by computing the minimum, maximum, mean, and standard deviation.

This results in a fixed-dimensional feature vector per sequence.

The classification labels were defined as:

- **Class 0:** Generated motion
- **Class 1:** Real motion

3.3.3 Model Selection: Logistic Regression

To evaluate the separability of real and synthetic gait distributions in feature space, we employed Logistic Regression, a linear probabilistic classifier suitable for interpretable binary discrimination.

Given a feature vector $\mathbf{x} \in \mathbb{R}^d$, the model estimates:

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b),$$

where $\sigma(\cdot)$ is the sigmoid function and $(\mathbf{w}, b)$ are learned parameters.

Logistic Regression was selected for three reasons:

1. **Interpretability:** Linear decision boundaries reveal whether biomechanical features alone are sufficient to distinguish data sources.
2. **Low model complexity:** Reduces risk of overfitting on moderate-sized datasets.
3. **Statistical transparency:** Provides clear probabilistic outputs.

3.3.4 Data Splitting and Training Procedure

The combined dataset (446 samples total) was stratified to preserve class balance and divided as follows:

- **Training set:** 70%
- **Validation set:** 15%
- **Test set:** 15%

Stratification ensured equal proportions of real and generated samples in each split.

The training set was used to estimate model parameters. Hyperparameter tuning was performed using the validation set. Final performance was reported on the held-out test set.

3.3.5 Evaluation Metrics

Model performance was evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

These metrics provide complementary perspectives: accuracy measures overall correctness; precision quantifies false positive rates; recall quantifies false negative rates; the F1-score balances precision and recall; and the confusion matrix reveals class-specific misclassifications.

3.3.6 Quantitative Results

The trained Logistic Regression model achieved:

- **Validation Accuracy:** 1.00

- **Test Accuracy:** 98.53%

Class-wise performance on the test set:

Class	**Precision**	**Recall**	**F1-score**
0 (Generated)	1.000	0.971	0.985
1 (Real)	0.971	1.000	0.986

Table 3.1: Class-wise test performance for Logistic Regression.

Confusion Matrix (Test Set)

$$\begin{bmatrix} 33 & 1 \\ 0 & 34 \end{bmatrix}$$

Interpretation:

- 33 generated sequences correctly classified.
- 34 real sequences correctly classified.
- 1 generated sequence misclassified as real.
- No real sequence misclassified as generated.

Thus, only one misclassification occurred.

3.3.7 Interpretation and Implications

The near-perfect classification performance indicates that the extracted gait features provide strong discriminative power between real and synthetically generated walking motions.

Key observations:

1. High linear separability suggests that synthetic motions exhibit systematic statistical deviations from real biomechanical patterns.
2. The single false positive (generated labeled as real) indicates that some synthetic samples approach realistic feature distributions.
3. The absence of false negatives (real labeled as generated) suggests the classifier does not incorrectly penalize authentic human gait.

These findings imply that, although visually plausible, generated motions still contain measurable biomechanical signatures that distinguish them from real human walking.

Importantly, since a simple linear classifier achieves $> 98\%$ test accuracy, the discrepancy between real and generated distributions exists at the level of fundamental spatial, temporal, and dynamic gait statistics rather than requiring complex nonlinear modeling to detect.

This quantitative evaluation therefore demonstrates that, while generative models produce semantically coherent walking motions, subtle yet statistically significant differences remain detectable through biomechanical feature analysis.

Chapter 4

Experimental Evaluation and Analysis

The goal of this chapter is to provide a comprehensive empirical evaluation of the text-to-motion generative model[GZZ+22b], focusing on its ability to produce realistic, diverse, and semantically faithful human motion sequences from natural-language descriptions. While previous chapters examined the model architecture and feature extraction pipeline, the present analysis addresses three essential dimensions of generative performance: variability, semantic controllability, and biomechanical quality. Together, these dimensions form a rigorous framework for assessing the reliability and practical viability of text-conditioned motion synthesis. First, we investigate whether the model can generate natural variability in its outputs when conditioned on identical textual prompts. Human motion is inherently stochastic; even repeated performances of the same action exhibit temporal and spatial differences. A generative model must therefore avoid deterministic repetition and instead produce diverse yet semantically consistent motions. This analysis evaluates both temporal variability (motion duration) and spatial variability (trajectory patterns) to determine whether the model accurately captures the inherent stochasticity of locomotion. Second, we examine semantic consistency, evaluating how the model responds to different textual descriptions that encode variations in speed, direction, and complexity of movement. By systematically comparing motions generated from prompts describing slow walking, normal walking, running, circular trajectories, and zig-zag paths, we assess the extent to which linguistic nuances translate into distinct kinematic and spatial behaviors. This investigation highlights the model's sensitivity to verbal cues and its ability to differentiate between qualitatively and quantitatively distinct forms of movement. Third, the chapter assesses temporal regularity and biomechanical plausibility by analyzing joint-level velocity patterns, specifically, the oscillatory behavior of ankle velocities that characterizes natural gait cycles. Through both single-sequence inspection and aggregated statistical analysis over

100 generated samples, we evaluate whether the model produces stable step cycles and consistent left–right alternations, or whether irregularities emerge across generations. Finally, we address the crucial question of prompting strategy effectiveness by comparing Numerical Prompting, which supplies explicit quantitative descriptors such as speed and frame count, to Descriptive Prompting, which expresses motion characteristics using qualitative, human-readable language. Using a feature-based Fréchet Distance metric [HRU+17], we assess how closely each prompting strategy reproduces the statistical distribution of real walking trials. This comparison provides insight into how the model interprets quantitative information and whether numerical conditioning can serve as a reliable means of motion control. Overall, the analyses presented in this chapter offer a detailed and multifaceted evaluation of the generative model's behavior. By examining variability, semantic alignment, biomechanical structure, and controllability, we establish a clear understanding of the model's strengths, limitations, and the conditions under which it produces motions that resemble real human locomotion.

4.1 Assessment of Generative Variability and Semantic Consistency

4.1.1 Evaluating the Variability of the Generated Sequences

The objective of this experiment is to evaluate the variability of motion sequences generated by the text-to-motion model when provided with identical textual inputs. Specifically, the analysis investigates whether the model introduces natural diversity in its outputs, both temporally and spatially, or whether it produces repetitive and deterministic results. This variability is an important indicator of the model's ability to capture the stochastic nature of human motion, even when the textual description remains constant.

Method

Six textual prompts were selected to represent simple but diverse walking behaviors. These prompts were randomly chosen from the HumanML3D [GZZ+22b] dataset's textual annotations, which describe short natural-language labels of human actions. Selecting prompts from an existing motion–language dataset ensures that the descriptions are semantically consistent with real motion data and reflect natural variations in walking patterns.

Each prompt was used to generate ten distinct motion sequences under identical model conditions, thereby isolating the variability arising solely from the generation process rather than from external parameters.

The prompts employed in this experiment were as follows:

- *The person walks forward in a slight diagonal.*
- *It is a person walking backwards.*
- *The person is walking from left to right.*
- *The person walks forward, turns around and walks back.*
- *A figure walks and spins on their heel.*
- *Walking forward and then stopping.*

For each generated sequence, two types of analyses were conducted:

Temporal analysis: the total number of frames was computed to quantify the duration of each motion.

Spatial analysis: the trajectories were projected onto the X–Z plane to visualize movement patterns, directions, and curvature.

Together, these analyses provide insight into both temporal and spatial variability within and across prompts.

Results

Table 4.1 reports the mean and standard deviation of motion lengths (in frames) for each prompt, calculated across the ten generated sequences.

Prompt	**Mean**	**Std**
The person walks forward in a slight diagonal.	94.8	12.4
It is a person walking backwards.	105.2	35.82
The person is walking from left to right.	180.4	18.37
The person walks forward, turns around and walks back.	128.4	20.04
A figure walks and spins on their heel.	94.0	27.44
Walking forward and then stopping.	108.4	46.77
Overall	**118.53**	**41.81**

Table 4.1: Mean and standard deviation of motion lengths (in frames) for each prompt.

The results demonstrate noticeable variability across the generated sequences, both within individual prompts and between different prompts. The mean motion length varies from

approximately 94 to 180 frames, reflecting how the model adapts the motion duration to the semantic content of each description. Prompts describing more dynamic or extended actions (e.g., walking from left to right) tend to yield longer sequences, whereas simpler or more localized motions (e.g., walking diagonally or spinning on the heel) produce shorter ones.

Figure 4.1 visualizes the 2D trajectories of all generated motions for the six prompts. Each subplot corresponds to one prompt, showing ten trajectories projected onto the X–Z plane. While the general motion direction remains consistent with the textual description, individual trajectories exhibit small deviations, variations in curvature, and differences in movement extent. For instance, in the "walk forward in a slight diagonal" case, the paths are relatively parallel and uniform, whereas in "walking from left to right," the trajectories are more complex and interwoven, indicating a higher degree of spatial variability.

Overall, these findings suggest that the text-to-motion model is capable of generating non-deterministic outputs that maintain semantic consistency while exhibiting natural variability in both temporal and spatial dimensions, an essential property for realistic motion synthesis.

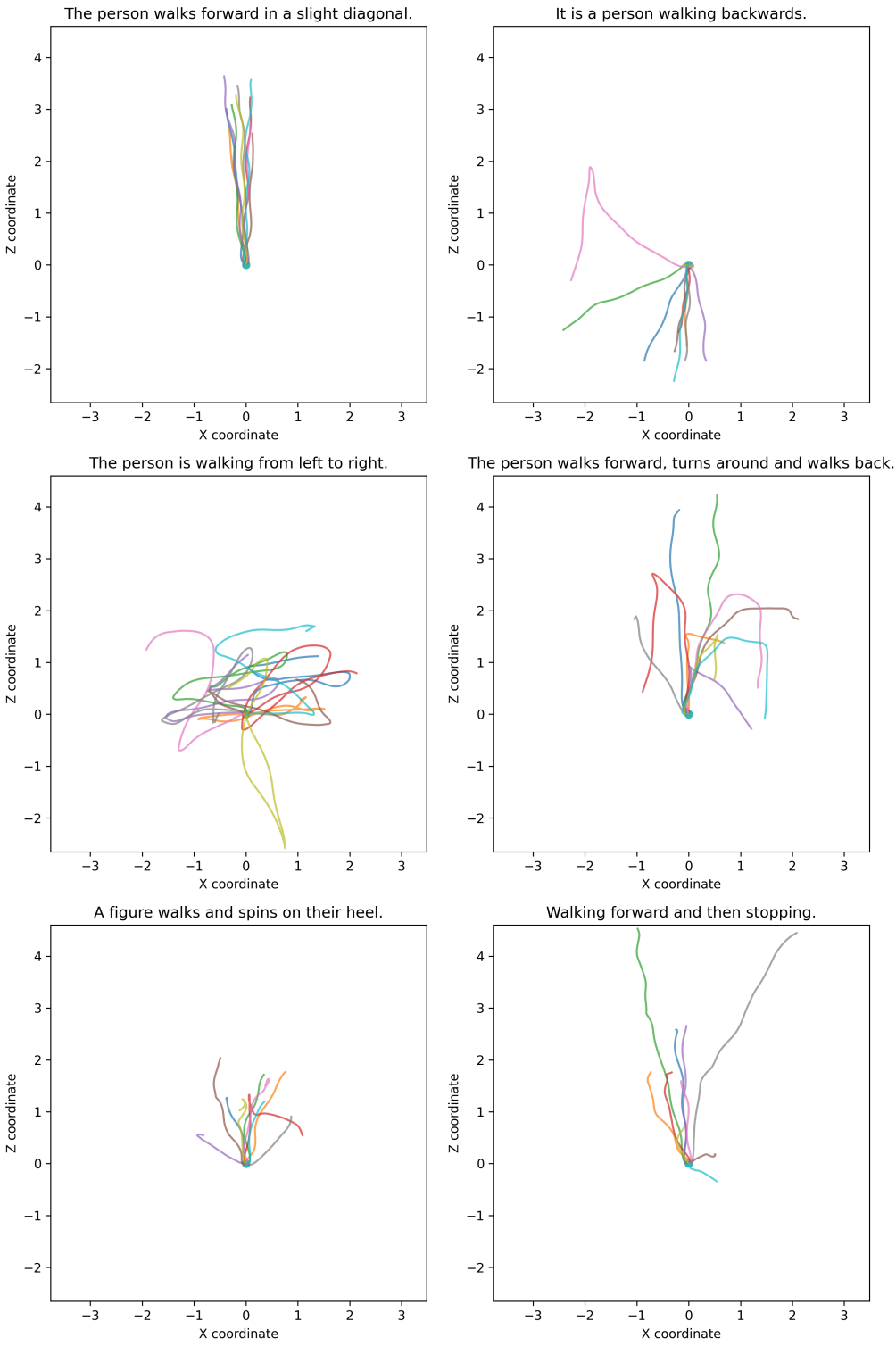

Figure 4.1: Walking trajectories of ten generated sequences for each of the six prompts, projected onto the X–Z plane.

4.1.2 Evaluating Variability with Different Prompts

The objective of this experiment is to systematically examine how the text-to-motion generative model responds to varying verbal descriptions of locomotion behaviors, specifically focusing on walking and running activities with different qualitative and quantitative characteristics. This analysis investigates the model's capability to differentiate between semantic nuances in textual prompts and translate these linguistic variations into correspondingly distinct motion patterns. The experiment evaluates both the quantitative aspects (such as movement speed and velocity profiles) and qualitative characteristics (including spatial trajectory patterns and directional behaviors) that should emerge from different textual descriptions.

Six carefully selected prompts were chosen to represent a spectrum of locomotion behaviors, encompassing variations in speed, direction, and movement patterns:

- *Walking forward at a slow pace.*
- *Walking forward at a normal pace.*
- *Running forward.*
- *Walking toward something.*
- *Walking in a circle.*
- *Walking zig-zag.*

These prompts were specifically designed to test the model's ability to capture semantic differences across multiple dimensions: speed gradations (slow, normal, running), directional complexity (straight forward, circular, zig-zag patterns), and goal-oriented behavior (walking toward something). The primary research question addresses whether the model can generate motion sequences that accurately reflect both the quantitative parameters (such as velocity magnitude) and qualitative spatial characteristics (trajectory patterns and directional behaviors) implied by these diverse textual variations.

Method

Each of the six selected prompts was used to generate ten distinct motion sequences under identical model conditions, ensuring that observed variations could be attributed to the model's interpretation of semantic differences rather than external parameters or random initialization effects. This systematic approach provides a robust statistical foundation for comparing the model's responses across different textual descriptions.

Two complementary types of quantitative analysis were conducted to capture both the kinematic and spatial characteristics of the generated motions:

1. Velocity Analysis: The average magnitude of sacrum velocity was computed for each generated sequence to quantify differences in locomotion speed. The sacrum marker, representing the pelvis and center of mass, provides a reliable indicator of overall body velocity during locomotion. This analysis enables direct comparison of movement speeds across different prompts, particularly for evaluating whether the model appropriately differentiates between speed-related descriptors (slow pace, normal pace, running).

2. Spatial Trajectory Analysis: The XZ-plane trajectories were extracted and visualized to examine spatial movement patterns and directional behaviors. By projecting the three-dimensional motion onto the horizontal plane, we can assess whether the generated motions exhibit trajectory characteristics consistent with the textual descriptions, such as linear paths for forward walking, circular patterns for circular walking, or alternating lateral movements for zig-zag walking.

For each prompt, statistical measures including mean and standard deviation were calculated across the ten generated samples to quantify both central tendencies and variability in the model's responses. This statistical approach enables assessment of consistency within each prompt category while highlighting differences between categories.

Table 4.2 reports the comprehensive velocity statistics for each prompt, while Figure 4.2 provides visual representations of the spatial trajectory patterns generated for all six locomotion behaviors.

Prompt	**Mean Velocity (units/sec)**	**Std. Deviation**
Walking forward at a slow pace.	0.3324	0.1321
Walking forward at a normal pace.	0.4097	0.2069
Running forward.	0.9306	0.2234
Walking toward something.	0.2622	0.0771
Walking in a circle.	0.8880	0.1255
Walking zig-zag.	0.7403	0.0703

Table 4.2: Average sacrum velocity (mean and standard deviation) for different walking and running prompts

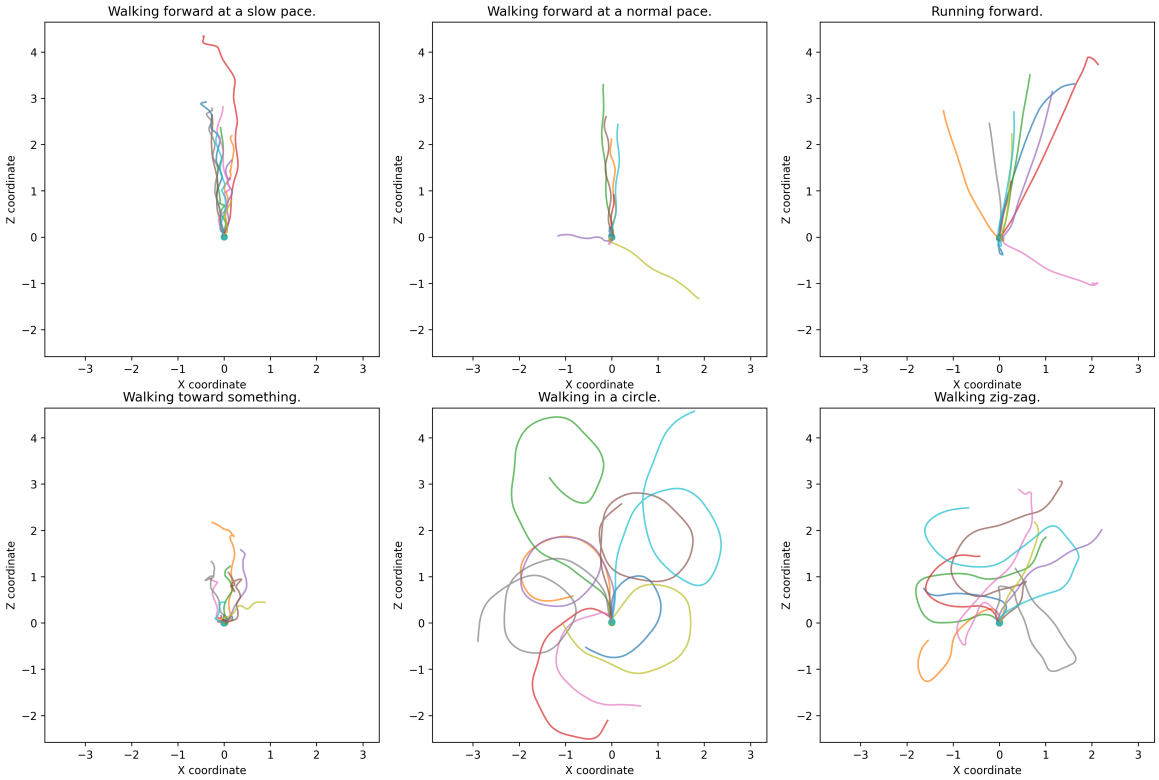

Figure 4.2: XZ-plane trajectories generated for each of the six prompts. Each subplot represents ten generated samples. The plots demonstrate distinct spatial and directional patterns that align with the semantics of the textual descriptions.

Results

The velocity analysis reveals that the text-to-motion model demonstrates strong semantic understanding of speed and movement descriptors. Most importantly, *Running forward* generated the highest mean velocity (0.9306 units/second), clearly distinguishing it from all walking behaviors. Among walking prompts, a logical speed hierarchy emerged: *Walking forward at a normal pace* (0.4097 units/second) exceeded *Walking forward at a slow pace* (0.3324 units/second), confirming the model's ability to interpret pace modifiers. Notably, complex movement patterns like *Walking zig-zag* (0.7403 units/second) and *Walking in a circle* (0.8880 units/second) produced higher velocities than simple forward walking, suggesting these intricate trajectories require increased movement speed to execute the prescribed patterns within the sequence duration. The lowest velocity was observed for *Walking toward something* (0.2622 units/second), indicating the model interprets goal-directed movement as cautious and deliberate.

The spatial trajectory analysis demonstrates remarkable alignment between textual semantics and generated motion patterns. *Walking in a circle* consistently produced characteristic looping trajectories with varying radii, while *Walking zig-zag* displayed clear alternating lateral movements with directional changes as expected. *Walking toward something* generated direct, goal-oriented linear paths with minimal deviation, reflecting purposeful target-directed movement. All forward walking prompts maintained predominantly linear trajectories, differing primarily in length and minor lateral variations rather than fundamental directional changes. The standard deviation values (ranging from 0.0703 to 0.2234 units/second) indicate that spatially constrained movements like zig-zag and goal-directed walking show lower variability, while open-ended prompts like running demonstrate higher diversity, suggesting the model appropriately balances consistency with natural motion variation across different prompt types.

4.2 Assessment of the quality of the sequences

4.2.1 Analysis of temporal consistency in Generated Sequences

This experiment aims to examine the temporal consistency and biomechanical plausibility of synthetically generated walking motions by analyzing the velocity magnitude of the left and right ankle joints. Since natural human gait exhibits alternating rhythmic foot movements, periodic peaks in ankle velocity serve as an indicator of realistic step cycles. By visualizing velocity patterns both at the single-sequence level and across multiple generations, this experiment evaluates whether the model produces stable, repeatable walking dynamics or displays irregular motion patterns.

Methodology

A total of 100 motion sequences were generated using the prompt *"He slowly walks forward towards something"* from a text-to-motion model. To ensure uniformity across sequences, all samples were truncated to the minimum observed sequence length of 68 frames, enabling consistent frame-wise statistical comparison.

For each sequence, the velocity magnitude of both ankle joints was computed based on their positional displacement between consecutive frames. The analysis proceeded in two stages:

Single-sequence qualitative analysis: One representative generated sequence was selected to visualize both the foot velocity profiles and corresponding motion frames. The plot overlays the left and right foot velocity curves alongside snapshots of the motion near key

velocity peaks, allowing visual verification of whether the peaks correspond to expected gait events such as toe-off and swing-phase acceleration.

Cross-sequence statistical analysis: For each frame, the mean and variance of the velocity magnitude were computed over all 100 generated sequences for each foot. These metrics were visualized using shaded line plots, where the solid line indicates the mean and the shaded region represents standard deviation.

Results

Single-Sequence Analysis

Figure 4.3 presents the velocity magnitude for one generated motion sample together with rendered body-pose frames. The velocity curves exhibit distinct alternating peaks: the right foot velocity rises sharply when the right leg enters the swing phase, while the left foot shows mirrored behavior. These peaks align with the visualized frames below the plot, where the corresponding leg is lifted and advanced forward.

The periodic alignment of peaks between the two feet indicates that the model captures the basic cyclic structure of gait, acceleration during swing and near-zero velocity during stance. However, the amplitude of the peaks varies, suggesting some inconsistencies in the height and speed of steps within a single generated sequence.

Cross-Sequence Analysis

Figures 4.4 and 4.5 show the aggregated mean and variance of ankle velocity magnitude across the 100 generated sequences. Both the left and right foot display an initial rise in velocity within the first frames, corresponding to the onset of walking. This is followed by oscillatory patterns characteristic of alternating steps.

The shaded variance regions reveal a noticeable spread across sequences, especially in the mid-sequence frames, where variability increases substantially. This indicates that although the model consistently produces walking-like behavior, the exact timing and amplitude of step cycles vary between generations.

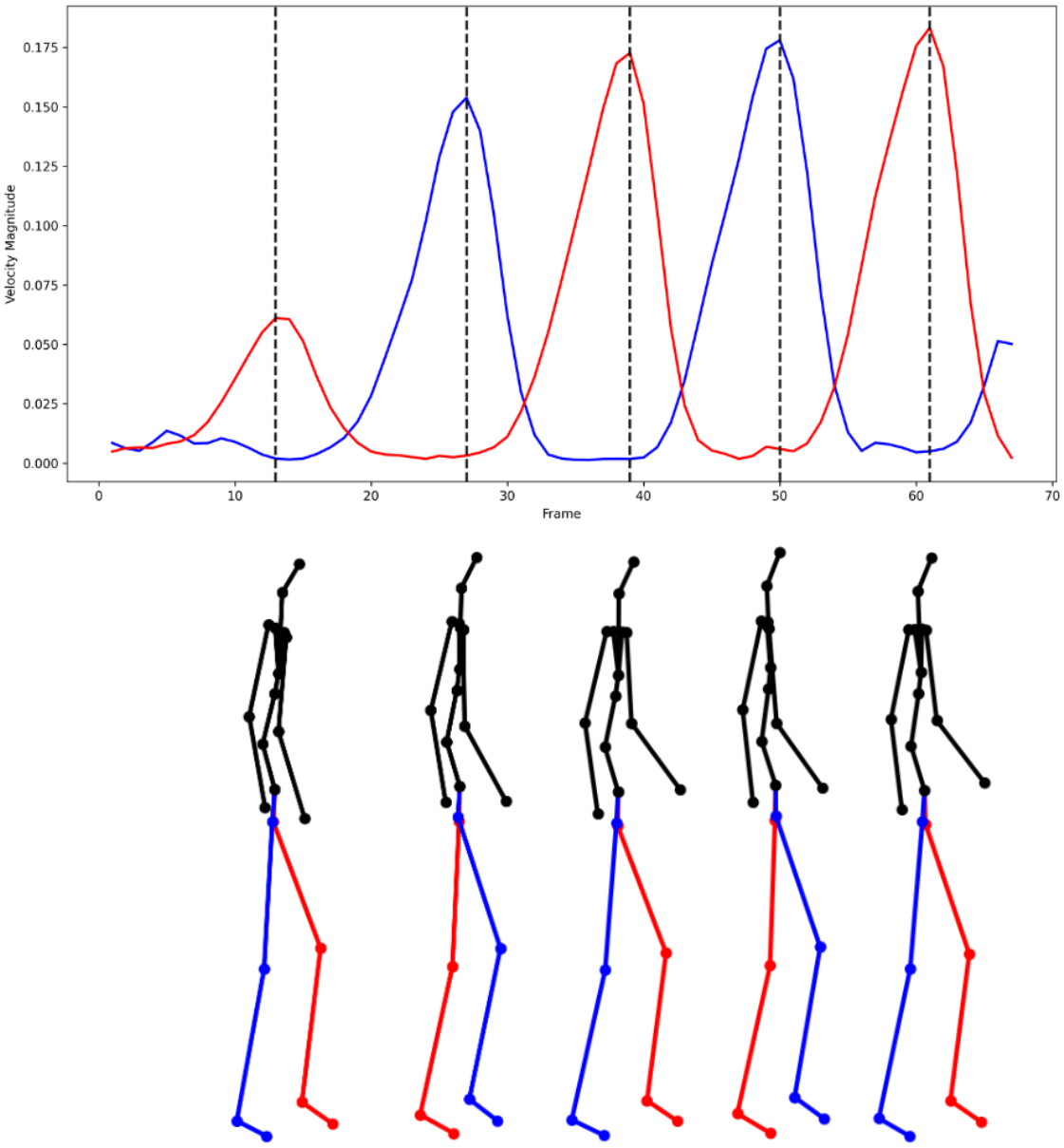

Figure 4.3: Velocity magnitude of left (red) and right (blue) foot for one generated sequence, with snapshots of the motion at selected peak frames. The peaks correspond to the swing phases of each leg.

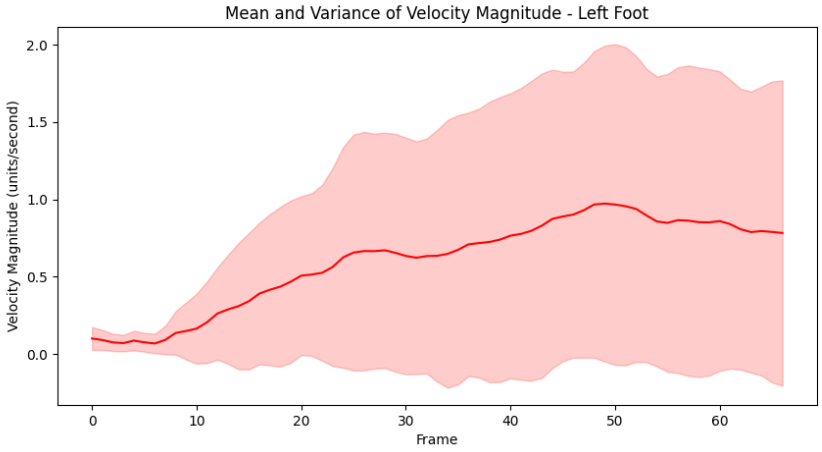

Figure 4.4: Left foot velocity magnitude across 100 generated walking sequences. The red line denotes mean velocity, and the shaded region represents standard deviation.

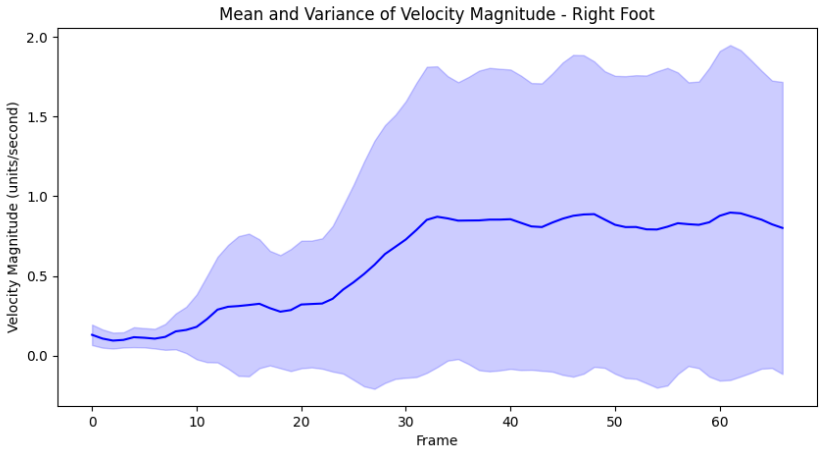

Figure 4.5: Right foot velocity magnitude across 100 generated walking sequences. The blue line denotes mean velocity, and the shaded region represents standard deviation.

The overall mean and variance of ankle velocity across all 100 sequences are presented in Table 4.3. The left and right foot show highly similar mean velocities (0.601 vs. 0.592 units/second), suggesting balanced bilateral motion. However, the variances (0.580 and 0.573) remain relatively high, indicating substantial diversity in the generated gait patterns.

Joint	Mean Velocity (units/second)	Std. Deviation
Right Ankle	0.592	0.573
Left Ankle	0.601	0.580

Table 4.3: Summary Statistics of Ankle Velocity Across 100 Generated Sequences

Overall, the results indicate that the model successfully captures the fundamental rhythmic structure of walking, but with notable variability in the magnitude and timing of steps across different generated sequences. While this variability may reflect natural diversity in human walking, the large variance values also suggest that the model's motion generation lacks strict temporal regularity.

4.2.2 Comparative Efficacy of Numerical versus Descriptive Prompting

The objective of this experiment is to examine the semantic controllability of the text-to-motion generative model by attempting to reproduce real human walking trials. In particular, we aim to identify the prompting strategy that yields synthetic motions whose statistical properties most closely resemble the underlying ground-truth distribution. To this end, we compare two contrasting prompting approaches: Numerical Prompting, which provides explicit quantitative attributes, and Descriptive Prompting, which provides qualitative linguistic descriptions.

Methodology

Data Preparation: The dataset consists of real walking trials stored in C3D format. After filtering out incomplete or corrupted recordings, a final set of 223 validated trials remained. For each trial, we extracted subject-specific metadata (age, mass, height) and motion-related attributes (number of frames, average velocity, and walking direction).

Prompt Engineering Strategies: The extracted attributes were converted into two textual conditioning strategies:

1. **Numerical Prompts:** Raw quantitative values were directly injected into the prompt to test whether the model can interpret and adhere to precise numerical constraints.
    - Example: "An 86-year-old person with a weight of 63 kg and height of 158 cm walking at 113 cm/s in 0 degrees direction for 337 frames."

2. **Descriptive Prompts:** Quantitative values were mapped into qualitative categories that resemble natural-language descriptions typically present in motion-language datasets (e.g., HumanML3D).

    - Example: "An elderly person walking in a straight line for a normal duration."
    - Mapping Logic: Age (< 30: young, 30–60: middle-aged, > 60: elderly); velocity (slow/normal/fast); direction (left/right/straight).

Generation Process: For each of the 223 real trials, we generated one synthetic sequence using the numerical prompt and one using the descriptive prompt. This produced three datasets:

1. Real Data: 223 sequences
2. Generated Numerical: 223 sequences
3. Generated Descriptive: 223 sequences

Baseline Real-Data Split: Following supervisory feedback, we additionally computed a baseline Fréchet Distance using real-world data only. To ensure that no subject appears in more than one group, the 223 real trials were split at the midpoint, producing two disjoint subsets:

- Batch A: 111 sequences
- Batch B: 112 sequences

This subject-exclusive split provides a meaningful estimate of natural variability within the dataset. The FD between Batch A and Batch B was 17.157782, representing the expected distance when comparing two real distributions.

Evaluation Metric: To compare the generated and real distributions, we computed the Fréchet Distance using the 37 handcrafted biomechanical features defined in Section 3.2 (e.g., step length statistics, cadence, sacrum velocity, foot acceleration). These features form a specialized embedding space suitable for gait analysis.

The FD was computed using the standard formulation:

$$FD = ||\mu_r - \mu_g||^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right)$$

Lower FD indicates a closer match between real and generated motion distributions.

Results

The Fréchet Distance scores for the two prompting strategies are presented in Table 4.4.

Comparison Pair	**Fréchet Distance (FD)**
Real vs. Generated *Descriptive*	389.62
Real vs. Generated *Numerical*	1403.33
Real (Batch A) vs. Real (Batch B)	17.16

Table 4.4: Fréchet Distance scores comparing the distribution of Real data against Descriptive and Numerical generations, as well as a baseline comparison between two real-data batches. Lower is better.

The Descriptive prompting approach substantially outperformed the Numerical prompting approach. The FD score of 389.62 for the Descriptive method represents a 72.24% improvement compared to the Numerical method (1403.33). The magnitude of this difference becomes even clearer when contrasted with the real-data baseline FD of 17.16, which represents the intrinsic variability of human gait within the dataset. The Descriptive prompts, while not perfectly aligned with the real distribution, remain far closer than the Numerical prompts.

These results reinforce that the model operates primarily within a semantic latent space, responding more effectively to qualitative linguistic cues (e.g., "elderly", "slow pace", "straight direction") than to explicit numeric constraints (e.g., "113 cm/s"). Numerical values appear to be interpreted as uninformative tokens, leading to synthetic motions that diverge significantly from the statistical structure of real gait. Thus, the model does not behave as a parameter-controlled generator but instead relies on high-level linguistic concepts to shape motion outputs.

Chapter 5

Conclusion

This thesis investigated a central question in contemporary generative modeling: How credible are the movements of synthetically generated humans? While recent advances in deep generative architectures, ranging from VAEs[KW13], GANs[GPAM+14], diffusion models[HJA20], and autoregressive Transformers[VSP+17], have significantly improved the visual realism and semantic controllability of text-to-motion systems, their physical plausibility has not been equally scrutinized through biomechanical analysis.

To address this gap, this thesis proposed a structured and reproducible evaluation framework grounded in classical gait analysis. Synthetic walking sequences generated by a state-of-the-art text-to-motion model[GZZ+22a] were directly compared against real motion capture data from a large-scale full-body gait dataset[VCST+23]. Rather than relying exclusively on embedding-based metrics such as R-Precision or FID[HRU+17], we extracted interpretable spatial, temporal, and dynamic descriptors of locomotion, including step length, step width, cadence, gait speed, sacrum acceleration, and foot-level velocity and acceleration profiles.

By framing realism assessment as a supervised discrimination problem, we quantitatively measured whether synthetic gait statistics are statistically indistinguishable from real human gait. A simple Logistic Regression classifier trained solely on biomechanical features achieved over 98% classification accuracy in separating real from generated walking sequences. This result demonstrates that, despite strong perceptual plausibility and semantic alignment, synthetic motions retain measurable statistical signatures that deviate from authentic locomotor dynamics.

Several insights emerge from these findings.

First, generative models excel at capturing high-level semantic intent and producing temporally coherent motion trajectories. Diffusion-based approaches[HJA20, TRG+22] in par-

ticular demonstrate strong mode coverage and smooth sequence generation. However, the distribution of low-level kinematic variables, especially dynamic descriptors such as center-of-mass acceleration and foot trajectory derivatives, reveals systematic discrepancies. These deviations are often subtle and not immediately perceptible to human observers, yet they are statistically consistent and reproducible.

Second, our experiments comparing descriptive prompts versus numerically constrained prompts indicate that generative models operate primarily within a semantic latent space. When conditioned on qualitative descriptions (e.g., "walking briskly"), the models reproduce realistic statistical distributions more effectively than when constrained by explicit numerical targets. This suggests that current architectures are optimized for semantic consistency rather than strict biomechanical control. The learned latent space encodes stylistic and contextual motion patterns, but does not guarantee adherence to precise physical constraints unless explicitly enforced during training.

Third, the results highlight the limitations of commonly used evaluation metrics[ZMR+23]. Distance-based measures such as ADE or MSE penalize valid alternative motions in one-to-many generation settings, while embedding-based metrics such as FID[HRU+17] or R-Precision primarily assess distributional similarity or semantic retrieval accuracy. While these metrics are valuable for assessing visual quality and semantic alignment, they do not directly quantify physical plausibility, which is the gap our feature-based classification protocol addresses. Our feature-based classification protocol complements existing benchmarks by providing an interpretable, physically grounded diagnostic tool capable of identifying specific motion artifacts.

Importantly, the objective of this thesis was not to discredit current generative models, but to characterize their limitations in a principled manner. The high classification separability does not imply that synthetic motion is unrealistic in practice; rather, it indicates that the statistical structure of generated gait does not yet fully replicate the fine-grained dynamics of human locomotion. In many application domains, such as animation or virtual agents, these discrepancies may be acceptable. However, in contexts requiring biomechanical fidelity, including clinical simulation, rehabilitation modeling, or ergonomic analysis, such differences become critical.

In conclusion, while contemporary text-to-motion models represent a remarkable advancement in generative AI, their outputs remain distinguishable from real human motion. By shifting the evaluation focus from perceptual and semantic criteria toward physically grounded metrics, this thesis provides a clearer understanding of the current capabilities and limitations of synthetic human motion generation.

Bibliography

[AHC+17] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. *CoRR*, abs/1710.05298, 2017. URL: `http://arxiv.org/abs/1710.05298`, `arXiv:1710.05298`.

[AM19] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. *CoRR*, abs/1907.01108, 2019. URL: `http://arxiv.org/abs/1907.01108`, `arXiv:1907.01108`.

[Car03] Carnegie Mellon University. CMU graphics lab motion capture database. Online dataset, 2003. URL: `http://mocap.cs.cmu.edu/`.

[CRZ+22] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022.

[GPAM+14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems*, 2014. URL: `https://api.semanticscholar.org/CorpusID:261560300`.

[GZW+20] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. *CoRR*, abs/2007.15240, 2020. URL: `https://arxiv.org/abs/2007.15240`, `arXiv:2007.15240`.

[GZZ+22a] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, June 2022.

[GZZ+22b] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL: https://github.com/EricGuo5513/HumanML3D.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. URL: https://arxiv.org/abs/2006.11239, arXiv:2006.11239.

[HRU+17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. URL: http://arxiv.org/abs/1706.08500, arXiv:1706.08500.

[IPOS14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[KKL+22] Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual information divergence: A unified metric for multimodal generative models. *ArXiv*, abs/2205.13445, 2022. URL: https://api.semanticscholar.org/CorpusID:249097749.

[KW13] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL: https://api.semanticscholar.org/CorpusID:216078090.

[LSP+19] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *CoRR*, abs/1905.04757, 2019. URL: http://arxiv.org/abs/1905.04757, arXiv:1905.04757.

[LZL+23] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 2023.

[MGFT+19] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. URL: https://amass.is.tue.mpg.de.

[PBV21] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer VAE. *CoRR*, abs/2104.05670, 2021. URL: https://arxiv.org/abs/2104.05670, arXiv:2104.05670.

[PBV22] Mathis Petrovich, Michael J. Black, and Gul Varol. Temos: Generating diverse human motions from textual descriptions. *ArXiv*, abs/2204.14109, 2022. URL: https://api.semanticscholar.org/CorpusID:248476220.

[PCA+21] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021.

[PMA16] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, dec 2016. URL: http://dx.doi.org/10.1089/big.2016.0028, doi:10.1089/big.2016.0028.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL: https://arxiv.org/abs/2103.00020, arXiv:2103.00020.

[TLYK17] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. *CoRR*, abs/1707.04993, 2017. URL: http://arxiv.org/abs/1707.04993, arXiv:1707.04993.

[TRG+22] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. *ArXiv*, abs/2209.14916, 2022. URL: https://api.semanticscholar.org/CorpusID:252595883.

[VCST+23] T. Van Criekinge, W. Saeys, S. Truijen, L. Vereeck, L. Sloot, and A. Hallemans. A full-body motion capture gait dataset of 138 able-bodied adults across the life span and 50 stroke survivors. figshare. Collection, 2023. URL: https://springernature.figshare.com/collections/A_full-body_motion_capture_gait_dataset_of_138_able-bodied_adults_across_the_life_span_and_50_stroke_survivors/6503791/1.

[vdOVK17] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *CoRR*, abs/1711.00937, 2017. URL: http://arxiv.org/abs/1711.00937, arXiv:1711.00937.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL: http://arxiv.org/abs/1706.03762, arXiv:1706.03762.

[ZCP+22] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:4115–4128, 2022. URL: https://api.semanticscholar.org/CorpusID:251953565.

[ZHL+24] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7368–7376, Mar. 2024. URL: https://ojs.aaai.org/index.php/AAAI/article/view/28567, doi:10.1609/aaai.v38i7.28567.

[ZMR+23] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:2430–2449, 2023. URL: https://api.semanticscholar.org/CorpusID:263796023.

[ZZC+23] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xiaodong Shen. Generating human motion from textual descriptions with discrete representations. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14730–14740, 2023. URL: https://api.semanticscholar.org/CorpusID:255942203.