The Thesis Committee for Jordan Guy Voas
certifies that this is the approved version of the following thesis:

What is the Best Automated Metric for Text to Motion Generation?

SUPERVISING COMMITTEE:

Raymond Mooney, Supervisor

Qixing Huang

What is the Best Automated Metric for Text to Motion Generation?

by

Jordan Guy Voas

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Computer Science

The University of Texas at Austin

December 2023

Acknowledgments

To everyone who has supported me along my way, thank you very much. Whether it was my family, a driven educator, or colleague I would not be where I am if each of you had not been present throughout the journey.

Additionally, this research was partially supported by NSF NRI Grant IIS-1925082 and NSF IIS-2047677.

Abstract

What is the Best Automated Metric for Text to Motion Generation?

Jordan Guy Voas, MSCompSci
The University of Texas at Austin, 2023

SUPERVISOR: Raymond Mooney

There is growing interest in generating skeleton-based human motions from natural language descriptions. While most efforts have focused on developing better neural architectures for this task, there has been no significant work on determining the proper evaluation metric. Human evaluation is the ultimate accuracy measure for this task, and automated metrics should correlate well with human quality judgments. Since descriptions are compatible with many motions, determining the right metric is critical for evaluating and designing meaningful training losses for supervising generative models. This paper systematically studies which metrics best align with human evaluations and proposes new metrics that align even better. Our findings indicate that none of the metrics currently used for this task show even a moderate correlation with human judgments on a sample level. However, for assessing average model performance, commonly used metrics such as R-Precision and rarely used coordinate errors show strong correlations. Several recently developed metrics are not recommended due to their low correlation compared to alternatives. Additionally, multiple novel metrics which exhibiting improved correlation and potential for future use.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

High-quality human motion generation in animation has a wide range of applications, from creating realistic CGI in cinema to enabling context-aware character movement in video games. The increasing interest in generating human motions from natural language descriptions (text-to-motion) is evident Lin et al. (2018); Ahuja and Morency (2019); Punnakkal et al. (2021); Ghosh et al. (2021); Zhang et al. (2022); Guo et al. (2022b); Delmas et al. (2022). Natural language offers a convenient and expressive means for controlling generative models, similar to image Ramesh et al. (2022) and video Singer et al. (2022) generation. Users can specify the desired actions or poses they want the motion to exhibit, such as global transitions like running, jumping, and walking, or localized actions like throwing or kicking. They may also indicate concurrent sub-motions or sequential motions with fluid or distinct transitions. The generated motion sequence should accurately match the prompt while appearing natural.

Determining the best automated metric for human motion generation from natural language prompts is crucial for developing effective models. Although human judgment is considered the gold standard, comparing large sample sizes is time-consuming and expensive. Stochasticity in recent models adds to this challenge, necessitating extensive repetitions for accurate results.

1.1 Research Statement

Our objective is to identify the best automated metric for evaluating language-conditioned human motion generations, with "best" referring to the metric most closely correlated with human judgments. While various automated metrics have been proposed Ahuja and Morency (2019); Ghosh et al. (2021); Guo et al. (2022a) and some works have conducted comparative human evaluations Guo et al. (2022a); Petrovich

et al. (2022), none have directly addressed this question. Developing appropriate automated metrics correlated with human judgments has been vital in fields such as machine translation Papineni et al. (2002); Zhang et al. (2019), and we believe it is essential for advancing text-to-motion methods.

To complement existing metrics, we propose novel ones that improve correlation with human judgment. Some of these metrics are differentiable and could enhance optimization when integrated into training losses.

Multiple distinct aspects should be considered when assessing the quality of generated human motions. We evaluate human motion quality by focusing on the following:

- **Naturalness**: How realistic is the motion to a human viewer? Unnatural motions exhibit inhuman or improbable poses and transitions or display global transitions without appropriate actions.
- **Faithfulness**: How well does the generated motion align with the natural language prompt? Unfaithful motions will omit key components or include irrelevant ones.

1.2 Contributions

Our main contributions are:

- A dataset of motion-text pairs with human ratings of *Naturalness* and *Faithfulness* for evaluating automated metrics.
- A critical evaluation of existing text-to-motion automated metrics based on correlation with human judgments.
- The development of several high-performing automated metrics for future architecture comparison and development.

Chapter 2: Related Work

We review prior research on human motion generation, which includes both unconditioned and conditioned generation, and discuss the evaluation metrics used in previous studies.

2.1 Human Motion Generation

Early unconditioned human motion generation approaches employed statistical generative models Ikemoto et al. (2009); Mukai and Kuriyama (2005), while more recent models have adopted deep learning techniques. Some studies have applied Variational Autoencoder (VAE) models Kingma and Welling (2013) for motion forecasting based on historical fragments Tulyakov et al. (2017); Aliakbarian et al. (2020); Ling et al. (2020); Rempe et al. (2021). Others have used Generative Adversarial Networks (GAN) Goodfellow et al. (2014) to enhance the quality of generated motion Barsoum et al. (2017). Normalization Flow Networks have also been explored Henter et al. (2020). The majority of these methods employ joint-based frameworks, utilizing variants of the SMPL Loper et al. (2015b) body model, which represents the body as a kinematic tree of connected skeletal segments.

2.1.1 Conditioned Human Motion Generation

For conditioned motion generation, various types of conditioning exist. Some studies have conditioned on fixed action categories, which simplifies the task compared to natural language conditioning but limits diversity and controllability. Action2Motion Guo et al. (2020) employs a recurrent category conditional VAE, while ACTOR Petrovich et al. (2021) uses a category-conditioned VAE with Transformers Vaswani et al. (2017).

Natural language conditioning allows for fine-grained motion control, enabling

temporal descriptions and specifying actions for different body parts. Early efforts utilized a Seq2Seq approach Lin et al. (2018). Other studies learned a joint embedding space projection for both modalities Ahuja and Morency (2019); Ghosh et al. (2021) and generated motions using a decoder. Some research applied auto-regressive methods Guo et al. (2022a), encoding text and generating motion frames sequentially. Recent approaches, such as TEMOS Petrovich et al. (2022), use stochastic generation to produce diverse outputs. The most recent works employed diffusion-based models like FLAME Kim et al. (2022), MotionDiffuse Zhang et al. (2022), or MDM Tevet et al. (2022) and achieved top performance on current evaluation metrics.

Related tasks have also been investigated, such as Music-to-Dance Li et al. (2020) and EDGE Tseng et al. (2022), which conditions motion generation on music. Some models treat the task as reversible, captioning motions and generating them from language prompts Guo et al. (2022b). Others generate stylized character meshes to pair with the generated motions, conditioned on language prompt pairs Youwang et al. (2022); Hong et al. (2022).

2.2 Metrics for Automated Evaluation of Human Motions

Various metrics have been used to evaluate text-to-motion. Language2Pose Ahuja and Morency (2019) employed Average Position Error (APE) and pioneered the practice of dividing joints into sub-groups for different versions of APE. Ghosh et al. (2021) introduced Average Variance Error and also considered versions dependent on which joints (root versus all) are being used and whether global trajectories are included. TEMOS Petrovich et al. (2022) and FLAME Kim et al. (2022) adopted similar methods, but recent works have moved away from these metrics despite no study establishing them as poor performers.

Guo et al. (2022a) developed a series of metrics based on their previous work for category-conditioned motion generation, advocating for Frechet Inception Distance (FID) Heusel et al. (2017), which is commonly used in image generation and measures

output distribution differences between datasets. Guo et al. (2022a) also included R Precision, a metric based on retrieval rates of samples from batches using embedded distances, metrics to evaluate diversity, as well as one measuring the distance of co-embedding in each modality. These metrics have become standard, used by multiple works Guo et al. (2022b); Tevet et al. (2022); Kim et al. (2022); Zhang et al. (2022).

The metrics established by Guo et al. (2022a) rely on a text and motion co-encoder and depend on the quality of the embedding space. Thus, proving the effectiveness of the embedding space is crucial for these metrics if they are to be used for judging competitive model performance.

The GENEA Challenge Kucherenko et al. (2021) provides a collective assessment of co-speech motion generation methods through standardized human evaluations. It divides human judgments into *Human-likeness* and *Appropriateness*, corresponding to our *Naturalness* and *Faithfulness*. Recent findings by the challenge Yoon et al. (2022) indicate that current methods generate natural motions at or above rates for baseline captures but underperform in faithfulness. While not directly applicable to text-to-motion, this research provides valuable data for understanding the performance of current methods and guiding future work in the area, including novel metrics.

Chapter 3: Dataset Collection

3.1 Baseline Models Evaluated

We evaluate four implementations to assess a range of motion qualities and focus on issues relevant to top-performing models: Guo et al. (2022a), TM2T Guo et al. (2022b), MotionDiffuse Zhang et al. (2022), and MDM Tevet et al. (2022). These models, trained on the HumanML3D dataset Guo et al. (2022a), support generating 22 joint SMPL body models Loper et al. (2015a), enabling consistent animation methods for human ratings. We also include reference motions from HumanML3D as a baseline for non-reference-based evaluation metrics.

Although our results provide comparative evaluations of these models, drawing conclusions should be done cautiously, as detailed below. Our aim is not to establish a state-of-the-art system.

3.1.1 Baseline Model Selection

In selecting the text-conditioned motion generation models to include in our study, our primary focus was to obtain a diverse sampling of motions, as each model may be prone to its own distributional differences. However, we also imposed additional constraints on the models studied to ensure their relevance.

Although several older models exist that could potentially increase the diversity of our motion samples, we excluded them due to their subpar performance, which renders the diversity less valuable for future models. Studying these less effective models might be interesting, but their inclusion could compromise our findings by reducing the number of samples we could incorporate for better-performing models and possibly leading to a false impression of which metric is most suitable for them.

The models included in our study were also required to be compatible with and have pretrained weights available for motion generation using 22 joint SMPL

body models. Some models that might have performed well enough to be considered relevant for our study cannot use 22 joint SMPL body models and would, therefore, be unsuitable for scoring against 22 joint reference motions or with 22 joint-based motion encoders.

3.2 Motion Prompt Sample Collection

We sourced motion prompts from the HumanML3D test set. To ensure diverse and representative prompts, we encoded them using the RoBERTa language model's CLS outputs Liu et al. (2019), projected the embeddings onto a low-dimensional space, and randomly sampled from the normal distribution, obtaining 400 unique sample prompts. We discarded duplicates and very short motions.

These prompts generated a dataset of 2000 motions, with 400 motions for each of the five baseline models (including HumanML3D). For models generating fixed-length motions, we used a length of 120 motion frames. All models were generated at the 20 Hz frequency used in HumanML3D.

3.3 Motion Visualization

Recent studies Guo et al. (2022a); Petrovich et al. (2022) utilized stick figure renderings for evaluation, but this approach has limitations. Evaluating motion *Naturalness* using stick figures can be challenging, as they may not be relatable to human observers. Moreover, they often lacked realistic environments, such as walls, floors, lighting, and textures.

To address these limitations, we created high-quality renders using Blender Community (2018), focusing on environmental details and controlled camera movements for smooth and natural motion perception. See Figure 3.1 for examples.

We collected human quality ratings using Amazon Mechanical Turk and a custom UI (Figures 3.2, 3.3, 3.4). To ensure quality, we implemented qualification re-

Figure 3.1: Sampled motion frames with paired descriptions, as used in our human evaluations. Our rendering framework generates pseudo-realistic environments with skin, wall, and floor textures as well as environment lighting and steady camera motions.

quirements, in-tool checks, and post-quality criteria. We hand-picked 25 motion-text pairs from the 2000 motion samples we generated and used them as gold test questions. The remaining annotations were divided into 20-pair batches, each containing five randomly placed gold test samples. We collected three ratings per sample and discarded batches that failed qualification checks.

3.4 Human Quality Ratings Collection

Ratings were presented as natural language descriptions corresponding to Likert Scale ratings (0 to 4). Annotators had access to a tooltip with detailed descriptions during the task. *Naturalness* descriptions, as provided to the annotators, were:

- **Very Unnatural**: Does not maintain a human body shape. Majorly glides without taking any appropriate actions such as walking or jumping. Possess very jerky movements.

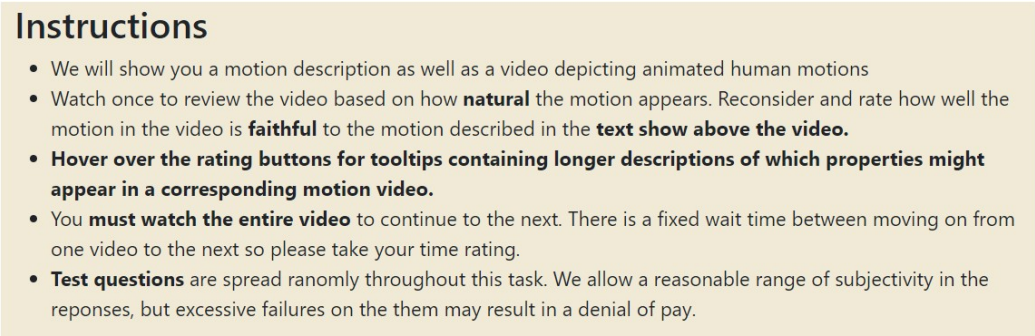

Figure 3.2: Instructions for raters in human judgment evaluations.

- **Unnatural**: Maintains a mostly human body shape but moves unnaturally. Glides moderately without taking any appropriate actions such as walking or jumping. Possess some jerky movements.

- **Neutral**: Possess some amount of gliding without taking appropriate actions such as walking or jumping. Moves its limbs or body in ways that are not humanly likely but are possible.

- **Realistic**: Possess very slight amounts of gliding without taking appropriate actions such as walking or jumping. Moves its limbs or body in a slightly rigid fashion.

- **Very Realistic**: Always moves in a human like way. Does not move without taking appropriate actions such as walking or jumping.

Faithfulness descriptions were:

- **Dose Not Describe**: Completely different motion which corresponds in no way to the given description.

- **Slightly Describes**: Has very slight resemblance to the given description (such as standing up when its describing walking) but otherwise does not correspond with the given description

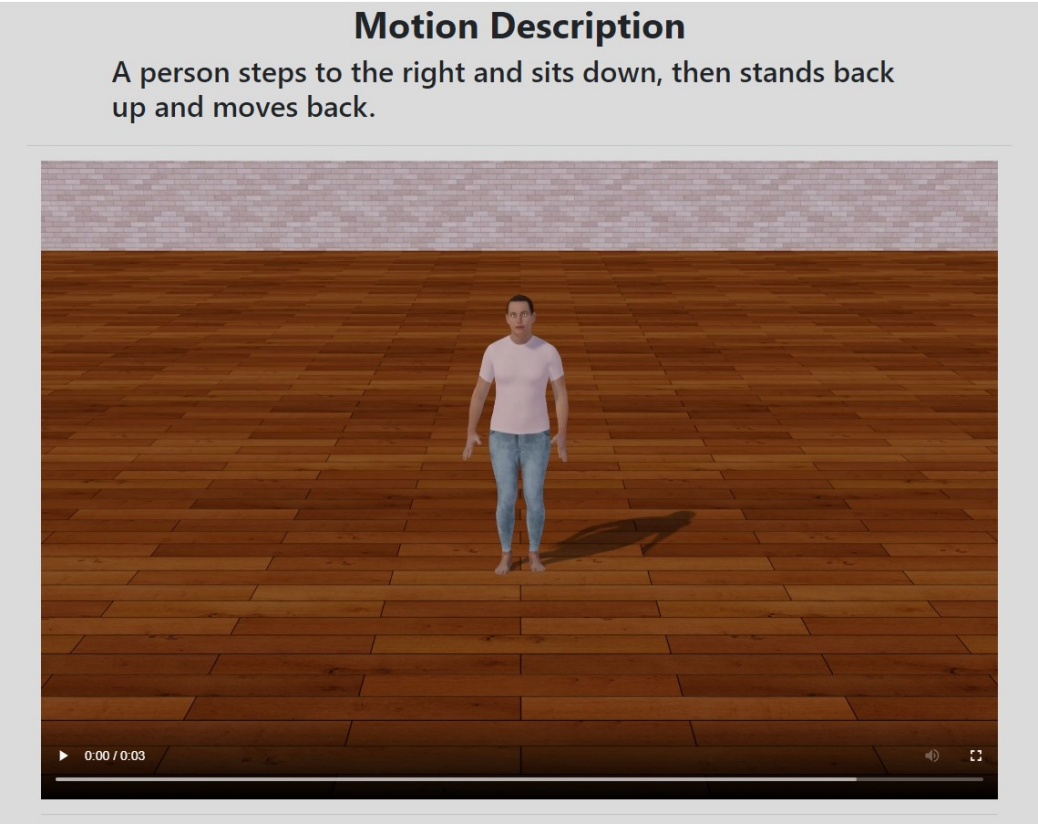

Figure 3.3: UI motion viewing section, situated below the instructions and above the rating selection.

- **Moderately Describes**: Follows the description moderately, but deviates from it in significant ways such as adding major actions or leaving out a portion of the description.
- **Greatly Describes**: Leaves out minor details of the description or includes minor unmentioned actions. For example, the person may be walking as mentioned in the description but also doing actions with their hands that were unmentioned.
- **Perfectly Describes**: Perfectly described by the description, with no actions left out or incorrectly included.

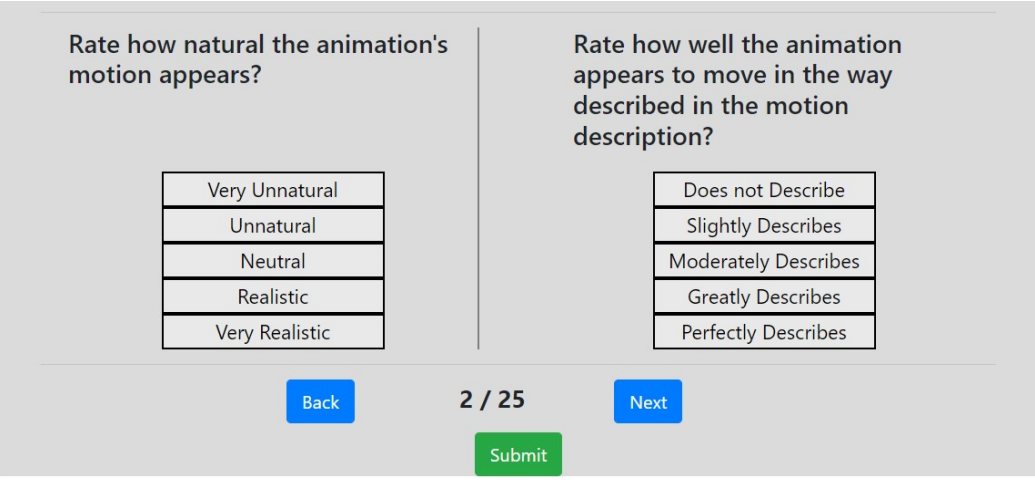

Figure 3.4: Rating selection UI, located below the motion viewing section. Detailed descriptions for each rating option were provided as tooltips upon hovering.

Ratings were rejected if more than two of the five gold test questions deviated by more than one from the "correct" answer. This leniency allowed for subjectivity, missed details on either our or their side, and slight rating scale understanding differences. Random guessing would pass a single question forty to sixty percent of the time, but over the ten independent ratings would be detected with a high likelihood. Significant deviations in rating scale understanding would also be flagged and filtered out with this approach.

We removed samples with less than three ratings for each of the five model types, resulting in 1400 rated motion-text pairs (280 distinct motion prompts for each baseline model). Averaging the three independent ratings provided overall *Naturalness* and *Faithfulness* values. Figure 3.5 displays the dataset's distribution to be generally normal, while Table 3.1 shows high inter-annotator agreement (Krippendorff's Alpha) was obtained.

In-tool quality checks required watching the entire video before progressing, capped the rate of progression to 12 seconds per sample, and ensured all ratings were entered. These measures aimed to prevent rushing and encourage thoughtfulness. Qualification requirements included residing in the U.S., completing over 1000 hits,

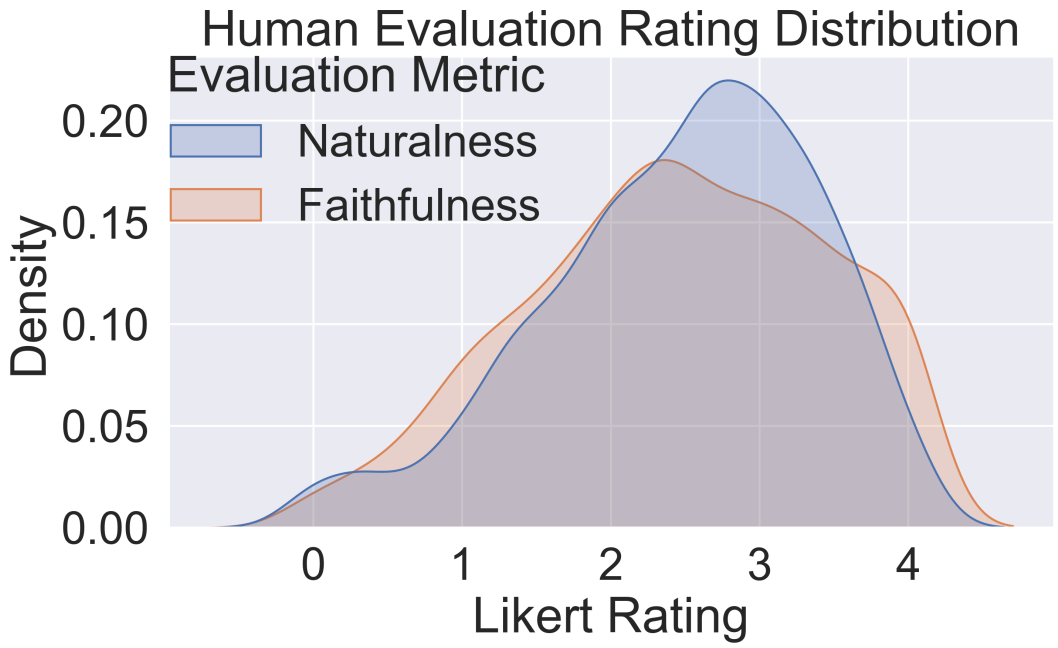

Figure 3.5: Human judgment distribution for all samples. Averages from three annotations are shown with a KDE smoothing filter (bandwidth 0.85) applied. Pearson's correlation between metrics is found to be 0.63 at the sample level.

and a minimum 98% acceptance rate.

3.4.1 Ethics and Compensation

Quality checks were disclosed in the task instructions (Figure 3.2). We paid $1.25 per HIT, equating to at least $12 per hour, considering 25 samples per HIT and a 15-second expected per sample completion rate.

3.5 Model Level Comparisons

The primary goal of this study was not to compare the performance differences between the baseline models included. Consequently, we did not take measures to ensure that our collected data would constitute a suitably valid sample set for

IAA (Krippendorff's Alpha)	
Naturalness	Faithfulness
0.647	0.701

Table 3.1: Inter-annotator agreement across all replicated MTurk samples. Results indicate substantial but non-perfect agreement.

assessing the overall performance of each model. This is due to two main reasons: insufficient sample size and inadequate repetitions per sample. To make our study comparable, at the model level, to the evaluations conducted in each work's independent studies, we would need to evaluate the entire HumanML3D test set. Moreover, we would need to execute multiple (typically 10) generation repetitions per sample for models with stochastic generative properties. However, our study did not fulfill these requirements.

Keeping these caveats in mind, we present the mean distributions of human judgments for each of the baseline models in Figure 3.6. As anticipated, the ground-truth motions in the HumanML3D dataset exhibit a clear advantage in terms of *Naturalness*. Their lead in *Faithfulness* is less pronounced when compared to the best-performing generative baselines, Guo et al. (2022b) and MotionDiffuse Zhang et al. (2022). This might indicate quality limitations in the ground-truth descriptions of the HumanML3D dataset, which were human-annotated, and qualitative analysis by the authors revealed some poor descriptions.

Another intriguing observation is the minimal observed performance difference between the two leading generative baseline models Guo et al. (2022b) and MotionDiffuse Zhang et al. (2022), despite MotionDiffuse reporting gains in R-Precision, FID, and Multimodal Distance over Guo et al. (2022b). These findings may stem from the model coverage limitations we discussed earlier. Alternatively, they could arise from noise in our human judgment ratings or be indicative of the imperfect correlation of these metrics observed in our study.

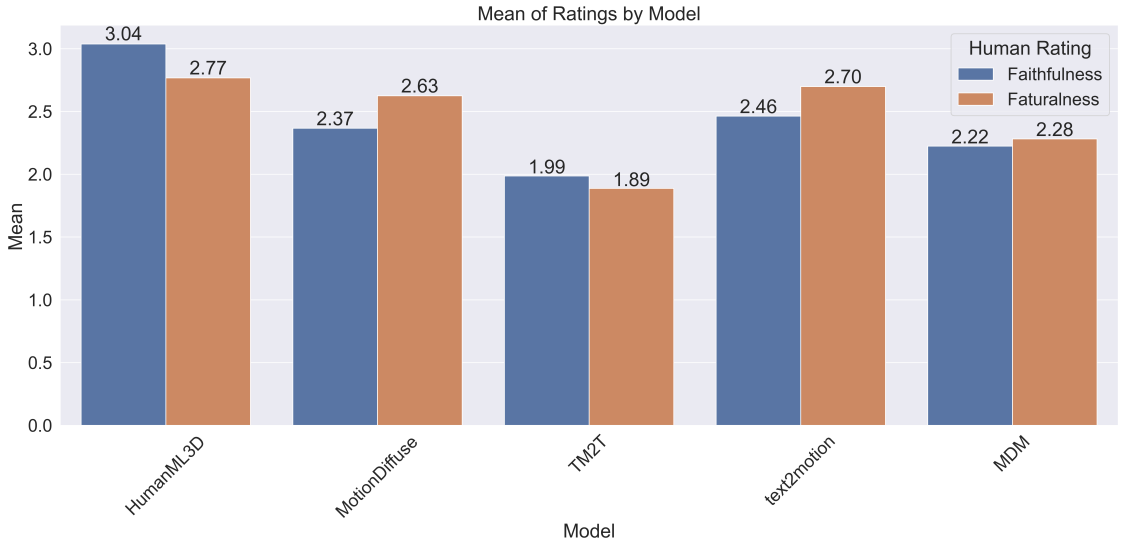

Figure 3.6: Mean rating scores from human judgment for each model utilized (including ground truth). As expected, ground truth achieves a slight lead. Note that these results are not conclusive for evaluating the capabilities of these models, as the sample sizes are relatively small and do not accurately account for stochastic generation.

3.6 Data Availability

We provide all collected human motion judgments for both *Naturalness* and *Faithfulness* as supplemental data to this work. Additionally, we include the prompts used to generate each motion sample and the motion generations themselves. Our data zip files contain the following files and formats:

- ratings_and_captions.csv: A comma-separated file with each line containing, in order; restricted sample index, model name, original sample index, the mean value for the sample for human judgment of Naturalness, the mean value for the sample for human judgment of Faithfulness, and the lowercase textual prompt it corresponds with/was generated from. The model names are either HumanML3D, MotionDiffuse, text2motion, TM2T, or MDM. The original sample index is the index from a larger superset of 400 samples for each model, which were initially included in the human annotation collection. After quality control

and restricting to samples with valid judgments across all baseline models, this number was reduced to 280 samples per model, as indicated by the restricted sample index.

- AMASS_motion_ModelName_SampleIndex.npy: The motion sequences utilized in this study, each in AMASS format for a 22 Joint SMPL body model. All other joints have zeroed out values. The value SampleIndex in the file name corresponds with the original sample index value from the ratings_and_captions.csv file.

We anticipate that this data will be valuable for reproducing our results and facilitating future testing of innovative text-conditioned motion generation metrics.

Chapter 4: Evaluated Metrics

We incorporate most automated metrics from recent works as well as several new ones. We assess each metric's correlation with samples on both individual and model levels, whenever possible. Sample level correlations are computed on individual sample scores across baselines, reflecting the metric's capability to evaluate single generations. Model level correlations are determined using the mean metric score for all samples generated by a specific baseline model, which are then correlated with the mean human rating for the corresponding baseline model. This assesses how well the metric can judge model performance ranking. These levels can be distinct since metrics with outlier failures may negatively impact sample level evaluation but have reduced effects when averaged over many samples. Ideal metrics should excel at both levels.

4.1 Coordinate Error (CE) Metrics

Average Error (AE), also known as Average Position Error (APE) when applied to joint positions Ahuja and Morency (2019), and Average Variance Error (AVE) Ghosh et al. (2021) are reference-based metrics employed in early works but have become less common recently. They calculate the mean L2 errors between reference and generated values, either absolute or as variance across frames, for each joint in the motion. We refer to these as coordinate error (CE) metrics, defined as:

$$AE = \frac{1}{JT} \sum_{j \in J} \sum_{t \in T} \|X_t[j] - \hat{X}_t[j]\|_2 \tag{4.1}$$

$$\sigma[j] = \frac{1}{T-1} \sum_{t \in T} (X_t[j] - \hat{X}_t[j])^2 \tag{4.2}$$

$$AVE = \frac{1}{J} \sum_{j \in J} \|\sigma[j] - \hat{\sigma}_t[j]\|_2 \tag{4.3}$$

Where j represents a joint from all 22 joints J, and t denotes a motion frame from the motion sequence T. We matched frame lengths for reference and generated motions by clipping the longer one.

We investigated CE metrics on positional values and their variations on positional derivatives, such as velocity and acceleration, calculated using frame-wise differences. Additionally, we evaluated these metrics on combinations of position and its derivatives. Similar to Ghosh et al. (2021), we examined three joint groupings for CE metrics: root only, all joints excluding the root (Joint), and all joints (Pose). Prior works Ghosh et al. (2021); Ahuja and Morency (2019) suggested that AE on the root joint best aligns with human judgments.

We hypothesized that this effect might stem from scaling issues when the root translations are included in combined calculations with other joints, causing their errors to dominate the metric. To test this, we explored potential root joint scaling factors, altering their transitions contribution to the metric's final score for the mean. We also examined the impact of scaling factors on each component when calculating combined position-velocity (PV) or position-velocity-acceleration (PVA) CE. These methods act as a weighted average, with scaling factors increasing or decreasing the root joint or component errors.

4.2 Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID) Heusel et al. (2017) is a widely used metric for generative tasks, which measures the alignment between two distributions. To compute FID, one must first obtain the mean and variance of each distribution from a large sample size. In generative tasks, these typically correspond to the reference samples (a valid distribution) and the generative model samples. A lower FID indicates better alignment between the generative and reference distributions. FID is calculated as follows for distributions D_1 and D_2:

$$FID(D_1, D_2) = |\mu_1 - \mu_2| + tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}}) \tag{4.4}$$

As FID is only accurate with large sample sizes, we report correlations for FID at the model level only and do not report correlation scores for individual samples.

4.3 R-Precision

R-Precision is a distance-based metric that measures the rate of correct motion-prompt pair matchings from a batch of random samples. Both motions and prompts are projected into a co-embedding space, and Euclidean Distance calculations are used to rank pair alignments. Scores of one are received if the correct matching is made within a rank threshold (Retrieval Allowance), and zero otherwise. Averaged over numerous samples, this provides a precision of retrieval metric.

Higher Retrieval Allowance thresholds yield higher R-Precision scores, as they are more forgiving of imperfect embedding spaces and account for multiple motions described by the same prompt randomly being included in the batch. R-Precision scores for thresholds of 1-3 are commonly reported. We analyze the correlation for R-Precision scores with thresholds of 1-20 and hold the batch size to 32, following common practice Guo et al. (2022a).

4.4 Multimodal Distance

This metric measures the distance between the generated motion embedding and the co-embedding of the prompt used for generation. When the two encoders (text and motion) are well-aligned in the embedding space, low scores suggest motions closely matching the prompt, while high scores indicate significant deviations in features Guo et al. (2022a).

4.5 Nearest Neighbor Captioning Methods

Our novel metric, Nearest Neighbor Captioning (NNC), retrieves the closest motions in a reference dataset based on Euclidean Distance in a motion embedding space. By identifying the nearest motions in the reference dataset to those produced by the generative model, we can evaluate these nearest neighbor captions using a language similarity metric, comparing them to the original prompt. The general workflow of these metrics is shown in Figure 4.1 and an example operation can be seen in Figure 4.2.

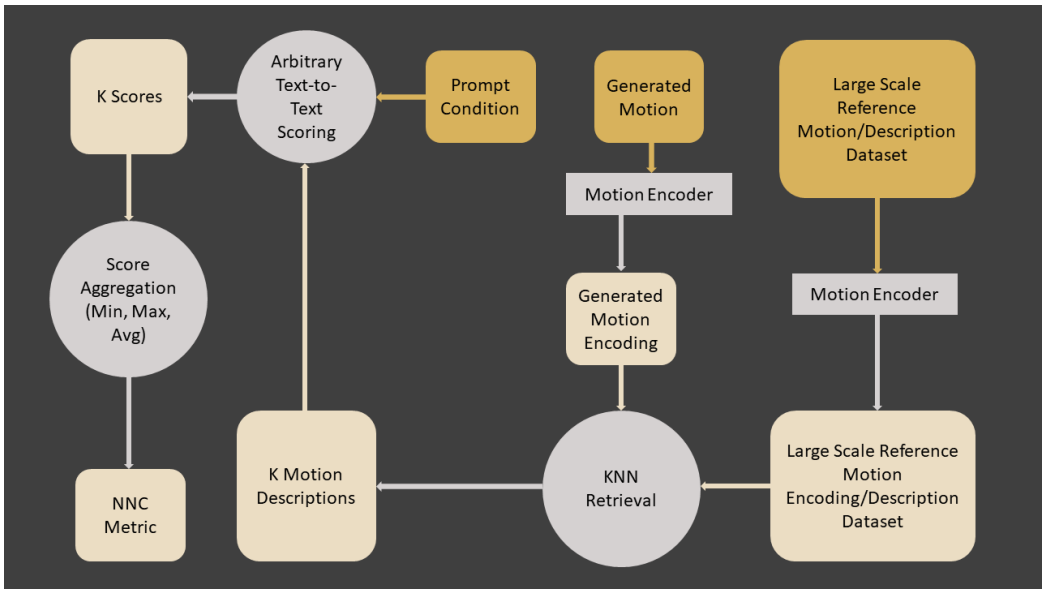

Figure 4.1: General operation of NNC metrics. Yellow items are given as input, greys are operations or models used, and tan is the intermediate results of the process.

NNC offers several key benefits. It leverages well-established text-to-text evaluation methods, which have been more extensively studied than motion-to-motion evaluation metrics. Furthermore, it creates a virtuous cycle of opportunities for future advancements, such as advanced motion encoders, expanded reference datasets, and improved text-to-text scoring methods. While direct motion-to-motion comparisons may seem logical, they are hindered by under-specification in the generative prompt. A single motion can correspond to multiple captions, and a single caption

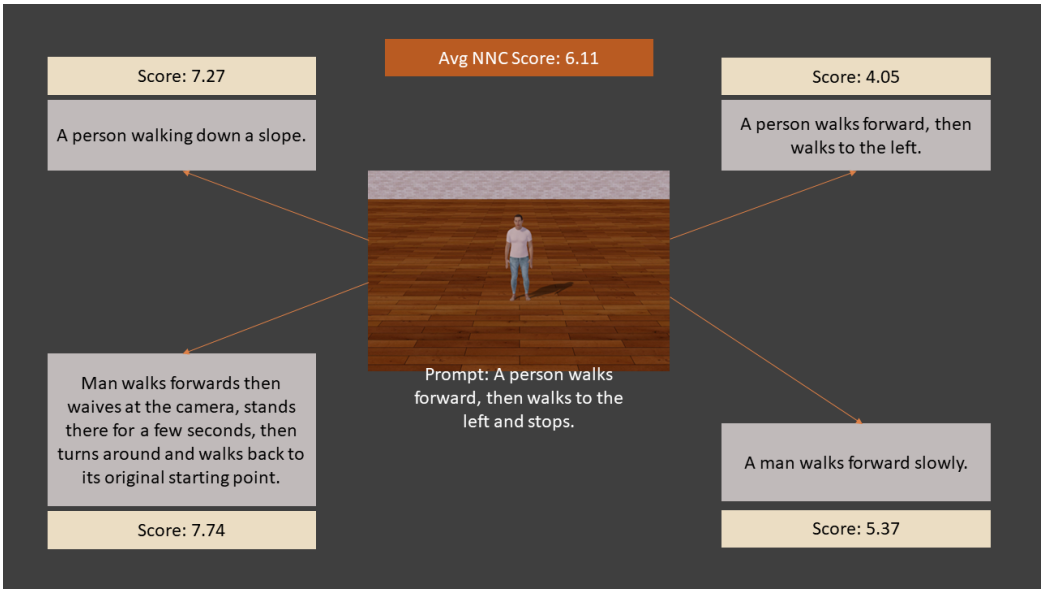

Figure 4.2: A small scale NNC example from our evaluation set. In this case, Match Distance was used, so low scores indicate improved feature alignment.

can correspond to multiple motions. By retrieving a set of nearest neighbor captions from the reference dataset, NNC metrics can mitigate under-specification effects using min, max, or averaging operations across the scores of the motions in the set.

We explore high-performing text-to-text scoring methods, such as BERTScore (Recall) Zhang et al. (2019) and BLEURT Sellam et al. (2020), and examine scoring based on the Euclidean distance of text embeddings from CLIP Radford et al. (2021) or the aligned text encoder developed with the Standard motion encoder (described below), termed Match Distance.

Our reference dataset is the training split of HumanML3D. Since all sample prompts are drawn from the test set, there are no exact matches between the evaluation samples and NNC references.

4.6 Encoders Used

To calculate FID, R-Precision, and NNC, we first project motion features into an embedding space using an encoder model. The standard encoder commonly used was developed by Guo et al. (2022a) and exhibits some correlation with human judgments from prior studies. We refer to its embedding space as the Standard Embedding Space (Std. Emb.). We also investigate alternative encoders, testing the motion encoder from the Guo et al. (2022a) generative model. This encoder produces a high-dimensional output of $R^{49\times512}$, so we examine embeddings reduced to R^{512} via min, max, and average operations, or by using the last index of the 49. These spaces are called T2M Emb., with Min, Max, Avg, or Lst as suffixes indicating the dimensionality-reduction operation used.

4.7 Metrics for Motion Diversity

The increasing popularity of FID, Multimodal Distance, and R-Precision metrics has led to a growing interest in the metric of Diversity, proposed alongside them. However, in this study, we did not examine Diversity for several reasons. Measuring diversity necessitates a distinct experimental setup compared to other metrics, as multiple generations for a single prompt are required. Furthermore, there is currently no well-defined method for assessing diversity from a human perspective. Present metrics evaluate the diversity of embedded representations, but it remains unclear how annotators should be instructed to measure diversity. A separate study should investigate measures of diversity, their alignment with human judgments, and the impact of enhancing a model's diversity of outputs on its *Faithfulness* and *Naturalness* judgments.

4.8 Hyperparameters

Our evaluation involved several hyperparameters, which are listed here. Whenever possible, we utilized pre-existing code, such as the evaluation script that implements FID, R-Precision, and Multimodal Distance from Tevet et al. (2022). R-Precision employed the standard batch size of 32. In our proposed NNC metrics search, we considered up to 32 nearest neighbors. For CE metrics, we searched root scalings ranging from 2^{-15} to 2^{14} in factors of 2. When combining position, velocity, and acceleration components for CE metrics, we performed a grid search to explore all possible combinations of component scalings from 2^0 to 2^9.

Chapter 5: Results Analysis

This section highlights the key findings from our evaluation. Due to the large number of values obtained from grid testing root and component scaling factors for CE metrics and neighbor counts for NNC metrics, we cannot display all results. Instead, we focus on commonly used metrics or those that yield the best results.

We employed Pearson's Correlation Coefficient Sedgwick (2012) to correlate metrics with human judgments, measuring rank ordering and relative distance between metrics as most of our data is interval rather than ordinal. We present model and sample level correlations between *Faithfulness* and *Naturalness* in Table 5.1.

We present the uncorrected correlation values for all metrics. Negative correlations are expected for certain metrics, such as FID or CE, since our human judgment ratings suggest better outcomes with opposing directions. High P-values are observed in many reported correlation values, which is anticipated as they were calculated (for model level results) based on only five samples. Our high performing metrics achieved P-Values near 0.05 at the model level, while our best performing sample level metrics a (Pearson's of 0.2 or above) had very low P-Values.

5.1 Coordinate Error Metrics Results

CE metric results are presented in Figures 5.2 and 5.1. Despite relying on only a single reference, CE metrics show weak correlations with human judgments for both *Faithfulness* and *Naturalness* at the sample level. Performance largely depends on non-Root transitions, with Joint POS AE and Joint POS AVE outperforming pure Root-based metrics. Root scaling cannot surpass Joint metrics, and our derivative-based methods do not match positional ones. Combining components only achieves results comparable to Joint POS-based metrics. Notably, AE performs better than AVE at the sample level with a significant margin (0.1 Pearson's).

Pearson's Correlation	
Sample Level	Model Level
0.62	0.83

Table 5.1: Correlation between *Naturalness* and *Faithfulness*.

At the model level, CE-based metrics strongly correlate with human judgments. Root-only traditional AE metrics achieve nearly 0.75 Pearson's, while Root AVE metrics surpass AE with approximately 0.91 Pearson's. Interestingly, Joint versions are unreliable on their own at the model level, suggesting that the main components of model evaluation can be derived from Root transitions alone. This supports claims by Ghosh et al. (2021). Root scaling enhances both metrics, with AVE nearing perfect correlation. Utilizing velocity derivatives benefits AE at the model level, and combining positions, velocity, and acceleration for both AVE and AE yields versions with greater than 0.99 Pearson's.

5.1.1 Root Scaling Exploration

We provide visualizations with scaling factors in Figures 5.3 and 5.4 to investigate the effects of root scaling on Pose CE metrics. Consistent with previous observations, model-level correlations improve (i.e., more negatively correlated) when additional weight is placed on Root transitions, except for PV and PVA AE versions. Alternatively, overemphasizing Root transitions significantly degrades performance at the sample level.

5.2 FID, R-Precision, and Multimodal Distance Results

We examine FID, R-Precision, and Multimodal Distance only at the model level for various reasons. FID cannot be calculated at the sample level as it requires distributional statistics over multiple samples. R-Precision provides binary values at the sample level, making it poorly suited for comparison with Likert Scale ratings. It becomes fine-grained only when averaged over samples. Multimodal Distance exhibits

Pearson's Correlation (Multimodal Distance)	
Naturalness	Faithfulness
-0.300	-0.211

Table 5.2: Multimodal Distance correlation with human judgments.

near-zero correlation at the sample level.

Focusing on model-level results, we present FID in Figure 5.5. It achieves acceptable results for *Faithfulness* with 0.71 Pearson's but significantly underperforms for *Naturalness*. Given the low correlation with *Naturalness* and model-level-only comparison, P-Values are notably high. Our samples may provide an unfavorable setting for FID, or it may improve with larger sample sizes.

R-Precision, shown in Figure 5.6 for various Retrieval Thresholds, demonstrates substantial correlations for both human quality judgments, approaching 0.8 in the typically tested range. Our results suggest current Retrieval Thresholds are suboptimally set, with thresholds of 4 and 5 yielding marginally better outcomes. Performance declines at higher thresholds. Since R-Precision and FID share an embedding space, strong R-Precision results may indicate that FID's poor performance is not due to sample selection. Multimodal Distance, presented in Table 5.2, displays weak correlations for both human quality judgments.

The results indicate that R-Precision, and possibly FID, are suitably correlated with human judgments. However, these metrics are less correlated than the CE metrics they replaced, and they preclude single-sample analysis, relying on many samples. Even if these metrics improved with larger sample sizes, an uncertain possibility, they would require substantial enhancements to match even traditional CE metrics such as Root POS AVE.

5.3 Nearest Neighbor Captioning Metrics

We evaluated various accumulation operations, natural-language scorers, and motion encoders for NNC, and present all results in Tables 5.3 - 5.8. Our focus was on model averages, as sample level correlations were close to zero.

Some NCC metric versions show a strong correlation with human judgments, matching or exceeding R-Precision. Despite variable results across neighbor counts for many, several strong performers demonstrate stable performance across all values, as shown in Figure 5.7. We tested mean values across sub-ranges of neighbor counts for stability (1-10, 11-20, and 21-30), which are displayed in the same tables. We observed many low scoring versions becoming increasingly stable at high neighbor counts, while high scoring ones remained insensitive to neighbor counts beyond the moderate range.

NNC Metrics seem more suitable for judging *Faithfulness* than *Naturalness*, in line with theoretical expectations. Our results also show all but one top performer (bolded in the tables) using either a medium or average accumulation function. Maximum and minimum accumulations likely are too sensitive to outliers in the retrieved neighbors. The standard motion encoder appears less suited for NNC metric than the T2M encoder versions.

5.3.1 Recommendations for NNC Parameters

Based on the results we presented, we recommend two NNC metric versions with high mean correlation, low standard deviation, and correct scoring function correlation sign: **Match CS with T2M Min Encoder and Average Accumulation Function**, or **CLIP Dist with T2M Lst Encoder and Average Accumulation Function**. Our recommendations aim to provide reliable performance in evaluating the naturalness and faithfulness of generated captions. According to Figure 5.7, neighbor counts of around 20 are optimal for both recommendations. However, users should choose a range of neighbor counts to indicate stability, similar to R-Precision.

We suggest reporting values for 10, 20, and 30.

5.3.2 NNC Metrics Parameter Stability

Our results show the inconsistency of many NNC metric versions across different neighbor counts, with the standard deviation for Pearson's correlation of *Faithfulness* typically ranging from 0.05 to 0.3. This indicates that improper hyperparameter choice could easily result in an NNC version with a high Pearson's correlation degrading to levels below R-Precision or FID, negating their usefulness.

Thus, we recommend users report the standard deviation of scores across a multiple neighbors (e.g., 32). High standard deviations, particularly ones that significantly alter sample rank order, may indicate poor performance, while low standard deviations tends to imply strong performance. Additionally, using accumulating functions that are resistant to outliers (average or medium) and scoring functions with evidence for multiple strong results (CLIP Dist, Match CS, or Match Dist) may provide stable performance.

However, additional work should be done to improve NNC metrics' stability and performance, as they fall behind the simpler CE metrics when even a single reference is available. Identifying strong and diverse reference datasets, better scoring functions, or motion encoders will cumulatively benefit NNC metrics and may make them an valuable metric in cases without reference motions.

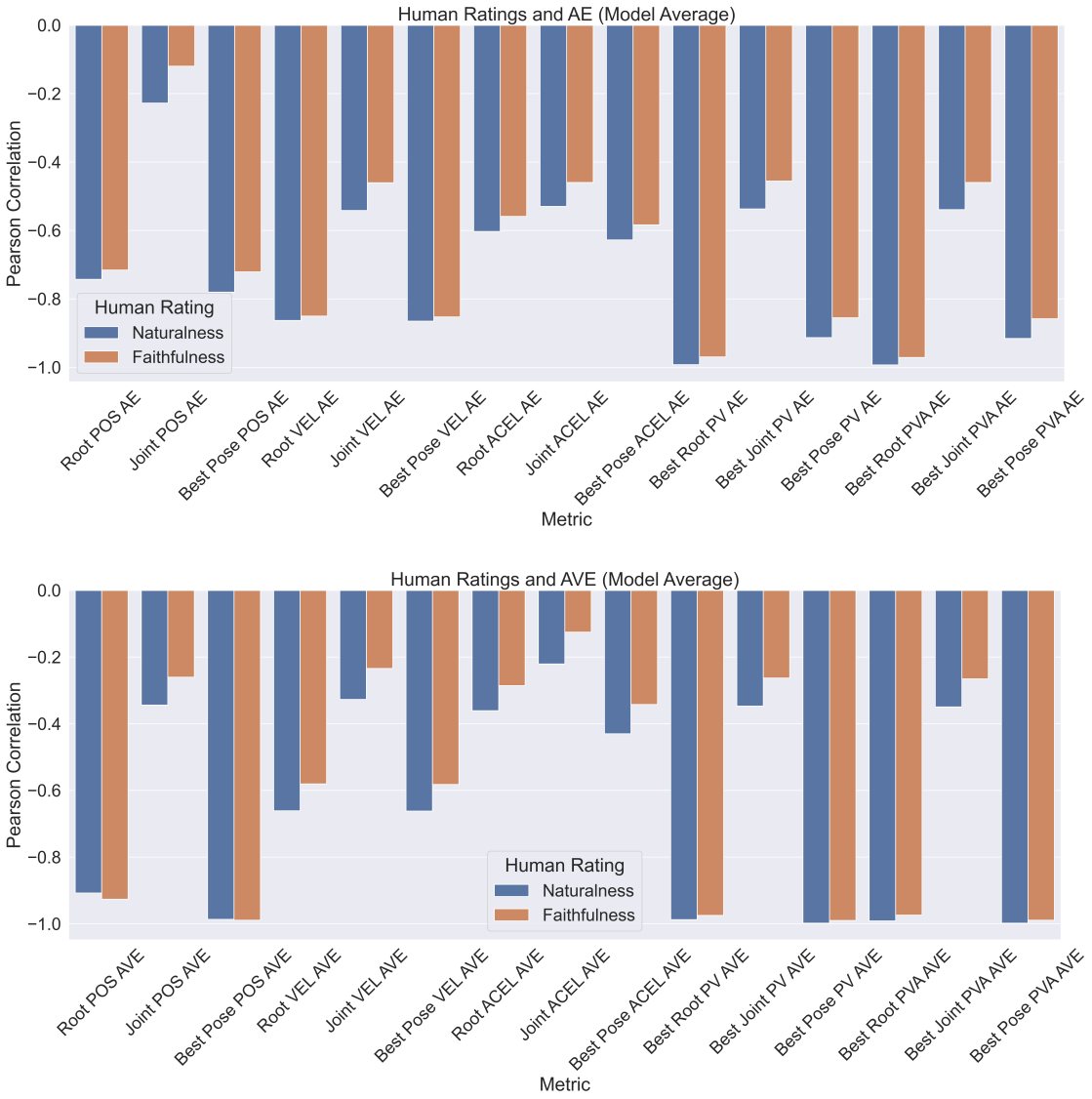

Figure 5.1: Model level correlations of CE metrics with human judgments. "Best" metrics use the highest performing settings for root joint or component scaling.

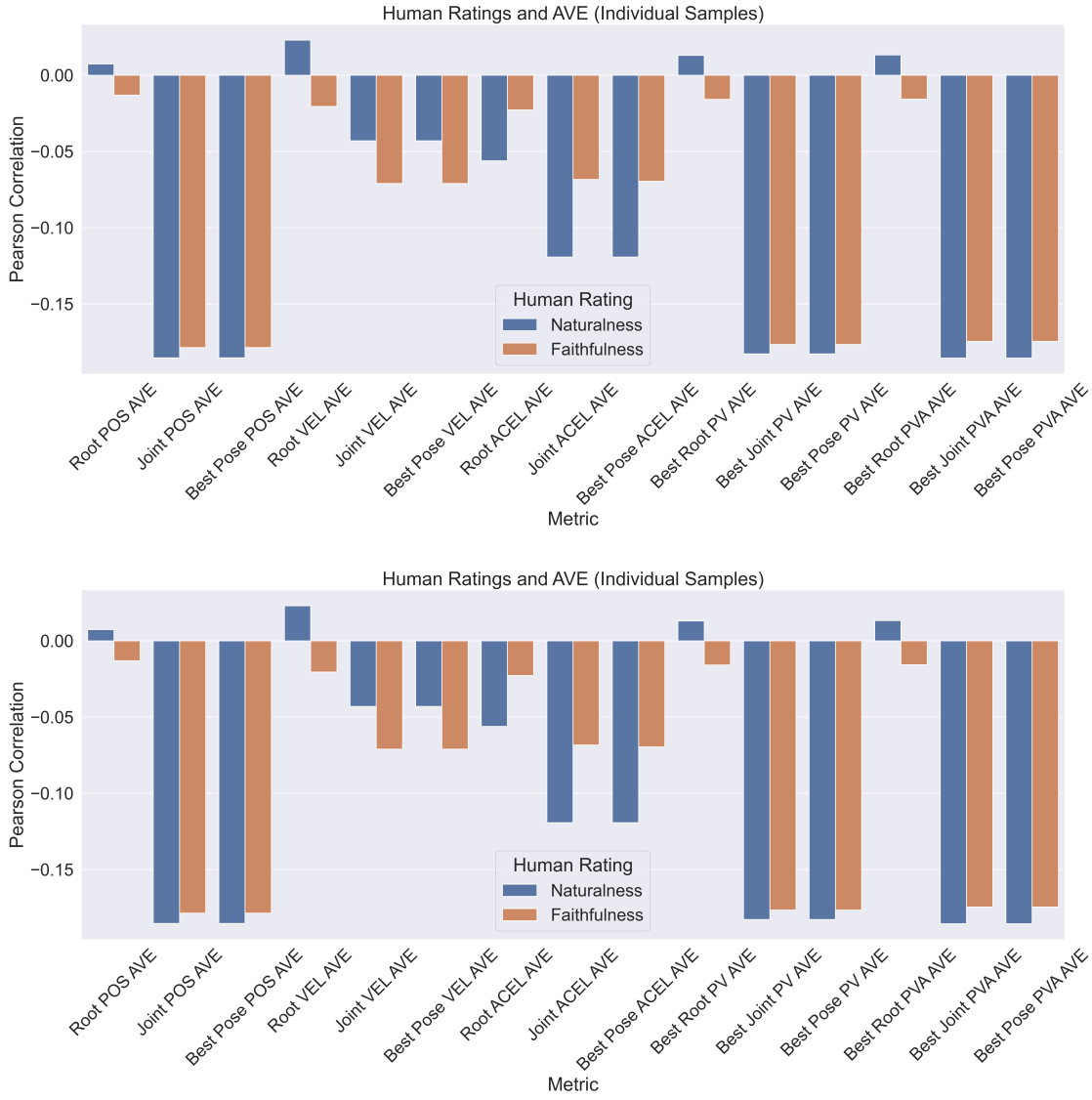

Figure 5.2: Sample level correlations of CE metrics with human judgments. "Best" denotes versions using highest performing settings for scaling root joint or components.

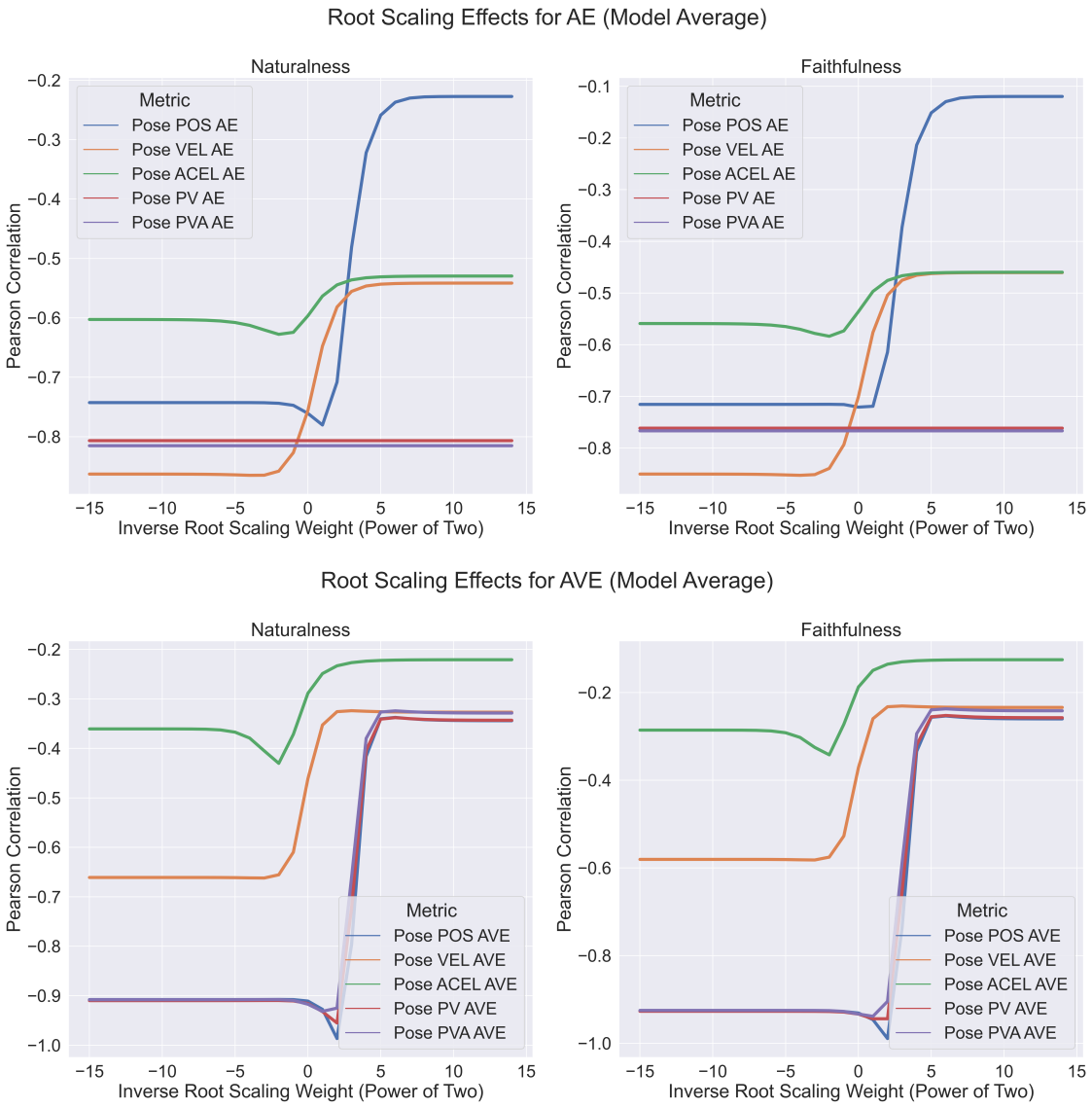

Figure 5.3: Model level examination of root joint scaling effects on CE metrics using all joints.

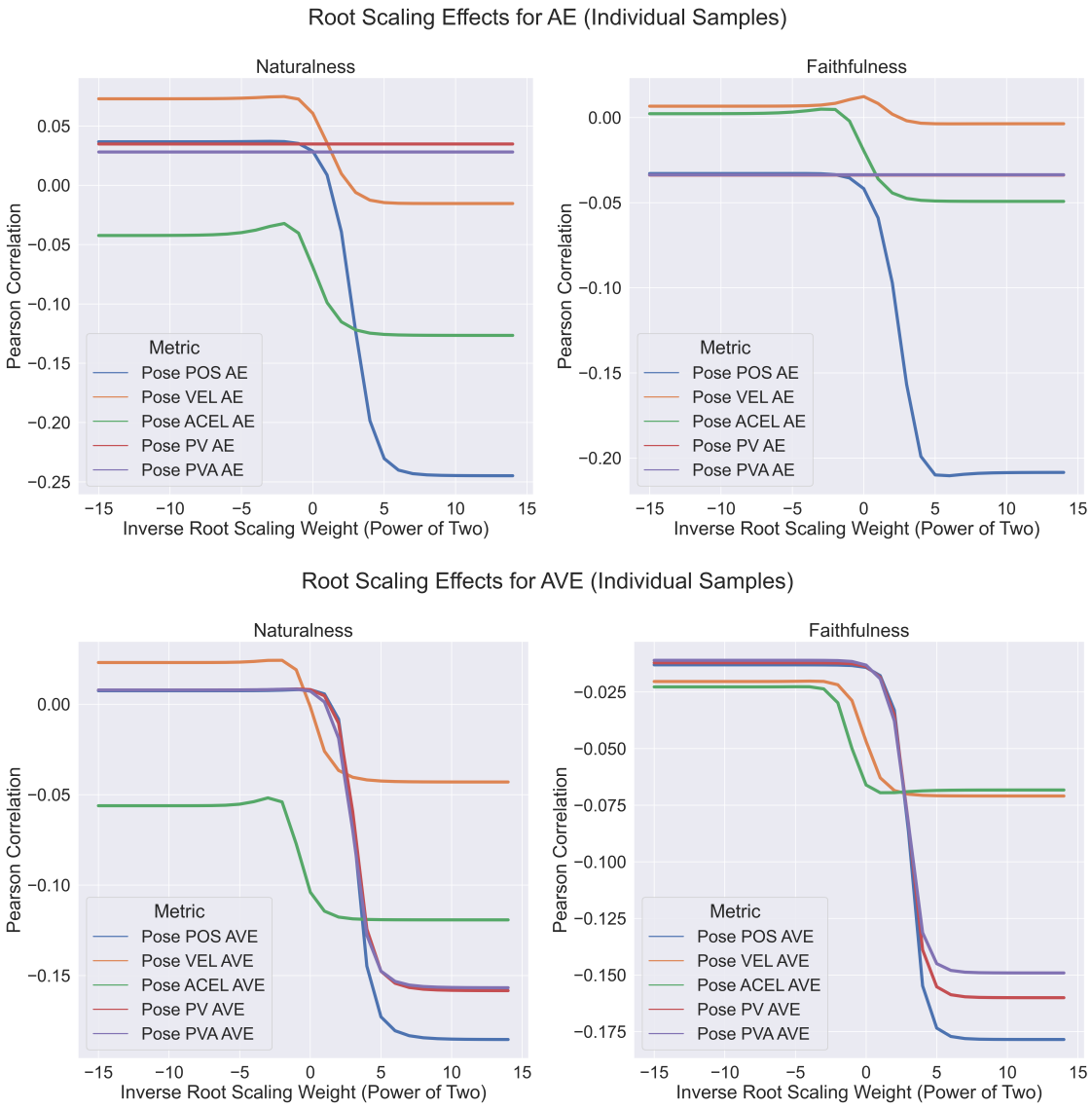

Figure 5.4: Sample level examination of root joint scaling effects on CE using all joints.

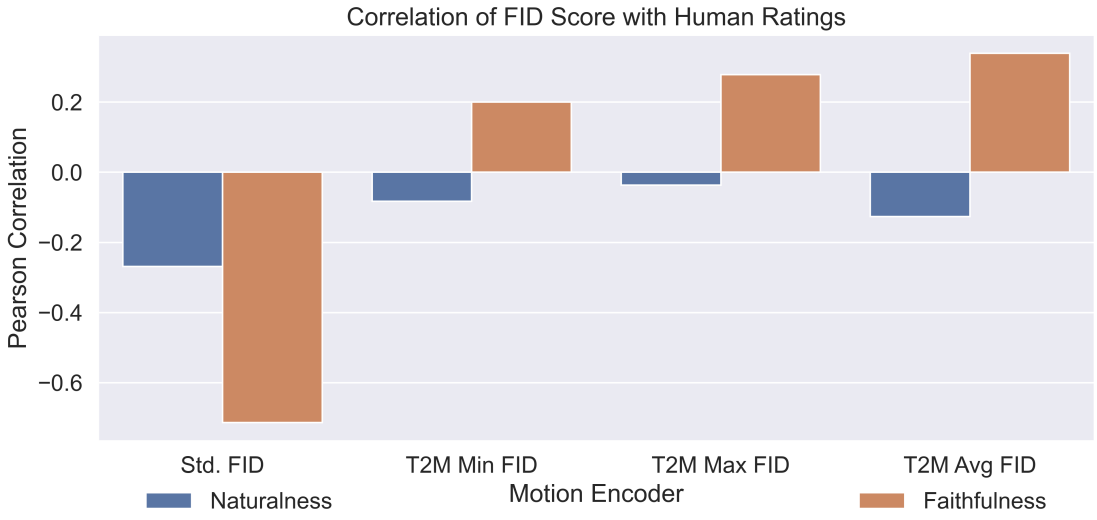

Figure 5.5: FID correlation with human judgments for various motion encoders.

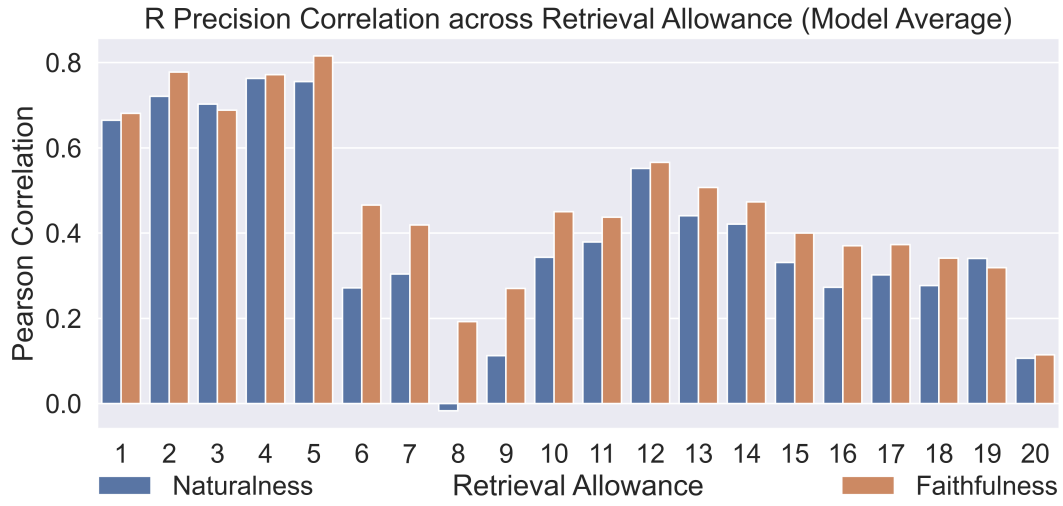

Figure 5.6: Model level R-Precision correlations with human judgments. Retrieval Allowance indicates the number of top samples (out of a batch size of 32) considered successful if the true match is found.

Match Dist Text-to-Text Scoring Function							
Encoder	Accum.	Naturalness			Faithfulness		
		Mean ρ	Std. ρ		Mean ρ	Std. ρ	
Std.	Min	-0.34	0.43	↑	-0.43	0.28	↑
	Max	-0.18	0.31	↑	-0.19	0.19	↑
	Avg	-0.11	0.27	→	-0.19	0.15	↑
	Med	-0.07	0.27	↑	-0.18	0.14	↑
T2M Min	Min	0.14	0.17	↑	0.38	0.35	↓
	Max	0.33	0.11	↑	0.61	0.05	↑
	Avg	0.35	0.12	↑	0.68	0.08	↑
	Med	0.4	0.17	↓	0.76	0.13	↑
T2M Max	Min	0.12	0.1	↑	0.35	0.03	↑
	Max	-0.32	0.11	↑	0.06	0.16	↑
	Avg	-0.11	0.02	↑	0.25	0.07	↑
	Med	-0.09	0.09	↓	0.27	0.08	↓
T2M Avg	Min	0.18	0.06	↑	0.59	0.03	→
	Max	-0.06	0.09	↑	0.43	0.07	↑
	Avg	0.06	0.02	↑	0.52	0.02	↑
	Med	0.09	0.02	→	0.54	0.02	↑
T2M Lst	Min	0.49	0.05	↑	0.85	0.02	↓
	Max	0.19	0.51	→	0.39	0.72	→
	Avg	0.51	0.04	→	**0.86**	**0.01**	→
	Med	0.51	0.05	↑	**0.86**	**0.01**	→

Table 5.3: Model level correlation scores of Nearest Neighbor Captioning metrics with human judgments when using Match Dist scoring. Mean correlation and standard deviation over all 32 neighbor counts tested. Up, right, and down arrows indicate best performance at high (21-30), intermediate (11-20), or low (1-10) neighbor counts. These arrows don't represent the magnitude of the difference, and several versions performed similarly across all neighbor counts. Bold metric values indicate high mean correlation (¿0.75) and low standard deviation (¡0.15) and are recommended for use.

Match CS Text-to-Text Scoring Function							
Encoder	Accum.	Naturalness			Faithfulness		
		Mean ρ	Std. ρ		Mean ρ	Std. ρ	
Std.	Min	**0.86**	**0.06**	↑	0.52	0.09	↑
	Max	0.71	0.05	↑	0.23	0.07	↑
	Avg	0.78	0.03	→	0.36	0.03	↑
	Med	0.75	0.06	↓	0.33	0.07	↑
T2M Min	Min	0.6	0.14	→	0.61	0.19	↑
	Max	0.78	0.11	↑	**0.86**	**0.06**	→
	Avg	0.8	0.08	→	**0.92**	**0.07**	↓
	Med	0.68	0.08	↑	0.84	0.09	→
T2M Max	Min	0.07	0.06	↑	-0.37	0.05	↑
	Max	0.03	0.08	→	-0.4	0.07	↑
	Avg	0.09	0.05	↑	-0.36	0.04	↑
	Med	0.13	0.05	↑	-0.34	0.05	↑
T2M Avg	Min	0.09	0.03	↑	-0.39	0.03	↑
	Max	0.15	0.02	→	-0.32	0.03	→
	Avg	0.12	0.02	→	-0.37	0.02	→
	Med	0.1	0.03	→	-0.38	0.03	↑
T2M Lst	Min	0.33	0.07	→	-0.17	0.07	→
	Max	0.31	0.04	→	-0.2	0.04	→
	Avg	0.35	0.02	↑	-0.16	0.02	↑
	Med	0.36	0.03	↑	-0.14	0.03	↑

Table 5.4: Model level correlation scores of Nearest Neighbor Captioning metrics with human judgments when using Match CS scoring. Mean correlation and standard deviation over all 32 neighbor counts tested. Up, right, and down arrows indicate best performance at high (21-30), intermediate (11-20), or low (1-10) neighbor counts. These arrows don't represent the magnitude of the difference, and several versions performed similarly across all neighbor counts. Bold metric values indicate high mean correlation (¿0.75) and low standard deviation (¡0.15) and are recommended for use.

CLIP Dist Text-to-Text Scoring Function							
Encoder	Accum.	Naturalness			Faithfulness		
		Mean ρ	Std. ρ		Mean ρ	Std. ρ	
Std.	Min	0.43	0.3	↑	0.6	0.39	↑
	Max	0.23	0.25	→	0.48	0.31	→
	Avg	0.28	0.32	→	0.57	0.38	→
	Med	0.23	0.46	→	0.49	0.48	→
T2M Min	Min	0.39	0.3	→	0.65	0.27	→
	Max	0.58	0.2	↑	0.24	0.23	↑
	Avg	0.66	0.2	↑	0.34	0.31	↑
	Med	0.45	0.29	→	0.22	0.36	→
T2M Max	Min	-0.54	0.37	↑	-0.48	0.11	→
	Max	-0.22	0.46	→	0.02	0.41	→
	Avg	-0.53	0.37	→	-0.32	0.4	→
	Med	-0.19	0.51	→	0.01	0.58	→
T2M Avg	Min	-0.46	0.29	↑	-0.25	0.34	↑
	Max	-0.46	0.28	↑	-0.25	0.26	↑
	Avg	-0.53	0.26	↑	-0.35	0.27	↑
	Med	-0.13	0.38	→	-0.25	0.26	→
T2M Lst	Min	-0.2	0.32	↑	-0.57	0.36	↑
	Max	-0.45	0.25	↑	-0.8	0.21	↑
	Avg	-0.5	0.1	↑	**-0.87**	**0.06**	→
	Med	-0.66	0.08	→	-0.85	0.07	→

Table 5.5: Model level correlation scores of Nearest Neighbor Captioning metrics with human judgments when using CLIP Dist scoring. Mean correlation and standard deviation over all 32 neighbor counts tested. Up, right, and down arrows indicate best performance at high (21-30), intermediate (11-20), or low (1-10) neighbor counts. These arrows don't represent the magnitude of the difference, and several versions performed similarly across all neighbor counts. Bold metric values indicate high mean correlation (¿0.75) and low standard deviation (¡0.15) and are recommended for use.

CLIP CS Text-to-Text Scoring Function							
Encoder	Accum.	Naturalness			Faithfulness		
		Mean ρ	Std. ρ		Mean ρ	Std. ρ	
Std.	Min	-0.26	0.25	$\rightarrow$	-0.53	0.29	$\rightarrow$
	Max	-0.44	0.32	$\rightarrow$	-0.51	0.36	$\rightarrow$
	Avg	-0.39	0.3	$\uparrow$	-0.64	0.33	$\rightarrow$
	Med	-0.39	0.44	$\rightarrow$	-0.59	0.43	$\rightarrow$
T2M Min	Min	-0.58	0.18	$\uparrow$	-0.4	0.27	$\rightarrow$
	Max	-0.21	0.27	$\uparrow$	-0.61	0.16	$\uparrow$
	Avg	-0.55	0.22	$\downarrow$	-0.3	0.3	$\rightarrow$
	Med	-0.16	0.41	$\rightarrow$	0.1	0.41	$\uparrow$
T2M Max	Min	-0.22	0.41	$\downarrow$	-0.27	0.48	$\rightarrow$
	Max	0.16	0.29	$\uparrow$	0.2	0.23	$\rightarrow$
	Avg	0.35	0.38	$\rightarrow$	0.16	0.36	$\rightarrow$
	Med	-0.07	0.4	$\uparrow$	-0.16	0.42	$\uparrow$
T2M Avg	Min	0.5	0.31	$\uparrow$	0.33	0.27	$\uparrow$
	Max	0.44	0.33	$\uparrow$	0.22	0.36	$\uparrow$
	Avg	0.42	0.3	$\uparrow$	0.28	0.28	$\uparrow$
	Med	0.08	0.4	$\rightarrow$	0.11	0.29	$\uparrow$
T2M Lst	Min	-0.37	0.39	$\uparrow$	-0.38	0.61	$\uparrow$
	Max	0.12	0.34	$\uparrow$	0.49	0.36	$\uparrow$
	Avg	-0.09	0.36	$\uparrow$	0.26	0.4	$\uparrow$
	Med	0.05	0.58	$\rightarrow$	0.19	0.75	$\rightarrow$

Table 5.6: Model level correlation scores of Nearest Neighbor Captioning metrics with human judgments when using CLIP CS scoring. Mean correlation and standard deviation over all 32 neighbor counts tested. Up, right, and down arrows indicate best performance at high (21-30), intermediate (11-20), or low (1-10) neighbor counts. These arrows don't represent the magnitude of the difference, and several versions performed similarly across all neighbor counts. Bold metric values indicate high mean correlation (¿0.75) and low standard deviation (¡0.15) and are recommended for use.

BERTScore R Text-to-Text Scoring Function							
Encoder	Accum.	Naturalness			Faithfulness		
		Mean ρ	Std. ρ		Mean ρ	Std. ρ	
Std.	Min	0.14	0.4	$\downarrow$	0.09	0.4	$\downarrow$
	Max	-0.11	0.4	$\uparrow$	-0.46	0.43	$\uparrow$
	Avg	0.26	0.33	$\rightarrow$	0.26	0.45	$\rightarrow$
	Med	0.18	0.3	$\rightarrow$	0.3	0.36	$\rightarrow$
T2M Min	Min	-0.39	0.33	$\uparrow$	-0.32	0.35	$\downarrow$
	Max	-0.25	0.46	$\uparrow$	-0.21	0.49	$\uparrow$
	Avg	0.14	0.39	$\rightarrow$	0.04	0.28	$\rightarrow$
	Med	-0.01	0.33	$\uparrow$	-0.21	0.27	$\uparrow$
T2M Max	Min	0.25	0.57	$\uparrow$	0.25	0.46	$\rightarrow$
	Max	0.13	0.42	$\uparrow$	0.27	0.21	$\uparrow$
	Avg	0.22	0.56	$\downarrow$	0.22	0.46	$\downarrow$
	Med	0.19	0.48	$\downarrow$	0.26	0.45	$\downarrow$
T2M Avg	Min	-0.56	0.4	$\rightarrow$	-0.63	0.35	$\rightarrow$
	Max	-0.19	0.21	$\uparrow$	0.05	0.23	$\uparrow$
	Avg	-0.57	0.16	$\uparrow$	-0.36	0.23	$\uparrow$
	Med	-0.64	0.17	$\uparrow$	-0.41	0.22	$\uparrow$
T2M Lst	Min	-0.34	0.14	$\uparrow$	-0.09	0.15	$\uparrow$
	Max	-0.36	0.28	$\uparrow$	-0.28	0.45	$\uparrow$
	Avg	-0.54	0.18	$\rightarrow$	-0.53	0.26	$\uparrow$
	Med	-0.48	0.23	$\uparrow$	-0.6	0.28	$\rightarrow$

Table 5.7: Model level correlation scores of Nearest Neighbor Captioning metrics with human judgments when using BERTScore R scoring. Mean correlation and standard deviation over all 32 neighbor counts tested. Up, right, and down arrows indicate best performance at high (21-30), intermediate (11-20), or low (1-10) neighbor counts. These arrows don't represent the magnitude of the difference, and several versions performed similarly across all neighbor counts. Bold metric values indicate high mean correlation (¿0.75) and low standard deviation (¡0.15) and are recommended for use.

BLEURT Text-to-Text Scoring Function							
Encoder	Accum.	Naturalness			Faithfulness		
		Mean ρ	Std. ρ		Mean ρ	Std. ρ	
Std.	Min	0.29	0.35	$\rightarrow$	0.5	0.33	$\rightarrow$
	Avg	0.17	0.43	$\rightarrow$	0.24	0.36	$\rightarrow$
	Med	0.08	0.52	$\downarrow$	0.05	0.5	$\downarrow$
T2M Min	Min	-0.64	0.25	$\uparrow$	-0.33	0.31	$\rightarrow$
	Max	0.55	0.25	$\rightarrow$	0.82	0.19	$\rightarrow$
	Avg	0.21	0.2	$\rightarrow$	0.6	0.16	$\rightarrow$
	Med	0.39	0.14	$\uparrow$	0.64	0.32	$\rightarrow$
T2M Max	Min	-0.35	0.35	$\uparrow$	-0.23	0.41	$\uparrow$
	Max	0.5	0.43	$\uparrow$	0.28	0.4	$\uparrow$
	Avg	-0.04	0.34	$\uparrow$	-0.08	0.39	$\uparrow$
	Med	-0.08	0.46	$\rightarrow$	-0.06	0.51	$\downarrow$
T2M Avg	Min	0.03	0.26	$\uparrow$	-0.07	0.32	$\uparrow$
	Max	0.29	0.43	$\downarrow$	0.31	0.43	$\uparrow$
	Avg	0.13	0.38	$\uparrow$	-0.12	0.54	$\uparrow$
	Med	0.01	0.46	$\uparrow$	-0.27	0.58	$\uparrow$
T2M Lst	Min	0.33	0.47	$\downarrow$	0.23	0.51	$\rightarrow$
	Max	-0.23	0.37	$\uparrow$	0.05	0.45	$\uparrow$
	Avg	0.5	0.21	$\rightarrow$	0.82	0.15	$\rightarrow$
	Med	0.6	0.21	$\rightarrow$	0.8	0.22	$\rightarrow$

Table 5.8: Model level correlation scores of Nearest Neighbor Captioning metrics with human judgments when using BLEURT scoring. Mean correlation and standard deviation over all 32 neighbor counts tested. Up, right, and down arrows indicate best performance at high (21-30), intermediate (11-20), or low (1-10) neighbor counts. These arrows don't represent the magnitude of the difference, and several versions performed similarly across all neighbor counts. Bold metric values indicate high mean correlation (¿0.75) and low standard deviation (¡0.15) and are recommended for use.

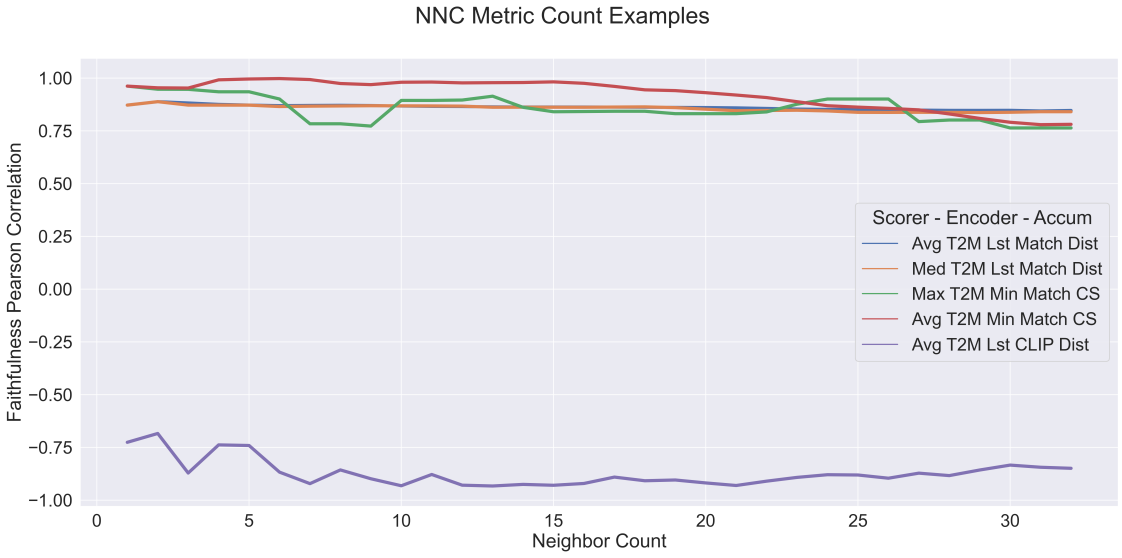

Figure 5.7: Pearson Correlations with human judgments of *Faithfulness* for our top performing NNC metrics.

Chapter 6: Discussion and Future Work

Our study provides valuable insights into the current state of automated evaluation in the text-to-motion domain. We discovered a strong correlation between R-Precision scores and human judgments, recommending its continued use with an increased Retrieval Threshold of 5, as our data suggests this as the optimal setting. FID demonstrated acceptable but inferior correlations with *Faithfulness* compared to R-Precision and poor correlation with *Naturalness*. We advise cautiously using FID, considering its performance may improve with larger sample sizes, but not prioritizing its results over better-performing alternatives. We discourage using Multimodal Distance due to its weak model-level correlations and near-zero sample-level correlations.

Our findings reveal that newer metrics result in a decline in evaluation quality compared to traditional CE-based metrics. CE-based metrics exhibited strong model-level performance and were the only metrics to achieve even weak sample-level correlations. Our novel versions with tuned root and component scaling achieved near-perfect correlations at the model level. While our samples' size limits the generalizability of these correlations, they provide compelling evidence supporting the continued use and study of CE-based metrics.

Our introduced NCC metrics show potential but exhibit noisy performance depending on the parameters. Some demonstrated stable high correlation values with room for improvement through enhanced motion encoders, larger reference datasets, or advanced text-scoring methods. In contrast, FID, R-Precision, and Multimodal Distance would benefit only from improved motion encoders. We suggest future research prioritize the development of superior encoders for both motion generation and evaluation.

Regarding sample-level correlations, new metrics are needed to reliably assess

individual samples with better than weak correlations. Current metrics fall short, with CE metrics outperforming motion encoder-based metrics despite their theoretical shortcomings due to reliance on a single reference.

We recommend using R-Precision 1-5, FID, and both Pose POS AVE and Pose ACEL AE when evaluating text-to-motion generation. Optimal root scalings for the latter two can be determined from Figure 5.3. Better results can be achieved by combining derivative components, but we do not currently suggest doing so due to diminishing returns. We propose reporting NNC metrics but refraining from drawing conclusions based on them for now. Finally, despite these findings, no suitable alternative to human evaluation currently exists, and text-to-motion evaluations should always include human studies when possible.

6.1 Towards an Ideal Metric

None of the studied metrics are ideal. Ignoring low sample-level correlations, each metric has unique theoretical limitations. FID cannot measure sample-level correlations, imposes strict assumptions on generated motion embedding spaces, and disregards the prompt. In contrast, CE metrics rely on a single reference, failing to represent the task's one-to-many nature.

NNC metrics come closest to an ideal metric since they can be computed at the sample level, do not require fitting embedding spaces to normal distributions, and can capture the one-to-many nature with large reference datasets. However, they currently underperform in sample-level correlation and exhibit stability issues in many versions.

Future metrics should:

- Evaluate both sample and model-level performance.
- Not depend on a single reference sample, addressing the task's one-to-many nature.

- Avoid using encoders trained solely on highly natural motions. Evaluation encoders should excel at encoding both reference and generated motions.

6.2 Loss Functions and Human Judgement Optimization

Certain generative models for human motions employ loss functions that resemble the CE metrics we evaluated, such as Pose POS AE. The moderate correlation with human judgments observed in our study raises questions about the advisability of this approach. We propose that higher-scoring differentiable metrics, such as Root POS AE or Pose VEL AE, might be more effective. Even non-differentiable metrics, like NNC metrics, could be incorporated into the learning process through reinforcement learning. This method could optimize highly-correlated metrics and potentially lead to improved human-judged performance.

Chapter 7: Conclusions

In this study, we compiled a dataset of human motions generated by recent text-to-motion models, accompanied by human quality assessments. By analyzing existing and newly proposed evaluation metrics, we identified those that best correlate with human judgments. While R-Precision is a reliable metric for evaluating model quality, both traditional and our novel CE metrics perform equally well or even better, suggesting that R-Precision should not be the sole metric relied upon. Several other metrics that have replaced CE metrics demonstrated suboptimal or even poor performance. Our novel NCC metrics show significant potential and should be considered for reporting in future research. However, neither current nor new metrics achieve satisfactory correlation at the sample level. Efforts to enhance encoder quality or develop novel metrics to improve sample-level evaluations are encouraged.

7.1 Limitations

Our dataset, containing 1400 motion annotations, is relatively small compared to typical automated evaluations and covers only a small fraction of the HumanML3D test set. Although our study presents strong findings for model-level averages, it includes only five models, making model-level correlations potentially vulnerable to chance.

Human annotation introduces noise, and while we achieved a high inter-annotator agreement (IAA), this does not eliminate annotation noise. We used a single instruction for annotation, which could introduce bias among annotators; alternative instructions might yield different results.

As motion generation techniques continue to advance, the samples used in our study may not accurately represent error distributions in future improved models, potentially affecting the determination of the best metric. Despite the strong

correlations observed between metrics and human judgments, independent human evaluations remain crucial for comparing model performance.

Works Cited

Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, pages 719–728. IEEE, 2019.

Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, June 2020. IEEE.

Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan, 2017. URL `https://arxiv.org/abs/1711.09561`.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL `http://www.blender.org`.

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *ECCV (6)*, volume 13666 of *Lecture Notes in Computer Science*, pages 346–362, New York, NY, USA, 2022. Springer.

Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *ICCV*, pages 1376–1386, Wasington, DC, USA, 2021. IEEE.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL `https://arxiv.org/abs/1406.2661`.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d hu-

man motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2021–2029, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413635. URL https://doi.org/10.1145/3394171.3413635.

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 *CVPR*, pages 5142–5151, Washington, DC, USA, 2022a. IEEE.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV (35)*, volume 13695 of *Lecture Notes in Computer Science*, pages 580–597, New York, NY, USA, 2022b. Springer.

Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. MoGlow. *ACM Transactions on Graphics*, 39(6):1–14, nov 2020. doi: 10.1145/3414685.3417836. URL https://doi.org/10.1145%2F3414685.3417836.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, Red Hook, NY, USA, 2017. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars, 2022. URL https://arxiv.org/abs/2205.08535.

Leslie Ikemoto, Okan Arikan, and David Forsyth. Generalizing motion edits with gaussian processes. *ACM Trans. Graph.*, 28(1), feb 2009. ISSN

0730-0301. doi: 10.1145/1477926.1477927. URL https://doi.org/10.1145/1477926.1477927.

Jihoon Kim, Jiseob Kim, and Sungjoon Choi. FLAME: free-form language-based motion synthesis & editing, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 11–21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380171. doi: 10.1145/3397481.3450692. URL https://doi.org/10.1145/3397481.3450692.

Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer, 2020.

Angela S. Lin, Lemeng Wu, Rodolfo Corona, Kevin W. H. Tai, Qixing Huang, and Raymond J. Mooney. Generating animated videos of human activities from natural language descriptions. In *Proceedings of the Visually Grounded Interaction and Language Workshop at NeurIPS*, Red Hook, NY, USA, 2018. Curran Associates, Inc.

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion VAEs. *ACM Transactions on Graphics*, 39(4), aug 2020. doi: 10.1145/3386569.3392422. URL https://doi.org/10.1145%2F3386569.3392422.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), nov 2015a. ISSN 0730-0301. doi: 10.1145/2816795.2818013. URL https://doi.org/10.1145/2816795.2818013.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015b.

Tomohiko Mukai and Shigeru Kuriyama. Geostatistical motion interpolation. *ACM Trans. Graph.*, 24:1062–1070, 2005.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae, 2021. URL https://arxiv.org/abs/2104.05670.

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: generating diverse human motions from textual descriptions. In *ECCV (22)*, volume 13682 of *Lecture Notes in Computer Science*, pages 480–497, New York, NY, USA, 2022. Springer.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: bodies, action and behavior with english labels. In *CVPR*, pages 722–731, Washington, DC, USA, 2021. Computer Vision Foundation / IEEE.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation, 2021.

Philip Sedgwick. Pearson's correlation coefficient. *BMJ*, 345, 2012. doi: 10.1136/bmj.e4483. URL https://www.bmj.com/content/345/bmj.e4483.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. Bleurt: Learning robust metrics for text generation, 2020. URL https://arxiv.org/abs/2004.04696.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. URL https://arxiv.org/abs/2209.14792.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022. URL https://arxiv.org/abs/2209.14916.

Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. Edge: Editable dance generation from music, 2022.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation, 2017. URL https://arxiv.org/abs/1707.04993.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 736–747, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393904. doi: 10.1145/3536221.3558058. URL https://doi.org/10.1145/3536221.3558058.

Kim Youwang, Ji-Yeon Kim, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *ECCV (3)*, volume 13663 of *Lecture Notes in Computer Science*, pages 173–191, New York, NY, USA, 2022. Springer.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model, 2022.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2019. URL https://arxiv.org/abs/1904.09675.

Vita

Jordan Guy Voas was born in Saint Cloud, Minnesota in 1995. He has two brothers, including a twin. Jordan earned a double major Bachelor of Science degree in Computer Science and Computer Engineering from the University of Minnesota - Twin Cities. Upon graduation in 2018, he secured a position as a test software engineer at Intel Inc.

For three years, Jordan contributed to the development of Intel's NAND non-volatile memory solutions group at their Folsom, California campus. In 2021, he decided to further his education and enrolled in a Master of Science program at the University of Texas at Austin. Jordan eventually parted ways with Intel in 2022 to fully focus on his graduate studies.

He has accepted admission into the University of Texas at Austin's Computer Science Doctoral program for the Fall 2023 semester.

Address: jvoas@utexas.edu

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.