DLHM: Motion Generation

Hannes Möhring
hannes.moehring@tum.de

September 8, 2025

Contents

1 Abstract

The task of generating realistic 3D human motion from natural language descriptions was addressed by using the SMPL body model [4]. While recent work has demonstrated impressive progress, challenges remain in ensuring semantic alignment, motion diversity, and naturalness, particularly due to limitations in training data. In this project, we design a modular API framework that integrates multiple state-of-the-art models, including TEACH [1] and T2M-GPT [12], allowing the same textual description to yield diverse motion realizations. The analysis highlights the importance of dataset-specific language patterns, showing that adherence to these conventions improves the fidelity of generated motions. By combining differently trained models with prompt adjustments, this approach mitigates data scarcity and enhances output quality. The work demonstrates that leveraging complementary strengths across models leads to more robust text-to-motion generation and provides insights into the role of dataset structure in shaping model performance.

2 Introduction

2.1 Problem Statement

Recent years have seen rapid progress in text-to-motion generation, with applications ranging from medical simulations to video games and animation. Despite this progress, existing solutions remain limited by the scarcity of high-quality motion data and the challenges of reusing or augmenting existing recordings. Even subtle modifications to recorded trajectories can result in perceptually unnatural motion, making data expansion costly and error-prone. As a result, models trained on constrained datasets often struggle to produce diverse and naturalistic human motion.

2.2 Motivation

To address above mentioned limitations within the scope of this project, the complementary strengths of different state-of-the-art text-to-motion models were explored. The aim is to improve the average quality and robustness of the generated motions, by generating motions from the same textual description across multiple models and selecting the most convincing outcome. While automatic evaluation metrics could provide valuable guidance, final quality assessment ultimately depends on user judgment, reflecting the subjective nature of natural motion perception.

2.3 Project Goals

The objectives of this project are as follows:

- Develop a modular framework for generating motions from textual descriptions using multiple models.
- Leverage existing datasets and architectures to maximize motion quality despite limited data availability.
- Evaluate the generated motions using established quantitative metrics [3], as well as qualitative inspection.
- Analyze the data used in existing approaches.

2.4 Project Scope

This project focuses on integrating and comparing previously established approaches, specifically TEACH [1], TEMOS [6], and T2M-GPT [12] rather than designing entirely new architectures or datasets. By concentrating on the modular combination of existing models, we aim to mitigate the challenges posed by limited training data and provide insights into the strengths and weaknesses of current methods.

2.5 Report Structure

The remainder of this report is organized as follows: Section 3 reviews relevant background and related work. Section 4 presents the proposed methodology and implementation details. Section 5 describes the evaluation methodology and discusses experimental results. Finally, Section 6 concludes the report and outlines potential future work.

3 Background and Related Work

3.1 Single-Dataset Models

Earlier approaches, such as TEMOS [6], are trained on individual datasets, which ensures consistency in both motion recordings and their textual descriptions. This homogeneity simplifies learning and improves alignment between language and motion. However, the reliance on a single dataset substantially reduces data volume, limiting both the diversity of motions that can be generated and the overall realism of the outputs.

3.2 Multi-Dataset Models

3.2.1 Datasets

To overcome data scarcity, more recent models rely on larger resources such as KIT-ML [7], HumanML3D [3, 5], and, in the case of TEACH [1], the

BABEL [8, 5] dataset. These collections of data expand the range of described actions while introducing greater variability in annotation style and recording conditions.

3.2.2 Independent Motion Generation

Models such as T2M-GPT [12], MotionCLIP [11], and GUESS [2] leverage the large-scale HumanML3D dataset. They encode a single textual description and generate a corresponding motion sequence in its entirety. While this enables coverage of a broad range of actions, each motion is produced independently, without explicit temporal context or continuity across scenes.

3.2.3 Sequential Motion Generation

In contrast, TEACH introduces a sequential generation paradigm, where complex scenes are decomposed into multiple prompts, each associated with a duration [1]. Motions are generated segment by segment, with the final frames of one sequence serving as context for the next through a post conditioned autoencoder [1]. This design allows for smoother transitions and coherent multi-action sequences, at the cost of increased model complexity and reliance on well-structured prompts.

4 Methodology

4.1 System Architecture

The here presented framework adopts a modular design based on an API service that provides a unified interface for motion generation. This design enables the integration of multiple text-to-motion models while maintaining flexibility for user interaction and evaluation.

User Interface. Users interact with the system through a set of dedicated API endpoints. The `/upload_model` endpoint allows the submission of custom SMPL [4] models, returning a unique *model_ id* for subsequent reference. Motion generation is initiated via the `/generate` endpoint, which accepts a textual description and, optionally, a *model_ id*. Each generation request is assigned a *request_ id*, whose progress can be monitored through `/status`. Upon completion, the generated motion can be retrieved using `/download`.

Backend. The backend consists of two main components: (i) the *Model Handler*, which manages user submissions, stores uploaded models, and orchestrates motion generation, and (ii) the collection of text-to-motion models. With this implementation, two representative approaches were integrated: TEACH, which employs post-conditioning for sequential generation with smooth inter-segment transitions [1]; T2M-GPT, which generates motions directly from a single description using large-scale training on HumanML3D. The modular API design enables the user to leverage the complementary strengths of these models within a unified workflow.

4.2 Dataset Analysis

In addition to motion generation, an analysis of dataset characteristics was conducted to better understand the linguistic and structural properties of the available corpora. Each dataset provides pairs of motion sequences and textual descriptions, often accompanied by motion-specific tags.

Corpus Statistics. Furthermore, this metric reports descriptive statistics including the number of motions, number of captions, vocabulary size, and the mean and median token counts. These values highlight the diversity and scale of each dataset.

Motion-Specific Terms. To examine the precision of descriptions, the distribution of action-specific words as provided by dataset tagging is inspected, in order to assess how well textual labels align with underlying motions.

Actor References. Moreover it was evaluated, how human actors are addressed in the captions. Nouns and pronouns such as *woman*, *man*, or *someone* are categorized into female, male, or neutral references, which can influence model bias and motion conditioning.

Intra-Motion Similarity. Finally, the semantic consistency of captions associated with the same motion was measured. Each caption x_i is embedded into a vector $v_i \in \mathbb{R}^d$ using a pre-trained sentence encoder (e.g., MPNet [10] or distillbert-uncased [9]). After L2-normalization, cosine similarity reduces to the dot product:

$$\cos(v_i, v_j) = v_i^\top v_j.$$

For a motion with n captions and embeddings $V = \{v_1, \ldots, v_n\}$, the mean pairwise cosine similarity is given by

$$s_{\text{intra}}(V) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} v_i^\top v_j.$$

Values close to 1 indicate semantically consistent descriptions, whereas values near 0 or negative reflect greater variability. The distribution of s_{intra} across motions serves as an indicator of dataset coherence.

Listing 1: Core of calculating Intra Motion Similarity

```
def intramotion_similarity(self, df, model="minilm"):
        df = self.ensure_parsed(df.copy())
        Z, label = self.sentence_embeddings(df, model=
            model)
        # maping row idx -> vector
        idx_to_vec = {i: Z[i] for i in range(len(Z))}
```

```python

        rows = []
        for mid, sub in df.reset_index(drop=True).groupby("motion_id"):
            idxs = sub.index.tolist()
            V = np.stack([idx_to_vec[i] for i in idxs], axis=0)
            mpc = self._mean_pairwise_cosine_from_normed(V)
            rows.append({"motion_id": mid, "mean_intra_caption_sim": mpc, "num_captions": len(idxs)})

        out = pd.DataFrame(rows).sort_values("mean_intra_caption_sim", ascending=False)
        summary = {
            "backend": label,
            "mean_of_means": float(out["mean_intra_caption_sim"].mean(skipna=True)),
            "median_of_means": float(out["mean_intra_caption_sim"].median(skipna=True)),
            "quantiles": out["mean_intra_caption_sim"].quantile([0.1, 0.25, 0.5, 0.75, 0.9]).to_dict(),
        }
        return out, summary
```

5 Evaluation and Results

5.1 Models

Three representative text-to-motion models,TEMOS, TEACH, and T2M-GPT were evaluated, by retraining them on standard datasets and comparing the outcomes to reported baselines. In case of T2M-GPT, performance is assessed using the established metrics of Guo et al. [3], including R-Precision (Top-1, Top-2, Top-3), Fréchet Inception Distance (FID), multimodality distance (MM-Dist), and Diversity.

5.1.1 TEMOS

TEMOS was trained on the KIT-ML dataset for 1000 epochs, requiring approximately 12 hours on an NVIDIA TITAN RTX GPU. The reproduced model achieved results closely matching the original publication [6], with negligible deviations across all evaluation metrics. This confirms the reliability of the training pipeline and the reproducibility of prior results.

5.1.2 TEACH

To enable scene-level generation with distinct motion segments, TEACH was trained on the BABEL dataset [8]. Training required roughly 80 GPU-hours on an NVIDIA RTX 3090. While no quantitative comparison is included here, the model qualitatively demonstrates its strength in producing temporally coherent multi-action sequences through its post-conditioning mechanism [1].

5.1.3 T2M-GPT

T2M-GPT was trained for single scene generation. Following the standard procedure, this first required pretraining a Variational Autoencoder (VAE). This VAE was trained for 300,000 epochs with batch size 256, reproducing the reported results on KIT-ML (Tab. 1).

Methods	**R-Precision ↑**			**FID ↓**	**MM-Dist ↓**	**Diversity ↑**
	Top-1	Top-2	Top-3			
Real motion	$0.424_{\pm.005}$	$0.649_{\pm.006}$	$0.779_{\pm.006}$	$0.031_{\pm.004}$	$2.788_{\pm.012}$	$11.08_{\pm.097}$
VQ-VAE (Recons.)	$0.399_{\pm.005}$	$0.614_{\pm.005}$	$0.740_{\pm.006}$	$0.472_{\pm.011}$	$2.986_{\pm.027}$	$10.99_{\pm.120}$
My VAE	0.424	0.615	0.750	0.470	2.960	11.51

Table 1: Evaluation metrics for real motion, baseline VQ-VAE [12], and our trained VAE on KIT-ML [7].

Subsequently, T2M-GPT was trained on HumanML3D for 300,000 epochs with batch size 128, requiring 50 GPU-hours on an NVIDIA RTX A5000. As shown in Tab. 2, the newly trained model reproduces the reported trends [12, 3], although minor deviations exist due to the smaller number of training runs compared to the original work.

Methods	**R-Precision ↑**			**FID ↓**	**MM-Dist ↓**	**Diversity ↑**
	Top-1	Top-2	Top-3			
T2M-GPT (baseline)	$0.417_{\pm.003}$	$0.589_{\pm.002}$	$0.685_{\pm.003}$	$0.140_{\pm.006}$	$3.730_{\pm.009}$	$9.844_{\pm.095}$
My T2M-GPT	0.527	0.701	0.793	0.21	3.042	9.616

Table 2: Evaluation metrics for baseline and our trained T2M-GPT on HumanML3D.

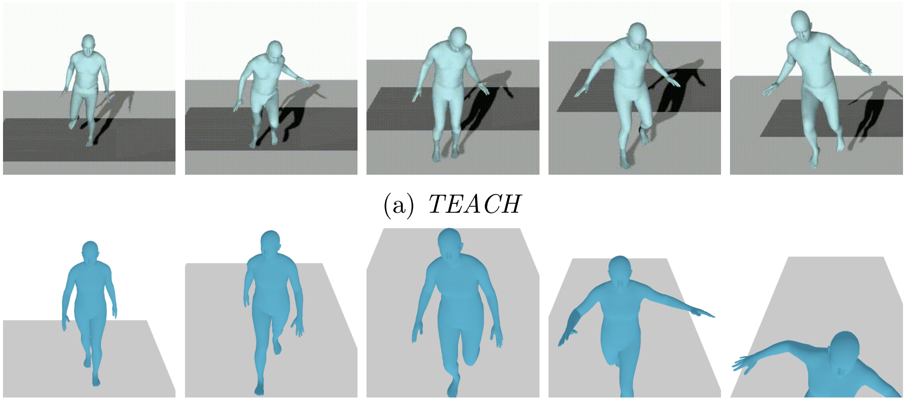
Finally, training on KIT-ML confirmed that while performance is slightly below the baseline due to fewer training runs, results remain in close proximity to reported values (Tab. 2). Notably, training on HumanML3D yields substantial gains over KIT-ML, underlining the importance of dataset scale and diversity (Tab. 3).

Methods	**R-Precision ↑**			**FID ↓**	**MM-Dist ↓**	**Diversity ↑**
	Top-1	Top-2	Top-3			
T2M-GPT (baseline)	$0.416_{\pm .006}$	$0.627_{\pm .006}$	$0.745_{\pm .006}$	$0.514_{\pm .029}$	$3.007_{\pm .023}$	$10.92_{\pm .108}$
My T2M-GPT	0.427	0.642	0.764	0.430	2.960	10.92

Table 3: Evaluation metrics for baseline and our trained T2M-GPT on KIT-ML.

5.1.4 Qualitative Comparison

A qualitative comparison highlights distinct differences between TEACH and T2M-GPT. With the single-scene prompt *"a person sprints and jumps"*, shown in Fig. 1, the motion generated by T2M-GPT is smoother and more realistic than that of TEACH. This suggests that T2M-GPT is better suited for short, isolated actions, consistently outperforming both TEACH and TEMOS in such scenarios.

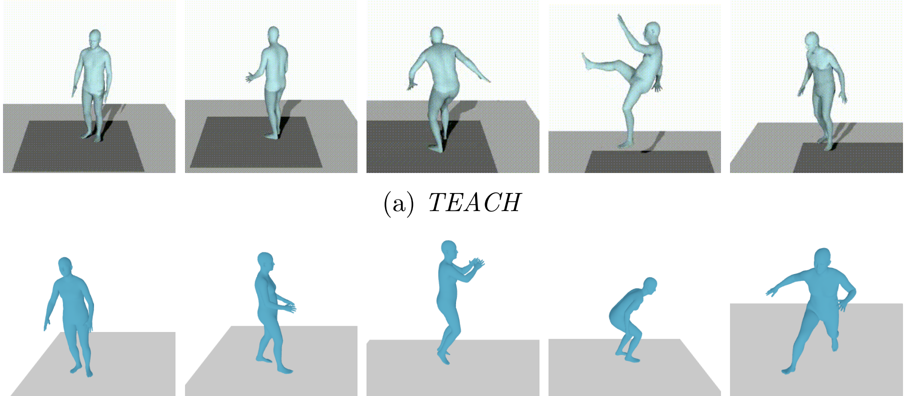


(a) *TEACH*

(b) *T2M-GPT*

Figure 1: Comparison of generated motions by T2M-GPT [12] and TEACH [1] with single-scene prompt.

In contrast, for more complex multi-action prompts such as *"a person turns around, then jumps and kicks, then turns around"*, TEACH better preserves the temporal structure of the description (Fig. 2). The generated sequence exhibits turning, a kicking motion, and a subsequent transition, aligning closely with the prompt. By comparison, T2M-GPT produces plausible local actions but fails to capture the intended sequence, instead repeating jumping behaviors with limited adherence to the specified order. These observations confirm that TEACH's sequential conditioning mechanism provides a clear advantage for multi-scene generation tasks.



(a) *TEACH*

(b) *T2M-GPT*

Figure 2: Comparison of generated motions by T2M-GPT [12] and TEACH [1] with multi-scene prompt.

5.2 Data Analysis

5.2.1 HumanML3D

The HumanML3D [3] dataset serves as a central benchmark for text-to-motion generation and thus strongly influences model performance. Captions in this dataset are typically obtained by assigning annotators to describe pre-recorded motions. While this procedure produces large-scale resources at relatively low cost, it is also prone to errors and inconsistencies. Several unique categories of issues were observed, ranging from typographical mistakes (e.g., *"twords"* instead of *"towards"*, sample 000193) to confusing or incomplete descriptions (e.g., *"he person jumps to the left. AND he cleans a so much"*, sample 008437), and in rare cases entirely uninformative captions (e.g., *"no idea i am so sorry"*, sample 000193). Such deviations reduce dataset quality and can negatively affect model training.

To systematically identify problematic entries, the previously defined **Intra-Motion Similarity** (IMS) was deployed. For example, in the case of the uninformative caption *"no idea i am so sorry"*, the normalized similarity to the other captions of the same motion was only 0.64, indicating clear semantic divergence. Another low score was calculated for the similarity of *"A person walks in a circle."* and *"A human performs a jump to the left."*. This highlights the utility of IMS for filtering noisy or inconsistent annotations. Providing a way to analyze and improve the quality of existing data.

5.2.2 Intra-Motion Similarity

Applying IMS across the entire dataset provides a measure of overall semantic coherence. The HumanML3D captions yielded an average similarity of average = 0.9035, with minimum and maximum values of 0.327 and 0.99, respectively. Setting thresholds on this score enables automatic filtering of inconsistent motions (Tab. 4).

IMS Threshold	**# Discarded motions**
0.6	2710
0.7	10423
0.8	22088
0.9	26104

Table 4: Number of discarded motions per IMS threshold, out of 29,216 total [3].

Filtering the training data by discarding motions whose descriptions fall below a similarity threshold can improve model quality, but it also reduces dataset size, creating a trade-off between quality and quantity. To examine this effect, T2M-GPT was retrained on multiple filtered subsets defined by different intra-motion similarity thresholds. Although the results are not statistically conclusive due to the limited number of training repetitions, they provide initial insights into the potential benefits of this approach. In our experiments, thresholds of 0.6 and 0.8 were applied, based on similarity scores obtained with the MPNet model [10].

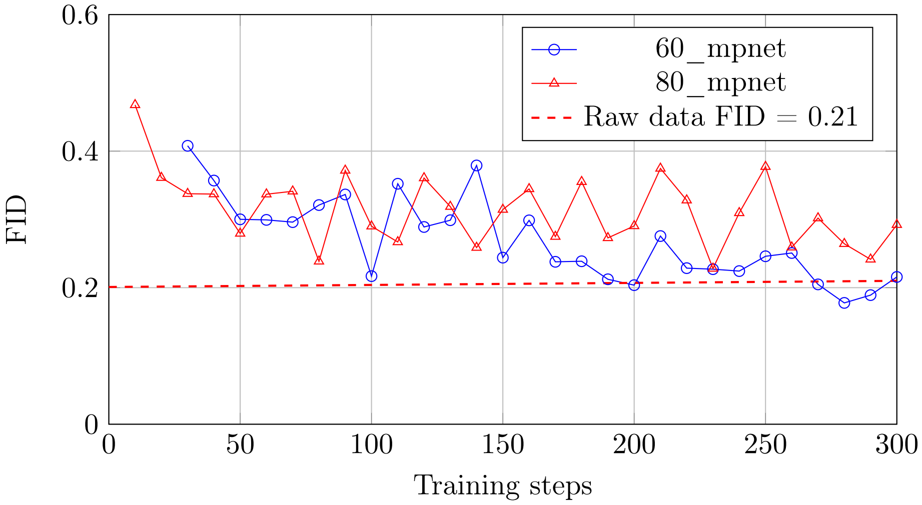


Figure 3: FID scores over training steps for filtered datasets compared to raw-data baseline.

The results in Fig. 3 indicate that the T2M-GPT models trained on filtered subsets (thresholds 0.6 and 0.8, using IMS with MPnet model, trained for 2 epochs on text-descriptions) achieve slightly better FID scores than the model trained on the full HumanML3D dataset. However, the improvements remain marginal and are not statistically significant. A promising future direction would be to refine the filtering process by discarding only inconsistent or erroneous captions rather than removing entire motion sequences, thereby preserving more training data while still improving quality.

5.2.3 Gender Bias

Additionally, it is further analyzed, how actors are referenced in textual descriptions, using noun and pronoun categories (*man*, *woman*, *person*, etc.). As shown in Fig. 4, the distribution is strongly skewed towards neutral references, with male nouns occurring significantly more often than female ones.

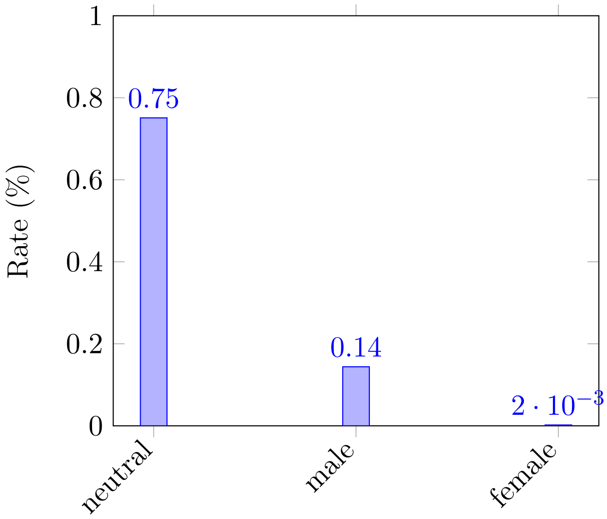
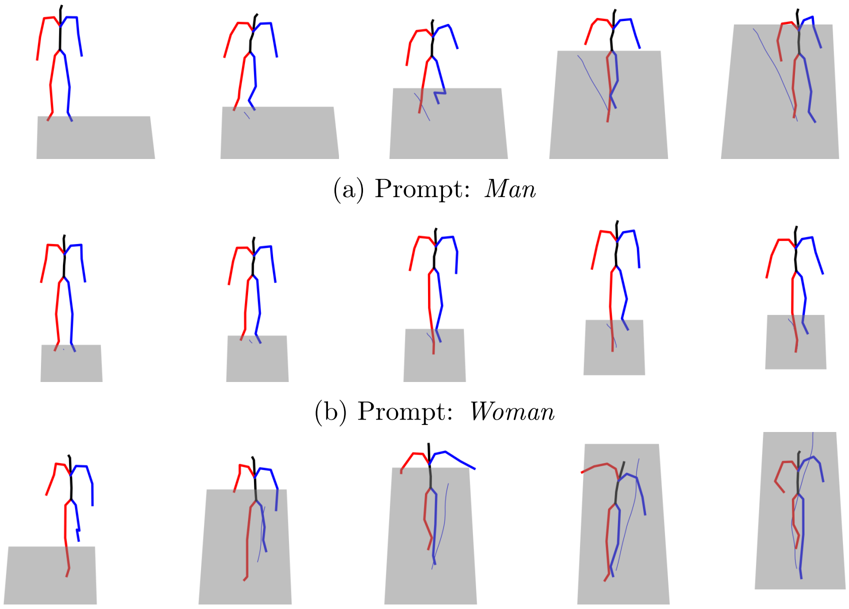


Figure 4: Distribution of actor labels across categories in HumanML3D.

This imbalance is not only quantitative but also leads to qualitative differences in generated motions. Figure 5 illustrates results obtained with the prompt *"a <gender> sprinting and jumping"*. While the neutral form *"person"* produces the intended behavior, the male variant *"man"* only exhibits sprinting, and the female variant *"woman"* fails to generate either sprinting or jumping. These discrepancies suggest that gendered references bias the model and can reduce output quality, highlighting a critical fairness challenge in text-to-motion generation.



(a) Prompt: *Man*

(b) Prompt: *Woman*

(c) Prompt: *Person*

Figure 5: Comparison of generated motions by T2M-GPT [12] under different gender references.

6 Conclusion and Future Work

In this work, the task of generating 3D human motion from natural language descriptions using the SMPL body model [4] was investigated. To address limitations of individual approaches, a modular API framework that integrates multiple state-of-the-art models was developed, TEACH, and T2M-GPT. This design enables users to generate motions across different paradigms—single-scene, sequential, and dataset-specific—and select the most suitable results.

The experiments demonstrated that retraining existing models on benchmark datasets such as KIT-ML, HumanML3D, and BABEL reproduces reported results with only minor deviations, thereby validating the reproducibility of prior work. Moreover, training on larger and more diverse datasets, particularly HumanML3D, substantially improves motion quality compared to smaller corpora such as KIT-ML.

Beyond model evaluation, dataset characteristics were analyzed and uncovered critical challenges. Using intra-motion similarity identified inconsistent or erroneous annotations in HumanML3D [3], illustrating the potential of semantic filtering for improving data quality. Furthermore, our analysis revealed strong gender imbalances in captioning, which directly influenced model outputs and exposed fairness concerns in text-to-motion generation.

Taken together, these findings highlight both the progress and the limitations of current methods. While modular integration of models can improve robustness and diversity, overall performance remains constrained by dataset quality and annotation bias. Future work should therefore focus on improving data curation practices, exploring bias-mitigation strategies, and extending model architectures to better generalize across heterogeneous training corpora. Such advances will be essential for deploying text-to-motion generation in real-world applications such as animation, virtual reality, and human–robot interaction.

During the course of this project, several limitations became apparent. The primary goal was to generate motion from textual descriptions using a custom SMPL model. However, many components of the existing research infrastructure, as well as SMPL itself, relied on dependencies that were often incompatible with one another. Implementations of prior work, such as TEACH [1], are highly complex, and in some cases even the maintainers appear uncertain about specific code segments, as indicated by comments like *"I'm too tired for this sorry"*. These issues made it challenging to develop a stable system that consistently supports custom SMPL models. The situation was further complicated by restrictive licenses in related projects, which limited the reuse of existing code. Although the required functionality is technically available, it is fragile in practice and appears to have been primarily tested with standard SMPL models, with only limited support for custom extensions that nevertheless conform to the SMPL specification [4].

A Appendix: Additional Materials

A.1 Repository for API Service Implementation

The codebase for the API Service is available here: `https://github.com/hannesmoehring/dlhm_prod`. Be sure to read the provided README.md file as it goes into more detail on how to use it.

A.2 Repository Data Analysis

To learn more about the means of the completed analysis, more results and the implementations are provided here: `https://github.com/hannesmoehring/data_analysis_smpl.git`.

References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. Teach: Temporal action compositions for 3d humans. In *International Conference on 3D Vision (3DV)*, September 2022.

[2] Xuehao Gao, Yang Yang, Zhenyu Xie, Shaoyi Du, Zhongqian Sun, and Yang Wu. Guess gradually enriching synthesis for text-driven human motion generation. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

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

[4] 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 2015.

[5] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, October 2019.

[6] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.

[7] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, dec 2016.

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

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[10] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020.

[11] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *arXiv preprint arXiv:2203.08063*, 2022.

[12] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.