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

Jordan Voas
University of Texas at Austin
Austin, TX, USA
jvoas@utexas.edu

Yili Wang
University of Texas at Austin
Austin, TX, USA
ywang98@utexas.edu

Qixing Huang
University of Texas at Austin
Austin, TX, USA
huangqx@cs.utexas.edu

Raymond Mooney
University of Texas at Austin
Austin, TX, USA
mooney@cs.utexas.edu

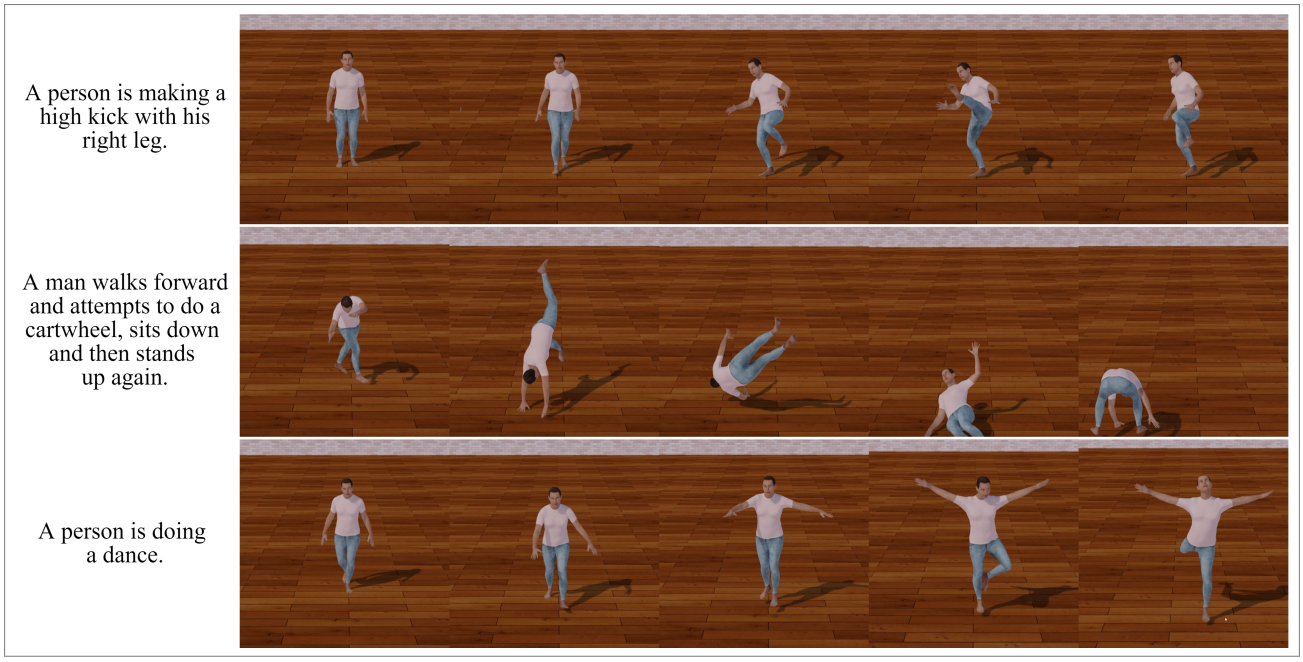

Figure 1: Sampled motion frames with paired descriptions, as used in our human evaluations. Our rendering framework generates pseudo-realistic environments with skin, wall, and floor textures as well as environment lighting and steady camera motions.

ABSTRACT

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 effective 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 less-used coordinate errors show strong correlations. Additionally, several recently developed metrics are not recommended due to their low correlation compared to alternatives. We also introduce a novel metric based on a multimodal BERT-like model, MoBERT, which offers strongly

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0315-7/23/12.
https://doi.org/10.1145/3610548.3618185

human-correlated sample-level evaluations while maintaining near-perfect model-level correlation. Our results demonstrate that this new metric exhibits extensive benefits over all current alternatives.

CCS CONCEPTS

• **Computing methodologies** → **Procedural animation**; **Motion capture**; **Natural language processing**; **Natural language generation**; **Temporal reasoning**; **Spatial and physical reasoning**; **Model verification and validation**; • **Human-centered computing** → **Visualization design and evaluation methods**.

KEYWORDS

Multi-modal, human evaluation

ACM Reference Format:

Jordan Voas, Yili Wang, Qixing Huang, and Raymond Mooney. 2023. What is the Best Automated Metric for Text to Motion Generation?. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3610548.3618185

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 [Ahuja and Morency 2019; Delmas et al. 2022; Ghosh et al. 2021; Guo et al. 2022b; Lin et al. 2018; Punnakkal et al. 2021; Zhang 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. 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.

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 while being differentiable and capable of enhancing optimization when integrated into training losses. One novel metric, a multimodal BERT-like model MoBERT, offers sample level evaluation scores with significantly improved human judgment correlations.

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 viewer? Unnatural motions exhibit inhuman or improbable poses 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.

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 novel high-performing automated metrics, including MoBERT, offering the first strongly human-correlated evaluation metric for this task. We also discuss how MoBERT addresses limitations of existing metrics, advancing future architecture comparison and development.

[1]

2 RELATED WORKS

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 [Aliakbarian et al. 2020; Ling et al. 2020a; Rempe et al. 2021; Tulyakov et al. 2017]. Others have used Generative Adversarial Networks (GAN) [Goodfellow et al. 2014] to enhance the quality of generations [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. 2015] body model, which represents the body as a kinematic tree of connected segments.

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. [Guo et al. 2020] employs a recurrent conditional VAE, while [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 specification of individual body parts. Early efforts utilized a Seq2Seq approach [Lin

[1]Our metric evaluation code and collected human judgment dataset are included as supplemental material to this work. Our novel evaluator model, MoBERT, is available at https://github.com/jvoas655/MoBERT.

et al. 2018]. Other studies learned a joint embedding 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 [Petrovich et al. 2022], use stochastic for diverse generations. Others employed diffusion-based models [Kim et al. 2022][Zhang et al. 2022][Tevet et al. 2022][Wei et al. 2023][Chen et al. 2022][Shafir et al. 2023][Zhang et al. 2023a][Han et al. 2023]. Recent models have taken inspiration from GPT-like LLM's through learned motion vocabularies and have competed with diffusion methods for SOTA performance [Zhang et al. 2023b][Jiang et al. 2023][Zhou and Wang 2022][Zhang et al. 2023c].

Related tasks have also been investigated, such as [Li et al. 2020] or [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 [Hong et al. 2022; Youwang et al. 2022]. Adjacent efforts have focused on scene or motion path-based conditioning, allowing for high-quality animation of character movements along specific paths in an environment [Holden et al. 2017][Ling et al. 2020b][Huang et al. 2023].

2.2 Metrics for Automated Evaluation of Human Motions

Various metrics have been used to evaluate text-to-motion. [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. [Petrovich et al. 2022] and [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 [Guo et al. 2022b; Kim et al. 2022; Tevet et al. 2022; Zhang et al. 2022]. These metrics rely on a text and motion co-encoder, so proving the effectiveness of the encoder is crucial for these metrics if they are to be used for judging model performance. [Yuan et al. 2022] expanded these metrics to measure factors of physical motion plausibility.

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 [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.

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 22 joint SMPL body models [Loper et al. 2015], enabling consistent animation methods for human ratings. We also include reference motions from HumanML3D as a baseline for non-reference evaluation metrics.

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]. The embeddings were projected onto a low-dimensional space and we randomly sample from the resulting dataset's distribution, taking the nearest unsampled entry, to obtain 400 unique sample prompts.

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 *Naturalness* using stick figures can be challenging, as they are not relatable to raters. 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 camera movements for natural motion perception (Figure 1).

3.4 Human Quality Ratings Collection

We collected human quality ratings using Amazon Mechanical Turk and a custom UI. To ensure quality, we implemented qualification requirements, 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[2]. 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.

Ratings were presented as natural language descriptions corresponding to Likert Scale ratings (0 to 4). Annotators had access to a tooltip with detailed descriptions for each rating level during the task, all shown in Figures 3, 4, and 5 of the supplement.

Ratings were rejected if more than two of the five test questions deviated by more than one from the "correct" answer. This allowed

[2]Gold test questions ground truth labels were judged by the Authors. Motions for which the ratings were deemed to be overly subjective were not included in the gold test set.

for subjectivity, missed details, and slight rating scale understanding differences. Significant deviations in rating scale understanding or guessing would pass a single question occasionally, but over the ten independent ratings would be detected with a high likelihood. In-tool quality checks required watching the entire video before progressing and capped the rate of progression to 12 seconds per sample. These measures aimed to prevent rushing and encourage thoughtfulness. Qualification requirements included residing in the U.S., completing over 1000 hits, and a minimum 98% acceptance rate. Quality checks were disclosed in the task instructions. We paid $1.25 per HIT, equating to at least $12 per hour.

We removed samples with less than three ratings for all models, resulting in 1400 rated motion-text pairs (280 distinct prompts for each baseline model). Averaging the three ratings provided final *Naturalness* and *Faithfulness* values. We show in Figure 6 the dataset's distribution to be generally normal, while Table 2 shows high inter-annotator agreement (Krippendorff's Alpha) was obtained.

4 EVALUATED METRICS

We evaluate most automated metrics from recent works as well as 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 individual 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 samples. 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.

To calculate FID, R-Precision, and Multimodal Distance the motion features must be projected into an embedding space using an encoder. The encoder used was developed by [Guo et al. 2022a] and is standard for these metrics.

4.1 Existing Metrics

4.1.1 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:

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

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

$$AVE = \frac{1}{J} \sum_{j \in J} \|\sigma[j] - \hat{\sigma}_t[j]\|_2 \tag{3}$$

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

We hypothesize 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 explore 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. Component-based scaling acts as a weighted average, with scaling factors increasing or decreasing the component errors, while root scaling adjusts the effects of root translation on all joint positions.

4.1.2 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:

$$FID(D_1, D_2) = |\mu_1 - \mu_2| + tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}}) \tag{4}$$

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.

4.1.3 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].

4.1.4 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 are

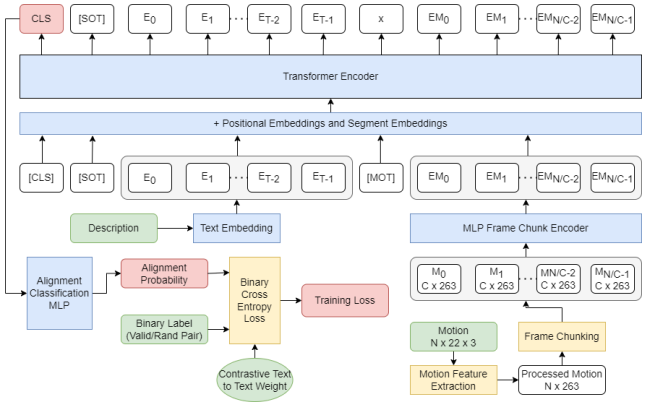

Figure 2: Our MoBERT architecture and process flow. Green items represent inputs, white items indicate intermediate steps, red items denote output/losses, and blue items contain learned model parameters.

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].

4.2 MoBERT: Multimodal Transformer Encoder Evaluator

Our novel evaluation method, MoBERT, is inspired by past learned metrics such as CLIPScore [Hessel et al. 2022], that score the alignment between a multimodal pair. However, MoBERT distinguishes itself by its ability to evaluate both modalities using a shared Transformer Encoder [Vaswani et al. 2017] through a multimodal sequence embedding. This approach, as shown in Figure 2, employs the attention mechanism of the Transformer to capture detailed relationships between the motion chunks and textual tokens.

Compared to CLIPScore, which uses separate encoders for each modality and combines the two modalities using cosine similarity, MoBERT's single Encoder approach allows for a richer understanding of the data. The Transformer Encoder's attention mechanism can learn to consider features across both modalities simultaneously, potentially capturing nuanced relationships between them that might be missed in a separate encoding scheme. In particular, this methodology allows MoBERT to consider the shared temporal aspects of motions and text prior to being collapsed to a single vector representation. This approach allows for more accurate prediction of correct and incorrect text pairings, allowing MoBERT to potentially outperform methods following CLIPScore's approach.

4.2.1 Encoding Motion Information. To better contextualize motion in our model, we preprocess our $N \times 22 \times 3$ motions into an $N \times 263$ representation following the approach in [Guo et al. 2022a]. This involves extracting motion transformations, such as root joint global transitions and rotations, to handle shifts in reference frames, as well as the linear velocities of each joint frame-to-frame and foot contact thresholding for a binary signal of foot-ground contact.

To utilize frame-to-frame motion information and mitigate redundancy in the motion domain, we downsample encodings by chunking consecutive frames into frame chunks before converting them into embeddings. Our dataset motions span up to 200 frames, processed at 20 Hz. We group these into 14-frame chunks, as 0.7 seconds of motion information offers adequate encoding and information differentiation. To account for the simplicity of our chunking algorithm, we apply an overlap factor of 4 frames, duplicating overlapped frames in consecutive motion chunks.

4.2.2 Multimodal Tokenization Process. For encoding text, we utilize a BPE [Gage 1994] vocabulary and learned embeddings. We generate sequence embeddings from the textual and motion processes and merge them into a single sequence (Figure 2). We incorporate special tokens for CLS, start indicators, and padding embeddings. With short one or two-sentence descriptions and motions limited to a chunk length of 20, we train using a max context size of 64. Learned segment and positional tokens are added to inputs.

4.2.3 Training Process. We used the HumanML3D dataset as the basis for our model's training. The model is trained through the task of **Alignment prediction** using Binary Cross Entropy loss. This task involves predicting a binary label that indicates whether a given motion corresponds to a specific textual description. For each motion-text pair in our training dataset, we randomly selected a contrastive textual description to serve as a negative label example. We then evaluate both valid and contrastive pairings with the model, resulting in alignment probability judgments. We employed a compact MLP model over the CLS output embeddings, terminating with sigmoid activation, to obtain an output alignment probability. Binary Cross Entropy loss is used to encourage the model to predict alignment labels for valid pairings and anti-alignment labels for incorrect pairings, as demonstrated in Equations 5 and 6.

$$H(q) = -\frac{1}{N}\sum_{i=1}^{N} y_q(i) \cdot log(p(y_q(i), q)) + (1 - y_q(i)) \cdot log(1 - p(y_q(i), q)) \tag{5}$$

$$\mathscr{L}1 = H(V) + H(R) \tag{6}$$

With N being all motions in a batch, $y_q(i)$ is the correct binary label for sample i given text grouping q (valid or contrastive), and p being the predicted binary label. V is the set of valid textual descriptions and R is the set of random contrastive descriptions. We found that this process could still present a difficult optimization landscape, and would often choose to predict all one label to minimize loss on one pairing despite increased losses on the other. To promote balancing each label's prediction, we achieved better results with the L2 balanced loss shown in Equation 7.

$$\mathscr{L}_2 = \sqrt{H(V)^2 + H(R)^2} \tag{7}$$

Additional tasks, in a multi-task learning framework, were trialed but did not improve performance and were not included in the version of MoBERT's we report in this work.

Improving Contrastive Examples. The HumanML3D dataset provides low diversity of descriptions, with many being very similar. Further, motions can be described in multiple ways, both of which make random contrastive textual samples provide low-quality guidance. To address this, we used Sentence Transformer similarity scores to weight contrastive training examples and adjust our loss

functions accordingly. Inverse similarity scores were applied as weights to the loss function, down-weighting similar descriptions to reduce label confusion. We employed the top-performing Huggingface "all-mpnet-base-v2" implementation. The contrastive loss was rescaled by the weights to maintain a consistent magnitude with the valid loss. The final loss function is shown in Equation 8, where alpha represents the similarity scores produced by the Sentence Transformer model score, confined to $[0, 1]$.

$$\mathscr{L}_f = \sqrt{H(V)^2 + \left(\frac{(1-\alpha)H(R)}{\sum_i^N (1-\alpha_i)}\right)^2} \tag{8}$$

Model Evaluation Process. We assess the correlation of our baseline models' raw Alignment Probability scores from our training process. Since this data lacks human rating guidance, we also test our model's performance when trained on a small set of human judgment data. We do this by discarding the output layers of our model, using an aggregation of output embeddings, and fitting a lightweight SVR or Linear Regression layer to predict human judgments. The best performance is achieved using a RBF Kernel SVR, with a Ridge regressor being the best fully differentiable. Sklearn's Python package is used for regression training and hyperparameters are reported in the supplemental materials section.

To avoid overfitting to the small human judgment dataset, we apply ten-fold cross-validation, fitting regressors on 90% of the dataset's samples to predict the remaining portion. These cross-validated predictions are collected, reordered, and Pearson's correlation is calculated against the human judgment ratings.

5 RESULTS ANALYSIS

This section highlights the key findings from our evaluation. We employed Pearson's Correlation Coefficient [Sedgwick 2012] to correlate metrics with human judgments, measuring the linear relationship 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 3.

All values are uncorrected, and negative correlations are expected for certain metrics (FID or CE) since our human judgment ratings suggest better outcomes with opposing directions. Weak P-values are observed for many reported correlations, which is anticipated as they were calculated (for model level results) based on only five samples. Our strongly-performing metrics achieved P-Values near 0.05 at the model level, while our best-performing sample-level metrics (Pearson's of 0.2 or above) had near zero P-Values.

5.1 Coordinate Error Metrics Results

The primary CE-metric results are presented in Table 1 with further details in Figures 8 and 9. Despite relying on only a single reference, CE metrics show weak but significant 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 does not surpass Joint metrics, and our derivative-based methods do worse than 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).

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 similar 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/or acceleration for both AVE and AE yields versions with greater than 0.99 Pearson's (Figure 9).

5.1.1 Root Scaling Exploration. We provide visualizations with scaling factors in Figures 10 and 11 to investigate the effects of root scaling on Pose CE metrics. Consistent with previous observations, model-level correlations improve (i.e., become more negatively correlated) when additional weight is placed on Root transitions. PV and PVA AE are the only versions that do not exhibit this trend. Alternatively, overemphasizing Root transitions significantly degrades performance at the sample level.

5.2 FID, R-Precision, and Multimodal Distance Results

Results for FID, R-Precision, and Multimodal Distance are also shown in Table 1, with additional detail for R-Precision across various Retrieval Thresholds in Figure 7. We examine FID only at the model level as it requires distributional statistics over multiple samples, preventing sample-level calculation. We present results for R-Precision at the sample level, but R-Precision provides only binary values at this level and so it is poorly suited for sample-level comparisons with Likert ratings unless averaged over multiple samples. Multimodal Distance scored near zero at the sample level so none of these metrics provide sample-level alternatives to CE metrics.

Regarding model-level results, FID achieves acceptable results for *Faithfulness* with 0.71 Pearson's but significantly underperforms for *Naturalness*. Given the weak correlation with *Naturalness* and model-level-only comparison, P-Values are notably weak. While these results are poor, it is possible our samples may provide an unfavorable setting for FID, or may improve with more samples.

R-Precision demonstrates substantial correlations for both human quality judgments, approaching 0.8 Pearson's with standard settings. Our results suggest current Retrieval Thresholds are suboptimally set, with thresholds of 4 and 5 being marginally better, and then declining at higher values. 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 is only weakly correlated with 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

Metric		Model Level		Sample Level	
		Faithfulness	Naturalness	Faithfulness	Naturalness
Root AVE	↓	-0.926	-0.908	-0.013	0.007
Root AE	↓	-0.715	-0.743	-0.033	0.037
Joint AVE	↓	-0.260	-0.344	-0.178	-0.185
Joint AE	↓	-0.120	-0.227	-0.208	-0.245
Multimodal Distance	↓	-0.212	-0.299	0.025	0.014
R-Precision	↑	0.816	0.756	0.036	0.042
FID	↓	-0.714	-0.269	-	-
MoBERT Score (Alignment Probability)	↑	**0.991**	0.841	0.488	0.324
MoBERT Score (SVR Regression)*	↑	0.962	**0.986**	**0.624**	**0.528**
MoBERT Score (Linear Regression)*	↑	0.951	0.975	0.608	0.515

Table 1: Pearson correlations with human judgments calculated for several existing metrics and our MoBERT model. The best-performing metric in each category is bolded. Models with (*) were judged through 10-fold cross-validation. R-Precision scores reported used the best settings identified (2 for sample level, 5 for model level). Arrows next to metrics indicate whether negative (↓) or positive (↑) correlation is expected.

possibility, they would require substantial enhancements to match even traditional CE metrics such as Root POS AVE.

5.3 MoBERT

Results for our novel learned metric are shown in Table 1, highlighting its performance against the best alternative metrics at the sample and model level. We observe that MoBERT substantially outperforms the best alternatives at both levels. The alignment probability outputs, without human judgment supervision, achieve a sample-level correlation of 0.488 for *Faithfulness*, up from a previous best of 0.208. As expected, the correlation with *Naturalness* is significantly weaker but still surpasses all other sample level correlations demonstrated by the baselines. Similarly strong results are observed for model-level performance.

Using a learned regression model over the output features further improves the results, highlighting the benefits of training on a small amount (approximately 1260 samples) of human-judgment. Our sample level correlations for *Faithfulness* and *Naturalness* increase to 0.624 and 0.528, respectively, reaching the strongly-correlated range for *Faithfulness* when using the SVR regression layer. Moreover, our model achieves near-perfect model-level correlations, verifying that its ability to signify improved model performance is highly reliable. We run additional experiments exploring MoBERTs ability to act as a text-free Naturalness evaluator in the supplemental materials.

5.4 Discussion and Future Work

Our findings underscore CE metrics as the most reliable baseline metric, demonstrating strong model-level performance supported by sample-level results. With the application of root/component scaling, CE metrics reached near-perfect model-level correlations, highlighting their significance when compared with newer metrics that showed weaker performance in our study.

Although R-Precision and FID demonstrate some correlation with human judgments, their relative significance should be evaluated in context. R-Precision reveals a solid correlation, yet fall short compared with CE metrics and should be considered supplemental. FID, while showing acceptable correlations with *Faithfulness* and some correlation with *Naturalness*, should be used with caution in consideration of its potential to improve with more samples, but not prioritized over more consistent metrics. We recommend against the use of Multimodal Distance due to its consistently weak correlations.

MoBERT significantly outperforms all competitors, presenting the first metric with robust model-level and sample-level performance. This metric also avoids reliance on any reference motions for evaluation, making it usable in more situations and alleviating concerns about the one-to-many nature of this task. Additionally, it is fully differentiable and could be used as a training objective for generative models in order to further enhance performance.

We recommend future evaluations employ our MoBERT evaluator alongside metrics such as R-Precision 1-5, FID, Pose POS AVE, and Root PV AE when assessing text-to-motion generation. Figure 11 can help determine optimal root scalings for Pose POS AVE.

5.4.1 MoBERT Out-of-Distribution (OOD) Robustness. MoBERT was pretrained exclusively on the HumanML3D dataset. Even though the regression versions are trained to fit human judgments using moderately OOD data produced by various generative models, these models were trained to emulate the HumanML3D data. The human judgment fine-tuning potentially learns to harness the most reliable MoBERT output features. These features, inferred from the distinct distributions produced by motion generation models, suggest a potential for MoBERT to withstand OOD scenarios. However, without a substantially OOD dataset, aligned to the 22-joint SMPL body model of HumanML3D, and coupled with human judgments this remains speculative. Low diversity in our training also may result in our vocabulary not being well covered for infrequent tokens.

To enhance MoBERT's adaptability, future efforts could retrain the regression versions with a growing, diverse dataset of human judgments as they are collected. This could enable MoBERT to better accommodate various motion types, textual inputs, or evolving concepts of **Naturalness** and **Faithfulness**. Nonetheless, when adapting MoBERT to OOD data, assessing its performance against relevant human judgments is recommended.

6 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. R-Precision is a reliable metric for evaluating model quality, but traditional CE metrics and our novel versions with root and component scaling perform equally well or even better, suggesting that R-Precision should not be relied upon as the sole metric. Some newer metrics that have replaced CE metrics in some publications demonstrated suboptimal or even poor performance. Our novel proposed MoBERT evaluator significantly outperforms all competitors, offering a reliable metric at all levels while being reference free. However, efforts to enhance encoder quality or develop novel metrics to improve sample-level evaluations are further encouraged as well as continued human studies whenever possible.

6.1 Limitations

Our dataset with 1400 motion annotations is fairly small for automated evaluation 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. Our interannotator agreement is high, but all human annotation has the potential to introduce biases and noise. We used a single instruction for annotation and 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 some metrics and human judgments, independent human evaluations remain crucial for comparing model performance.

ACKNOWLEDGMENTS

This research was partially supported by NSF NRI Grant IIS-1925082 and NSF IIS-2047677 as well as funding from Wormpex AI Research.

REFERENCES

Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2Pose: Natural Language Grounded Pose Forecasting. In *3DV*. IEEE, 719–728.

Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. 2020. A Stochastic Conditioning Scheme for Diverse Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Washington, DC, USA, 8 pages.

Emad Barsoum, John Kender, and Zicheng Liu. 2017. HP-GAN: Probabilistic 3D human motion prediction via GAN. https://doi.org/10.48550/ARXIV.1711.09561

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. 2022. Executing your Commands via Motion Diffusion in Latent Space. *ArXiv* abs/2212.04048 (2022).

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

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. 2022. PoseScript: 3D Human Poses from Natural Language. In *ECCV (6) (Lecture Notes in Computer Science, Vol. 13666)*. Springer, New York, NY, USA, 346–362.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal archive* 12 (1994), 23–38. https://api.semanticscholar.org/CorpusID:59804030

Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. 2021. Synthesis of Compositional Animations from Textual Descriptions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 1376–1386. https://doi.org/10.1109/ICCV48922.2021.00143

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. https://doi.org/10.48550/ARXIV.1406.2661

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022a. Generating Diverse and Natural 3D Human Motions from Text. In *CVPR*. IEEE, Washington, DC, USA, 5142–5151.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022b. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *ECCV (35) (Lecture Notes in Computer Science, Vol. 13695)*. Springer, New York, NY, USA, 580–597.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2Motion: Conditioned Generation of 3D Human Motions. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 2021–2029. https://doi.org/10.1145/3394171.3413635

Bo Han, Hao-Lun Peng, Minjing Dong, Changming Xu, Yi Ren, Yixuan Shen, and Yuheng Li. 2023. AMD: Autoregressive Motion Diffusion. *ArXiv* abs/2305.09381 (2023).

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv:2104.08718 [cs.CV]

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Red Hook, NY, USA. https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-Functioned Neural Networks for Character Control. *ACM Trans. Graph.* 36, 4, Article 42 (jul 2017), 13 pages. https://doi.org/10.1145/3072959.3073663

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. https://doi.org/10.48550/ARXIV.2205.08535

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. 2023. Diffusion-based Generation, Optimization, and Planning in 3D Scenes. *arXiv preprint arXiv:2301.06015* (2023).

Leslie Ikemoto, Okan Arikan, and David Forsyth. 2009. Generalizing Motion Edits with Gaussian Processes. *ACM Trans. Graph.* 28, 1, Article 1 (feb 2009), 12 pages. https://doi.org/10.1145/1477926.1477927

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint arXiv:2306.14795* (2023).

Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2022. FLAME: Free-form Language-based Motion Synthesis & Editing.

Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. https://doi.org/10.48550/ARXIV.1312.6114

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEA Challenge 2020. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 11–21. https://doi.org/10.1145/3397481.3450692

Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. 2020. Learning to Generate Diverse Dance Motions with Transformer.

Angela S. Lin, Lemeng Wu, Rodolfo Corona, Kevin W. H. Tai, Qixing Huang, and Raymond J. Mooney. 2018. Generating Animated Videos of Human Activities from Natural Language Descriptions. In *Proceedings of the Visually Grounded Interaction and Language Workshop at NeurIPS*. 4 pages.

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

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. 2020b. Character Controllers Using Motion VAEs. *ACM Trans. Graph.* 39, 4 (2020).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph.* 34, 6, Article 248 (nov 2015), 16 pages. https://doi.org/10.1145/2816795.2818013

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.

org/10.3115/1073083.1073135
Mathis Petrovich, Michael J. Black, and Gül Varol. 2021. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. https://doi.org/10.48550/ARXIV.2104.05670
Mathis Petrovich, Michael J. Black, and Gül Varol. 2022. TEMOS: Generating Diverse Human Motions from Textual Descriptions. In *ECCV (22) (Lecture Notes in Computer Science, Vol. 13682)*. Springer, New York, NY, USA, 480–497.
Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. 2021. BABEL: Bodies, Action and Behavior With English Labels. In *CVPR*. Computer Vision Foundation / IEEE, Washington, DC, USA, 722–731.
Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. https://doi.org/10.48550/ARXIV.2204.06125
Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. 2021. HuMoR: 3D Human Motion Model for Robust Pose Estimation. arXiv:2105.04668 [cs.CV]
Philip Sedgwick. 2012. Pearson's correlation coefficient. *BMJ* 345 (2012). https://doi.org/10.1136/bmj.e4483 arXiv:https://www.bmj.com/content/345/bmj.e4483.full.pdf
Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. 2023. Human Motion Diffusion as a Generative Prior. *ArXiv* abs/2303.01418 (2023).
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. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. https://doi.org/10.48550/ARXIV.2209.14792
Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human Motion Diffusion Model. https://doi.org/10.48550/ARXIV.2209.14916
Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. 2022. EDGE: Editable Dance Generation From Music. arXiv:2211.10658 [cs.SD]
Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2017. MoCoGAN: Decomposing Motion and Content for Video Generation. https://doi.org/10.48550/ARXIV.1707.04993
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762
Dong Wei, Xiaoning Sun, Huaijiang Sun, Bin Li, Sheng liang Hu, Weiqing Li, and Jian-Zhou Lu. 2023. Understanding Text-driven Motion Synthesis with Keyframe Collaboration via Diffusion Models. *ArXiv* abs/2305.13773 (2023).
Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A Large Evaluation of Data-Driven Co-Speech Gesture Generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 736–747. https://doi.org/10.1145/3536221.3558058
Kim Youwang, Ji-Yeon Kim, and Tae-Hyun Oh. 2022. CLIP-Actor: Text-Driven Recommendation and Stylization for Animating Human Meshes. In *ECCV (3) (Lecture Notes in Computer Science, Vol. 13663)*. Springer, New York, NY, USA, 173–191.
Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2022. PhysDiff: Physics-Guided Human Motion Diffusion Model. *ArXiv* abs/2212.02500 (2022).
Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xiaodong Shen. 2023c. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. *ArXiv* abs/2301.06052 (2023).
Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model.
Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. 2023a. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. *ArXiv* abs/2304.01116 (2023).
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. https://doi.org/10.48550/ARXIV.1904.09675
Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2023b. MotionGPT: Finetuned LLMs are General-Purpose Motion Generators. *arXiv preprint arXiv:2306.10900* (2023).
Zixiang Zhou and Baoyuan Wang. 2022. UDE: A Unified Driving Engine for Human Motion Generation. *ArXiv* abs/2211.16016 (2022).

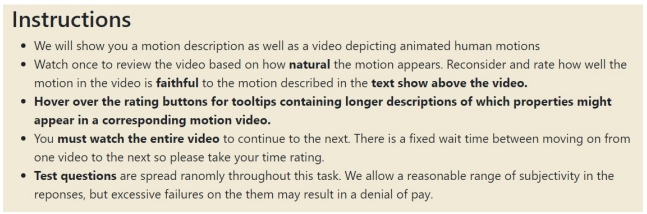

Figure 3: Instructions for raters in human judgment evaluations.

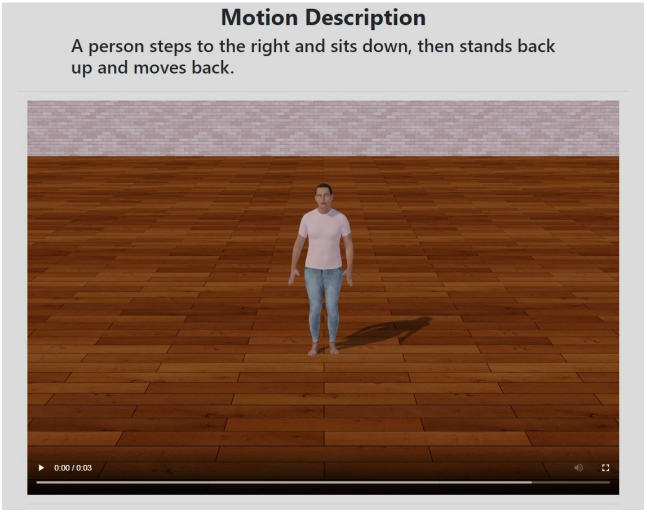

Figure 4: UI motion viewing section, situated below the instructions and above the rating selection.

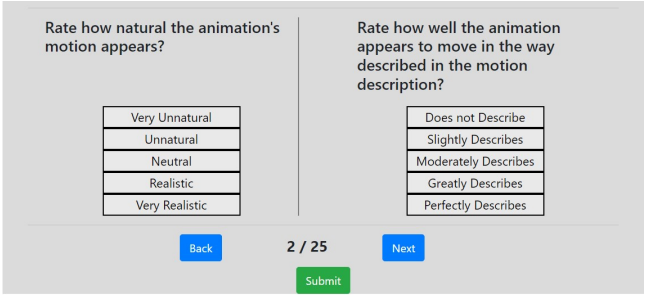

Figure 5: Rating selection UI, located below the motion viewing section. Detailed descriptions for each rating option were provided as tooltips upon hovering.

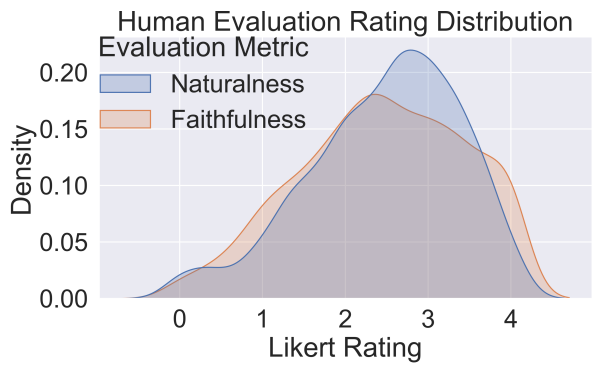

Figure 6: Human judgment distribution for all samples. Averages from three annotations are shown with a KDE smoothing filter (bandwidth 0.85) applied. Pearson's correlation between metrics is found to be 0.63 at the sample level.

IAA (Krippendorff's Alpha)	
Naturalness	Faithfulness
0.647	0.701

Table 2: Inter-annotator agreement across all replicated MTurk samples. Results indicate substantial but non-perfect agreement.

Pearson's Correlation	
Sample Level	Model Level
0.62	0.83

Table 3: Correlation between *Naturalness* and *Faithfulness*.

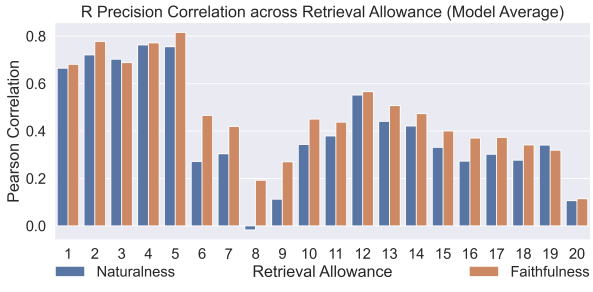

Figure 7: Model level R-Precision correlations with human judgments. Retrieval Allowance indicates the number of top samples (out of a batch size of 32) considered successful if the true match is found.

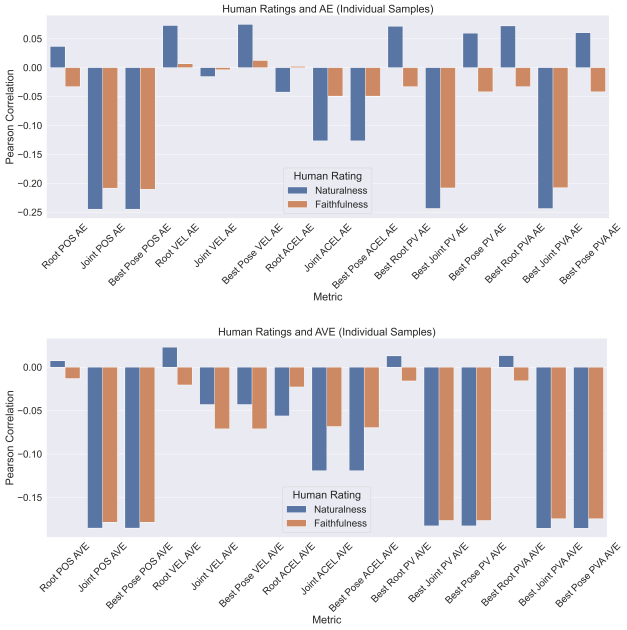

Figure 8: Sample level correlations of CE metrics with human judgments. "Best" denotes versions using highest performing settings for scaling root joints or components. Greater magnitude indicates better performance.

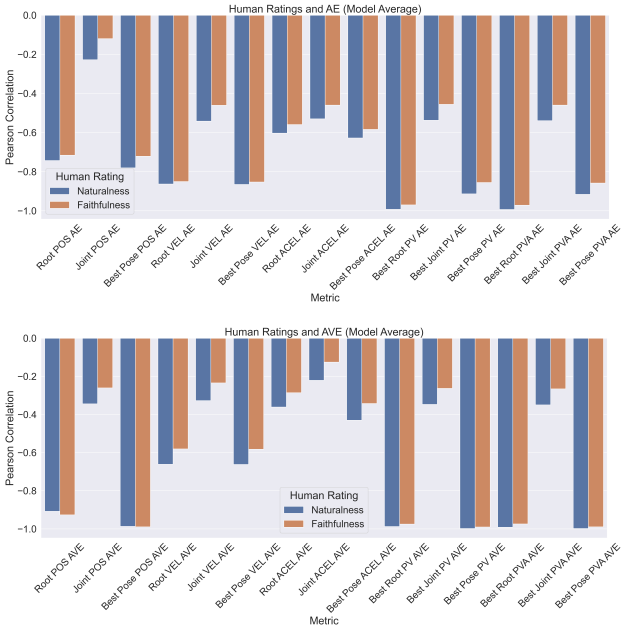

Figure 9: Model level correlations of CE metrics with human judgments. "Best" metrics use the highest performing settings for root joint or component scaling. Greater magnitude indicates better performance.

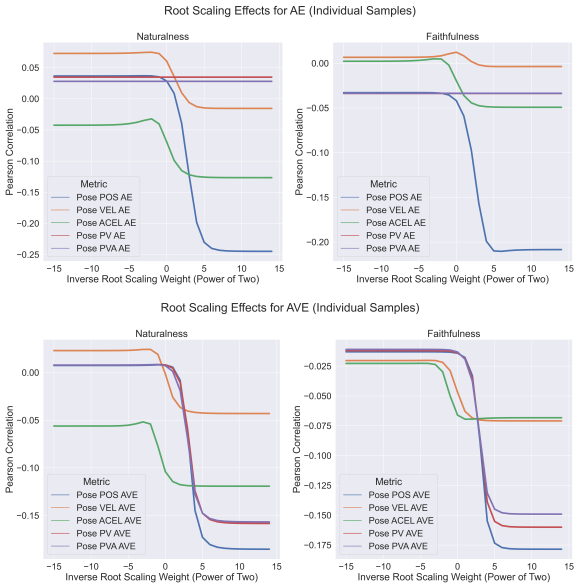

Figure 10: Sample level examination of root joint scaling effects on CE using all joints. Greater magnitude indicates better performance.

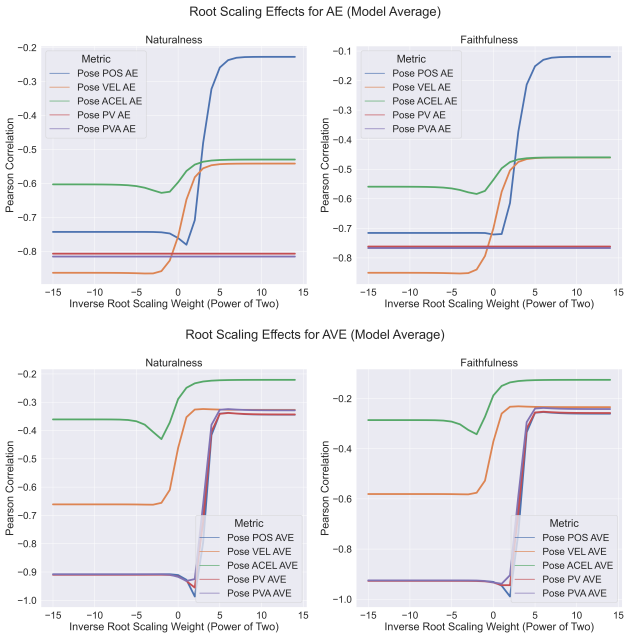

Figure 11: Model level examination of root joint scaling effects on CE metrics using all joints. Greater magnitude indicates better performance.