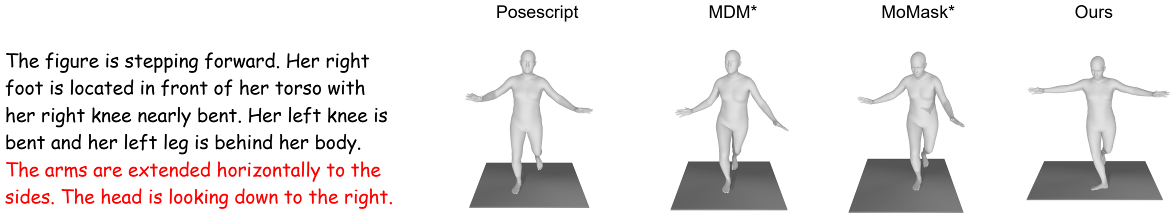
Semantic Mask Transformer for 3D Human Pose Generation with Detailed Text Description

Anonymous submission



The figure is stepping forward. Her right foot is located in front of her torso with her right knee nearly bent. Her left knee is bent and her left leg is behind her body. The arms are extended horizontally to the sides. The head is looking down to the right.

Figure 1: Our Semantic Mask Transformer (SMT) can generate high-quality 3D human poses aligned with the detailed text descriptions while existing methods suffer from semantic absences, highlighted in red.

Abstract

Previous methods for generating 3D human poses from detailed text descriptions often encounter semantic mismatches and struggle to produce precise local body part actions. These challenges primarily arise from the limited variety of body part action combinations in existing datasets. To address these issues, we introduce the Semantic Mask Transformer (SMT), a text-driven animation framework designed to synthesize 3D poses that closely align with detailed textual descriptions. The key innovation of SMT lies in its integration of semantic biases derived from a Large Language Model into the training objectives, thereby enhancing local semantic consistency. Equipped with mask data augmentation, body part modeling, and semantic bias training objectives, our SMT effectively generates high-quality poses while maintaining accurate semantic alignment with the input descriptions. Furthermore, the ablation study demonstrates that the semantic bias objectives can be applied across various backbone architectures.

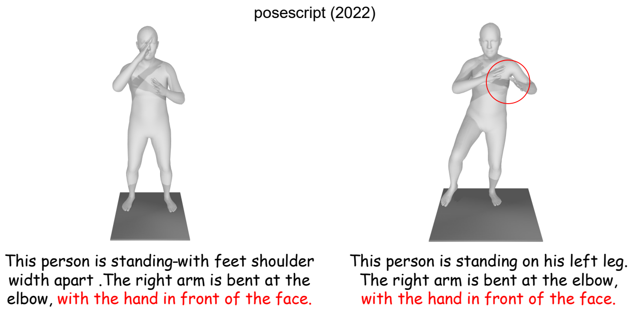
Introduction

This paper tackles the challenge of text-to-pose generation, which involves generating a plausible and realistic 3D human pose from a detailed textual description. This capability is highly desirable in the gaming and film industries, where automated and accurate pose generation from detailed text description can significantly enhance visual storytelling and character animation ((Lan et al. 2023), (Zhao et al. 2022), (Chen, Peng, and Zhou 2021)). The process requires extensive textual input, specifying an overall action label and detailed descriptions of individual body parts and their relative positions. The complexity of this task lies in generating a pose that is consistent with the overall action label and aligns with the semantic requirements of the detailed description.

To generate poses from detailed text descriptions, Posescript (Delmas et al. 2022) employs a contrastive Language-Pose training approach to align the pose latent space with the textual space. Chatpose (Feng et al. 2023) quantizes human poses into distinct signal tokens within a multimodal LLM, enabling the direct generation of 3D body poses from both textual and visual inputs. PRO-Motion (Liu et al. 2023) introduces a posture diffusion model to generate human poses with text descriptions from GPT-like Large Language Model(LLM) (Brown et al. 2020).

However, existing approaches typically face challenges such as semantic mismatches or overlooking body part actions, which limit models to simple text descriptions. When prompts involve complex descriptions that require the coordination of multiple body parts, the resulting poses often omit certain parts. Our analysis identifies that shortcoming of these methods mainly comes from the combination bias regarding body part actions in the dataset, which complicates the model's ability to handle unseen combinations of previously observed body part actions. For example, in Fig 2 consider the action *”The right hand is in front of the face.”*. This action is frequently paired with the detail that *”Stand with feet shoulder-width apart”*. Thus, when tasked with generating a pose from the combination of *”Stand on his left leg”* and *”The right hand is in front of the face.”*, the model fails to include the action of raising the right hand in the generated pose due to their training on more common, less varied combinations.

To mitigate the combination bias in the dataset and generate diverse animations, SINC (Athanasiou et al. 2023) employs LLM to decompose text descriptions for each body part, thereby creating a limited, unbiased spatial composition synthesis dataset for text-to-motion generation. Lgtm (Sun et al. 2024) further trains text-to-motion models for each body part to capture local semantic consistency. Given that textual descriptions for human pose generation are considerably more complex and often involve specific coordination and alignment of multiple body parts, they cannot be easily segmented into body part captions or simply recombined to yield consistent descriptions. The combination bias



This person is standing-with feet shoulder width apart .The right arm is bent at the elbow, with the hand in front of the face.

This person is standing on his left leg. The right arm is bent at the elbow, with the hand in front of the face.

Figure 2: Posescript (Delmas et al. 2022) suffers from the body part combination bias in the dataset (The action ”right hand in front of the face” combines more frequently with ”standing with feet shoulder width”) which complicates the model’s ability to handle unseen combinations of previously observed body part actions.

of body part actions hinders the model’s ability to learn the joint distribution of textual descriptions and body part actions. Consequently, we introduce the semantic bias from LLM into the loss function to strengthen the connection between body part actions and their corresponding textual descriptions. To achieve this, we randomly mask the textual descriptions at the sentence level to create semantically biased descriptions that involve limited body parts. The loss objective is then computed only on the semantic-related body parts to establish a more precise association between specific body part actions and their corresponding textual descriptions. The semantic bias is effectively represented as a binary loss mask generated by LLM for each body part, where semantically related body parts are assigned a value of 1, while all other body parts are assigned a value of 0.

Our method, named the Semantic Mask Transformer, integrates a mask transformer with a body part vector quantize autoencoder(BPVQ-VAE) to specifically address the challenges of generating human poses semantic consistent with detailed text description. Specifically, we tackle the issue of combination bias at three distinct levels. At the model level, we adopt the BPVQ-VAE which encodes body parts actions into separated tokens. At the data level, the mask transformer is trained as a generative masked model(Chang et al. 2022). The input of the mask transformer consists of a sequence of masked body part tokens sequence and independently masked text descriptions, which mitigates the combination bias at both the body part and textual description levels. Regarding the loss function, since we apply a text mask at the sentence level to accommodate a variety of descriptions, the loss function is computed exclusively on the semantically relevant body parts, which serve as the semantic bias.

To evaluate the effectiveness of our proposed semantic bias loss function, we conducted experiments using various backbone architectures. The ablation studies demonstrate that the semantic bias loss function can be applied across different pipelines to enhance the semantic consistency between body part actions and their corresponding text descriptions. Comparative experiments indicate that our proposed framework generates vivid human pose semantics that are consistent with detailed text descriptions and outperforms state-of-the-art methods. Notably, we are the first to utilize semantic bias extracted from LLM as a binary loss mask for animation generation.

In summary, our contributions are as follows:

- We are the first to integrate semantic biases derived from the semantic prior of LLM into training objectives, thereby enhancing body part semantic consistency in human pose generation.
- We propose a framework that addresses the challenge of pose generation based on detailed text descriptions by leveraging mask data augmentation, body part modeling, and semantic bias loss objectives. With body part modeling and masking, we mitigate the combination bias of body part actions. The application of text description masking and semantic bias objectives enables our model to establish accurate connections between body part actions and their corresponding descriptions.
- Our framework achieves state-of-the-art performance in the challenging task of maintaining semantic consistency between generated poses and detailed text descriptions.

Related Work

Generative Masked Modeling. BERT(Devlin et al. 2018) introduces a masked modeling approach for language tasks, where word tokens are randomly masked at a fixed ratio, and a bidirectional transformer is then tasked with predicting these masked tokens. MAE (He et al. 2022) introduces mask modeling into computer vision as a pre-training task. While MAE serves as a robust pre-trained encoder, it cannot synthesize novel samples. Addressing this limitation, MaskGIT (Chang et al. 2022) proposes an innovative approach where tokens are masked at a variable and traceable rate controlled by a scheduling function, allowing for the iterative synthesis of new samples following the scheduled masking. Muse (Chang et al. 2023) applies this to text-to-image editing, demonstrating the versatility of masked modeling in creative tasks. Magvit (Yu et al. 2023) introduces a versatile masking strategy tailored for multi-task video generation. MoMask (Guo et al. 2023) introduces generative masked modeling for human motion synthesis. An advantage of mask modeling is that it could generate a variety of masked body part combinations which could break the combination bias.

Text-conditioned Human Motion Synthesis. Posescript (Delmas et al. 2022) links the 3D human pose with the natural language by mapping 3D poses and textual descriptions into a joint embedding space. ChatPose (Feng et al. 2023) employs Large Language Models to understand and reason about 3D human poses from images or textual descriptions by embedding SMPL poses as distinct signal tokens within a multimodal LLM which can generate human poses from images and textual descriptions.

Human motion synthesis is a domain related to human pose generation, which aims to synthesize human motion sequences with text conditions. TEMOS (Petrovich, Black, and Varol 2022) leverages variational autoencoder to produce human motion distribution with textual descriptions. T2M (Guo et al. 2022) presents a temporal variational autoencoder to synthesize human motions of different lengths from text input. MDM (Tevet et al. 2022) introduces a diffusion-based generative model for human motion generation. MLD (Chen et al. 2023) applies the latent diffusion model to improve motion quality and speed up the generation process. Momask (Guo et al. 2023) proposes a mask transformer combined with a vector quantize autoencoder to generate human motion sequences based on the text description.

The primary distinction between detailed text-conditioned human pose generation and text-to-motion generation lies in the complexity and specificity of the text descriptions. For human pose generation, the text descriptions are highly detailed, focusing on specific body part actions and their relative positions, which necessitates a greater emphasis on body part semantic consistency. This raises higher demands for semantic accuracy and detail in pose generation compared to motion generation.

Part-based Motion Modeling. Separating the human body into distinct segments facilitates the control of motion synthesis at a more granular level, allowing for precise body part control and alignment. PAN (Hu et al. 2023) encodes body part motion into separate features to introduce body part correspondence prior information into motion retargeting. Motion Puzzle (Jang, Park, and Lee 2022) performed style transfer at the part level, utilizing a graph convolutional network to assemble different body part motions into new, coherent sequences, preserving local styles while transferring them to specific body parts without compromising the integrity of other parts or the entire body. LGTM (Sun et al. 2024) employs a large language model to decompose textual descriptions into part-specific narratives and train independent body-part motion encoders to ensure precise local semantic alignment.

Spatial Composition Synthetic Data Training with spatial composition synthetic data is another way to remove the influence of the dataset bias about the combination of body part action. SINC (Athanasiou et al. 2023) creates spatial composition synthesis data with the help of LLM for the text-to-motion generation BMSS (Soga et al. 2016) synthesized dance motions from existing datasets by focusing on body partitions. Chimera (Lee, Lee, and Lee 2022) compose part animations from a collection of source animations and refine it by a policy trained with reinforcement learning. However, it can only synthesize simple and low-quality human motion and has limited effectiveness in text-to-pose, particularly when faced with more challenging descriptions. For instance, consider the complex action described as "bending over with two hands touching feet". This pose involves specific coordination and alignment of multiple body parts which can not be synthesized with body part actions.

Method

Our goal is to generate a 3D human pose with a detailed textual description **c**. Our framework comprises two components: body part VQ-VAE(BPVQ-VAE) and mask transformer. To mitigate the combination bias present in the dataset, we employ semantic bias training for the mask transformer, which includes text description masking, semantic mask extraction, and semantic bias computation. This approach aims to reduce the influence of combination bias. We will provide an overview of each component in the order of their training.

Body-Part Quantization

Graph neural network has been proven to be a responsible structure to model human topology (Zhang et al. 2024). We model the human pose as a graph according to the skeleton hierarchy where each node corresponds to a joint and each edge represents a directed connection between joints. The pose data can be considered as node features $\mathbf{f_{node}} \in R^9$, which encompass the 6D joint rotation representation and 3D local joint positions.

The BPVQ-VAE consists of a graph encoder and a graph decoder, with unique quantization layers for each body part that map body-part features into tokens from learned codebooks. Specifically, We manually divide the human body into 6 parts: left arm, right arm, left leg, right leg, main body, and head. After the human pose is encoded into features $\mathbf{f} \in R^{J \times D}$ with number of joints **J** and latent dimension **D**, the features are grouped by body part group and construct body part features $\hat{\mathbf{b}}_{1:N} \in R^D$ by group pooling where **N** is the number of body parts. Each body part features are replaced with its nearest code in the codebook which is known as quantization $Q(.)$. The quantized body part code $\mathbf{b}_i = Q_i(\hat{\mathbf{b}}_i)$ is mapped to the origin graph where the joints in the same body part share the quantized code. The graph decoder projects node features into human pose $P \in R^{J \times 6}$ to get 6D rotation for each joint.

To reduce quantization errors, we adopt residual quantization(Yang et al. 2023) denotes as $RQ(.)$ for the quantization layers that iteratively quantize the body part features which is similar to (Guo et al. 2023). The residual quantization represents body part features $\hat{b}$ as **L** ordered code sequences $RQ(\hat{b}) = [b^l]_{l=0}^{L}$. It works as

$$b^{l+1} = Q(r^{l+1}), \quad r^{l+1} = r^l - b^l \tag{1}$$

The final quantized result is the sum of the entire sequence, which is fed into the graph decoder to reconstruct the human pose. Similar to the (Guo et al. 2023), we adopt the indices of the selected codebook entries (namely body part tokens) as the alternative discrete representation. The BPVQ-VAE is trained with reconstruction loss and commit loss. the reconstruction loss is computed as the geodesic loss between the reconstructed joint rotation matrix and the ground truth rotation matrix.

$$\mathcal{L}_{rec} = \sum_{j=1}^{J} \arccos\left(\frac{\mathrm{tr}(R_j \hat{R}_j^T) - 1}{2}\right) \tag{2}$$

Where **J** refers to the number of joints, $\hat{R}_j$ is the ground truth rotation matrix of j joint while R_j is the prediction. The commit loss is computed as:

$$\mathcal{L}_{com} = \sum_{i=1}^{N} \sum_{j=1}^{L} \|z_{i,j} - Q(z_{i,j})\|_2 \tag{3}$$

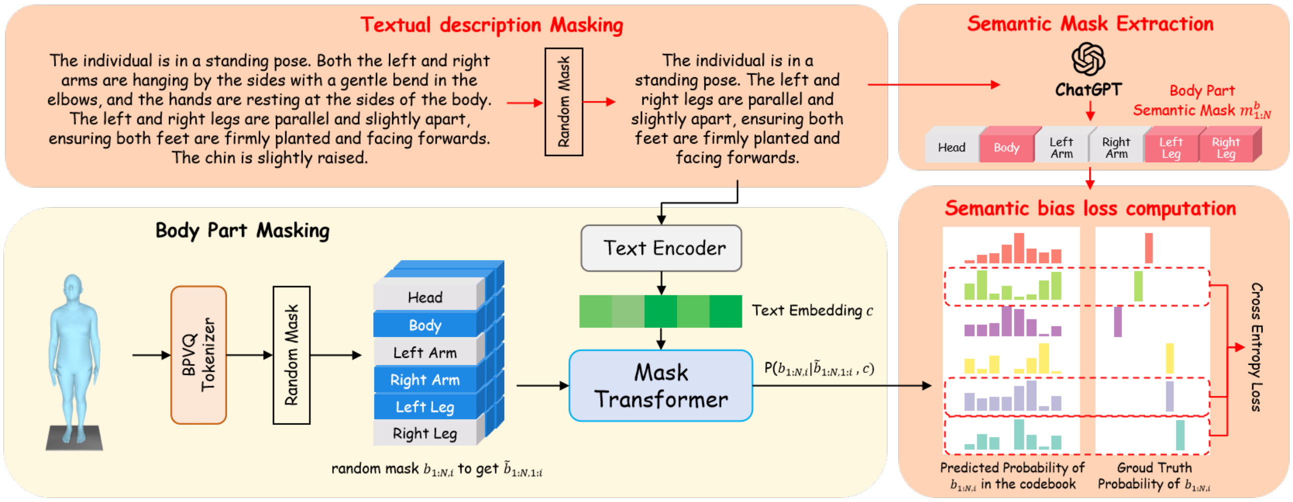

Textual description Masking
The individual is in a standing pose. Both the left and right arms are hanging by the sides with a gentle bend in the elbows, and the hands are resting at the sides of the body. The left and right legs are parallel and slightly apart, ensuring both feet are firmly planted and facing forwards. The chin is slightly raised.
Random Mask
The individual is in a standing pose. The left and right legs are parallel and slightly apart, ensuring both feet are firmly planted and facing forwards.
Semantic Mask Extraction
ChatGPT
Body Part Semantic Mask $m^b_{1:N}$
Head
Body
Left Arm
Right Arm
Left Leg
Right Leg
Body Part Masking
BPVQ Tokenizer
Random Mask
Head
Body
Left Arm
Right Arm
Left Leg
Right Leg
random mask $b_{1:N,i}$ to get $\tilde{b}_{1:N,1:i}$
Text Encoder
Text Embedding c
Mask Transformer
$P(b_{1:N,i}|\tilde{b}_{1:N,1:i}, c)$
Semantic bias loss computation
Cross Entropy Loss
Predicted Probability of $b_{1:N,i}$ in the codebook
Groud Truth Probability of $b_{1:N,i}$

Figure 3: Semantic Bias Training for the Mask Transformer. Through textual description masking, we create semantically biased descriptions that do not encompass all body parts. Subsequently, a semantic mask $\mathbf{m}^b_{1:N}$ is generated by LLM, indicating whether a body part is semantically related to the descriptions. This semantic mask is applied to the loss computation, resulting in the loss concentrating on the semantically related body parts. Additionally, we apply body part masking to the input of the mask transformer to further mitigate the combinatorial bias of body part actions.

where i refers to the number of body parts, and j is the index of the quantized result in code sequences.

The overall objective function is defined as follows:

$$\mathcal{L}_{bp} = \alpha \mathcal{L}_{rec} + \beta \mathcal{L}_{com} \tag{4}$$

Semantic Mask Extraction

To introduce semantic bias into the loss objective, we first randomly mask the textual description at the sentence level resulting in semantically biased descriptions that do not encompass all body parts. Subsequently, a semantic-related mask, denoted as $\mathbf{m}^b_{1:N}$, is extracted by LLM from the text descriptions corresponding to each body part. In contrast to traditional rule-based methods, LLM is capable of inferring not only explicit but also implicit body parts mentioned in the text, thereby outperforming rule-based methods when addressing high-level action labels. For instance, consider the description, "The character is looking backward." This action primarily involves the neck and head, which can be seamlessly identified by the language model. **More details on extracting semantic masks with LLM and a comparison between LLM and rule-based methods can be found in the supplementary material**.

Mask Transformer

The mask transformer is a decoder-only transformer. It takes the quantized body part token sequences and text embedding $\mathbf{c}$ from a pre-trained T5 encoder(Raffel et al. 2020) as inputs. Similar to (Guo et al. 2023), we randomly mask out body part tokens with a special [MASK] token and represent the masked body part tokens as $\tilde{b}$. It is distinguished that different from the previous masked generative method, our goal is to predict semantic-related body part tokens given text condition instead of masked tokens. To make textual description more suitable for computing semantic bias, we also apply random masking for textual description at the sentence level independently. And a semantic-related mask $\mathbf{m}^b_{1:N}$ for each body part is generated in which the related body part is 1 and 0 otherwise. The output of the mask transformer is a set of logits corresponding to the codebook size of BPVQ-VAE. The overall objective function is to minimize the negative log-likelihood for the target body part tokens where the negative log-likelihood is computed as the cross-entropy between generated probabilities with ground truth one-hot vector for each body part. The overall objective function contains only the cross-entropy loss for semantic-related body parts. Formally, the loss function is defined as follows:

$$\mathcal{L}_{mt} = \sum_{i=1}^{N} \mathbf{m}^b_i \, log\mathcal{P}(b_i|\tilde{b}, \tilde{c}) \tag{5}$$

Similar to the previous work, since we adopt residual quantization which represents body part features as $\mathbf{L}$ ordered code sequence $b \in R^{N \times L}$ where N refers to the number of body parts, We iteratively recover the body part tokens in order with the sum of previous body part tokens subsequence. Specifically, when recover the i^{th} body part tokens $b_{1:N,i}$ where $b_{n,i} \in \mathbf{R}^D$, the input to the mask transformer is $\tilde{b}_{1:N,1:i} = [\sum_{j=1}^{i} b_{1,j}, ..., \sum_{j=1}^{i} b_{N,j}]$ where $b_{n,i}$ is [MASK] if body part n is masked. The overall forward process of mask transformer $F(.)$ is defined as follows:

$$\mathcal{P}_{1:N,i} = F_\theta(\sum_{m=1}^{i} \tilde{\mathbf{b}}_{1:N,m}, c) \tag{6}$$

where c refers to the text embedding output by the pre-trained T5 encoder. A learnable position embedding and a body part embedding are added to the input body part tokens to make the mask transformer aware of the body parts and which index of the L sequence it operates on.

```
Algorithm 1: Iterative Inference with Semantic Bias
Input: sentences $\mathbf{c}_{1:k}$
       sentence level semantic mask $\mathbf{m}_{1:k}$
       mask transformer $F(.)$
       sample function $S(.)$
Parameter: length of the token sequence $\mathbf{L}$
Output: body parts tokens sequence $\mathbf{b} \in R^{N \times L}$
Let $\mathbf{b}_{1:N,1:L}$ =[MASK].
for i = 1:L do
    for j = 1:k do
        $\hat{\mathbf{b}}_{1:N,i} = S(F(\sum_{k=1}^{i} \mathbf{b}_{1:N,k}, c_{1:j}))$
        $\mathbf{b}_{1:N,i} = m_j \hat{\mathbf{b}}_{1:N,i} + (1 - m_j)\mathbf{b}_{1:N,i}$
    end for
end for
return b
```

The mask ratio to conduct masked body part input is sampled from a uniform distribution $\mathcal{U}[0, 1]$.

Iterative Inference with Semantic Bias

To close the gap between training and inference, we adopt iterative inference. Similar to (Chang et al. 2023) and (Guo et al. 2023), the process begins with all tokens masked and our mask transformer predicts body part probabilities for sampling the masked. Different from the previous method, we also introduce semantic bias into the inference. Specifically, given textual description $\mathbf{c}_{1:k}$ consists of k sentences, a sentence level semantic mask $\mathbf{m}_{1:k} \in \mathbf{R}^N$ is provided where $m_{i,n} = 1$ if i^{th} sentence is semantic related to n^{th} body part. The whole process can be summarized as Algorithm 1.

Experiment

Settings

Datasets. We train and evaluate our method on the Posescript dataset with 100,000 rule-based automatic captioned poses and 6283 human-labeled data with mirror augmentation. The partitioning of the dataset into training and test sets follows the same division as that used in the PoseScript, ensuring consistency and comparability in our evaluations. we have made several improvements to the automatic caption pipeline. These enhancements are designed to produce captions that more closely mimic human annotation, both in terms of linguistic naturalness and the accuracy with which they describe the corresponding poses. **The details on the improvements in the automatic captioning pipeline can be found in the supplementary material**.

Implementation details. We leverage a pre-trained T5 language model as our frozen text encoder to transform input textual descriptions into a sequence of 1024-dimensional embeddings. As the description for pose generation is often longer than 77 which is the max input length for the CLIP text encoder, the T5 language model can carry richer information about detailed body part action. **More implementation details on Body part VQ-VAE and Mask Transformer can be found in the supplementary**.

Objective	FID ↓	MPJPE ↓	ITM ↑	Diversity →
Ground Truth	-	-	0.89	-
Posescript(Delmas et al. 2022)	1.88	0.048	0.50	9.57
MDM*(Tevet et al. 2022)	0.90	0.042	0.62	9.71
MoMask*(Guo et al. 2023)	0.72	0.035	0.66	7.11
SMT(ours)	**0.23**	**0.018**	**0.79**	8.96

Table 1: Quantitative comparison with the state of the arts. FID is Fréchet inception distance of motion semantics. MPJPE denotes the mean per joint position error. ITM indicates the image-text matching score.

Objective	FID ↓	MPJPE ↓	ITM ↑	Diversity →
BERT-style (Devlin et al. 2018)	0.56	0.033	0.69	8.84
MASS-style (Song et al. 2019)	0.57	0.035	0.67	8.51
Semantic-biased-style	**0.23**	**0.018**	**0.79**	8.96

Table 2: Quantitative comparison on semantic bias. All three masking objectives involve randomly masking input tokens. The goal of the BERT-style objective is to predict masked tokens, while the MASS-style approach aims to predict all tokens. These two masking training strategies do not introduce semantic bias. In contrast, the goal of our semantic bias strategy is to predict tokens that are semantically related.

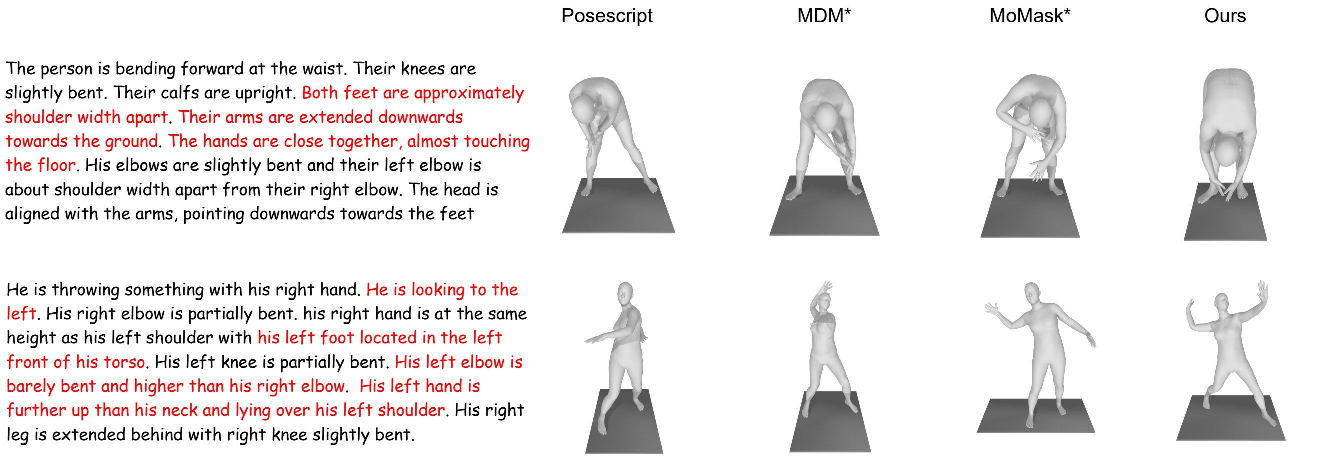
Evaluation metrics. We introduce Image-Text Matching (ITM) score, Fréchet inception distance (FID), Average Positional Error (MPJPE), and Diversity as evaluation metrics. The Image-Text Matching (ITM) score is proposed by (Zhang et al. 2023) which quantifies the visual-semantic similarity between the source textual description and the rendered generated pose. The generated pose is rendered into three images from different views. The pre-trained Vision Language Model BLIP2 (Li et al. 2023) outputs the logits whether the images and their corresponding descriptions are matched. Image-text matching (ITM) scores are the top match logits across the three views.

Comparison with State of the Arts

In this section, we conduct a comprehensive comparative analysis of our method against several state-of-the-art approaches in both human pose generation and human motion generation. Given that descriptions for pose generation typically require more detail than those for motion generation, we have adapted these human motion generation methods to better suit the specific needs of human pose generation. We replaced the original CLIP text encoder and transformer encoder layers with a T5 encoder and transformer decoder layers to improve the model's capacity to handle detailed text descriptions. The baseline methods includes Posescript (Delmas et al. 2022), MDM (Tevet et al. 2022), and MoMask (Guo et al. 2023).

Quantitative. The comparative analysis of our method against the state-of-the-art approaches is presented in Table 1. Posescript embeds the text descriptions and human poses into a shared latent space, neglecting the detail and demand in the text description. The MDM and MoMask also fail to maintain semantic consistency between body parts with detailed text descriptions. Notably, our model exhibits the best FID and ITM score among all methods, showcasing the capability of the proposed framework to produce high-quality poses with semantics consistency.

Qualitative. In Figure 4, we present a qualitative compari-



The person is bending forward at the waist. Their knees are slightly bent. Their calfs are upright. Both feet are approximately shoulder width apart. Their arms are extended downwards towards the ground. The hands are close together, almost touching the floor. His elbows are slightly bent and their left elbow is about shoulder width apart from their right elbow. The head is aligned with the arms, pointing downwards towards the feet

He is throwing something with his right hand. He is looking to the left. His right elbow is partially bent. his right hand is at the same height as his left shoulder with his left foot located in the left front of his torso. His left knee is partially bent. His left elbow is barely bent and higher than his right elbow. His left hand is further up than his neck and lying over his left shoulder. His right leg is extended behind with right knee slightly bent.

Figure 4: Qualitative comparison. The result demonstrates that our method can effectively preserve the body part semantic consistency with the description. From the first column to the last column are the text descriptions, posescript (Delmas et al. 2022), MDM* (Tevet et al. 2022), MoMask*(Guo et al. 2023) and our method respectively. **More results can be found in the supplementary**.

Objective	FID ↓	MPJPE ↓	ITM ↑	Diversity →
Posescript(Delmas et al. 2022)	1.88	0.048	0.50	9.57
Posescript + semantic bias	0.77	0.038	0.66	9.08
MDM*(Tevet et al. 2022)	0.90	0.042	0.62	9.71
MDM* + semantic bias	0.36	0.035	0.69	9.66

Table 3: Quantitative comparison with different backbones networks. All metrics improve when the semantic bias training objective is introduced

inference type	FID ↓	MPJPE ↓	ITM ↑	Diversity →
inference without bias	0.27	0.022	0.76	9.33
inference with bias	**0.23**	**0.018**	**0.79**	8.96

Table 4: Quantitative comparison for inference with and without semantic bias

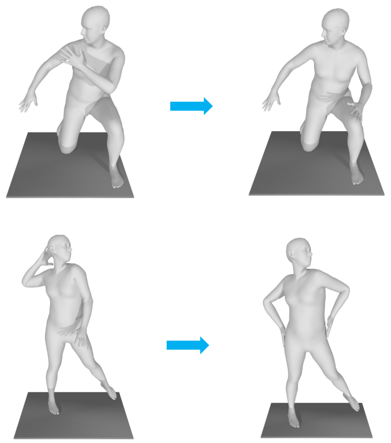
son between state-of-the-art methods and our approach. This comparison underscores the effectiveness of our method in preserving semantic consistency. PoseScript (Delmas et al. 2022) struggles to accurately capture essential semantic elements from detailed text descriptions. This limitation becomes particularly evident when addressing complex actions with numerous details, such as the second kicking pose. In such cases, PoseScript may not only overlook specific details but may also fail to grasp the overall semantics, resulting in significant errors in pose generation, such as producing a crouching pose instead of the intended kicking action. MDM (Tevet et al. 2022) and MoMask (Guo et al. 2023), on the other hand, are capable of preserving important semantic content but fall short in capturing crucial details. In contrast, our approach outperforms the other methods in terms of semantic preservation, achieving the most reliable and accurate pose generation based on detailed text descriptions.

Ablation Study

The effect of semantic bias. To rigorously evaluate the effectiveness of our semantic bias training objectives, we conducted a high-level ablation study where we compared our approach against two other masking objectives: BERT-style (Devlin et al. 2018) and MASS-style (Song et al. 2019), both of which lack a semantic bias component. Each of these three masking methods involves randomly masking input body part tokens, but they differ significantly in their reconstruction goals. The BERT-style objective focuses on reconstructing the masked tokens. In contrast, the MASS-style objective aims to reconstruct all tokens. Our semantic bias objective, however, specifically targets the reconstruction of semantic-related tokens. This means that the training process is particularly focused on ensuring that the model accurately interprets and generates those parts most closely related to the semantics of the input text. The quantitative results, as presented in Table 2, clearly demonstrate the benefits of incorporating semantic bias into the training objective. The introduction of semantic bias significantly enhances the semantic consistency of the generated poses.

Semantic bias with different backbone. To demonstrate the effectiveness of our proposed semantic bias training objective, we integrate this training strategy with different backbones. Specifically, we apply it to the Posescript and MDM pipelines. The MoMask(Guo et al. 2023), which embeds the entire human pose into a single token, does not support the body part semantic bias training objective and, therefore, is excluded. By incorporating semantic bias, the model is encouraged to establish associations between specific body part actions and their corresponding textual description, leading to more accurate and semantically consistent pose generation. The quantitative results, presented in Table 3, indicate that all metrics improve with the semantic bias training objective. This enhancement is observed across various baselines, substantiating that the semantic bias training objective can effectively enhance the semantic consistency of pose generation across different architectures. This finding highlights the potential of semantic bias as a powerful tool for refining the training process in text-to-pose synthesis tasks and potentially in other generative tasks.

Iterative inference with semantic bias As we incorporate semantic bias into inference, we also conduct an ablation study on the effectiveness of the proposed inference method. We compare our method with iterative inference without semantic bias (Chang et al. 2023) (Guo et al. 2023), where



Someone is kneeling on her right knee. Her right knee is bent at an angle of almost 90 degrees and is touching the ground. Her right foot is behind her torso, firmly planted on the ground. ~~Her left elbow is in front of her chest, bent in an L-shape with her left hand near her right shoulder~~ (Her left elbow is bent in an L-shape with her left hand reaching her left knee). Her left knee is bent at an angle of almost 90 degrees and her left thigh is parallel to the ground. Her left calf is erect. The right arm is stretched out to the right with the elbow slightly bent. Her head is turned to the right.

A person is standing on their right leg and looking to the left. The torso is straight and slightly turned to the right. Both legs are straight. The left shoulder is in front of the right shoulder, the left leg is stretched back to the left and the toe touches the ground. ~~The right elbow is flexed to the maximum with the hand on the back of the head. The left elbow is slightly bent and the left hand is in front of the left hip~~ (His both elbows are bent at an angle of almost 90 degrees with the hands putting on the hips).

Figure 5: Examples of body part editing. From the first column to the last column are the text descriptions, origin pose, and modified pose. The original pose is generated with the original text description. The modified pose is generated using the modified description, where the strikethrough text is replaced with the red text. **More results can be found in the supplementary**.

the update rules $\mathbf{b}_{1:N,i} = S(F(\sum_{m=1}^{i} \mathbf{b}_{1:N,m}, c))$. At each iteration, the masked body part tokens with the highest probability are preserved leaving remains still masked and the iteration repeats until all body part tokens are unmasked. The quantitative results, as shown in Table 4.

body part number	FID ↓	MPJPE ↓	ITM ↑	Diversity →
SMT$_{2bodyparts}$	0.31	0.031	0.72	9.01
SMT$_{3bodyparts}$	0.28	0.025	0.74	8.94
SMT$_{5bodyparts}$	0.24	0.021	0.78	8.87
SMT$_{6bodyparts}$	**0.23**	**0.018**	**0.79**	8.96
SMT$_{jointlevel}$	0.33	0.031	0.70	8.89

Table 5: Quantitative comparison for different numbers of body parts. as the body is divided more precisely, the preservation of semantic consistency improves. The underperformance of joint-level SMT can be attributed to the limitations of the LLM in accurately generating joint-level labels.

Body part division for semantic bias We further explored the body part division by varying the number of body parts and segmenting the human body into different levels of granularity. The SMT is tested using configurations of 2, 3, 5, and 6 body part groups, as well as at the joint level. In the 2 body part groups configuration, the human body is divided into an upper body group, encompassing the arms, head, and torso, and a lower body group, comprising the legs. The 3-group configuration further delineates the body into arms, the main body (including the neck and head), and legs. The 5-group configuration subdivides the arms and legs into left and right segments. The main body is further divided into the neck and head for 6 groups. The quantitative results, as presented in Table 5, demonstrate that as the body is divided more precisely, the preservation of semantic consistency improves. However, the model with joint-level semantic bias performed worse than the 2 groups. This underperformance can be attributed to the limitations of the LLM in accurately generating joint-level labels for text descriptions. Despite various attempts to prompt the LLM to recognize human body topology, it continues to encounter hallucinations when faced with stochastic descriptions. These hallucinations hinder the model's ability to learn the correct correspondences between body part descriptions and actions. **More details can be found in the supplementary**

Application: Body Part Editing

In Fig. 5, we illustrate the capability of the Semantic Mask Transformer (SMT) in body part editing tasks. The body part can be freely selected. Specifically, we modify the textual descriptions and then follow the same inference procedure described in Algorithm 1. During the generation process, we first generate poses based on the original descriptions. We then modify the descriptions for several body parts and regenerate the corresponding body part tokens. The quantitative results demonstrate that our model effectively captures the association between local pose semantics and their corresponding body parts, generating novel and accurate poses that align with the modified descriptions.

Conclusions

In this paper, we introduce a novel method for detailed text-conditioned human pose generation that leverages a mask transformer combined with a BPVQ-VAE. The mask transformer is trained using a semantic-related mask training objective, specifically designed to enhance the model's understanding of the semantic relationships between body part actions and their corresponding textual descriptions. By employing this training strategy, our Semantic Masked Transformer (SMT) model can generate poses that are not only detailed but also semantically consistent with the provided descriptions. Extensive experiments and comparisons with multiple baselines demonstrate that our proposed method effectively captures and maintains semantic consistency between the text descriptions and the generated poses, outperforming existing approaches in managing complex and detailed pose generation tasks.

References

Athanasiou, N.; Petrovich, M.; Black, M. J.; and Varol, G. 2023. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9984–9995.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11315–11325.

Chen, L.; Peng, S.; and Zhou, X. 2021. Towards efficient and photorealistic 3d human reconstruction: a brief survey. *Visual Informatics*, 5(4): 11–19.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Delmas, G.; Weinzaepfel, P.; Lucas, T.; Moreno-Noguer, F.; and Rogez, G. 2022. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, 346–362. Springer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feng, Y.; Lin, J.; Dwivedi, S. K.; Sun, Y.; Patel, P.; and Black, M. J. 2023. PoseGPT: Chatting about 3D Human Pose. *arXiv preprint arXiv:2311.18836*.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2023. MoMask: Generative Masked Modeling of 3D Human Motions. *arXiv preprint arXiv:2312.00063*.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5152–5161.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hu, L.; Zhang, Z.; Zhong, C.; Jiang, B.; and Xia, S. 2023. Pose-Aware Attention Network for Flexible Motion Retargeting by Body Part. *IEEE Transactions on Visualization and Computer Graphics*.

Jang, D.-K.; Park, S.; and Lee, S.-H. 2022. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 41(3): 1–16.

Lan, C.; Wang, Y.; Wang, C.; Song, S.; and Gong, Z. 2023. Application of ChatGPT-Based Digital Human in Animation Creation. *Future Internet*, 15(9): 300.

Lee, S.; Lee, J.; and Lee, J. 2022. Learning virtual chimeras by dynamic motion reassembly. *ACM Transactions on Graphics (TOG)*, 41(6): 1–13.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, J.; Dai, W.; Wang, C.; Cheng, Y.; Tang, Y.; and Tong, X. 2023. Plan, Posture and Go: Towards Open-World Text-to-Motion Generation. *arXiv preprint arXiv:2312.14828*.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Soga, A.; Yazaki, Y.; Umino, B.; and Hirayama, M. 2016. Body-part motion synthesis system for contemporary dance creation. In *ACM SIGGRAPH 2016 Posters*, 1–2.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Sun, H.; Zheng, R.; Huang, H.; Ma, C.; Huang, H.; and Hu, R. 2024. LGTM: Local-to-Global Text-Driven Human Motion Diffusion Model. *arXiv preprint arXiv:2405.03485*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Yang, D.; Liu, S.; Huang, R.; Tian, J.; Weng, C.; and Zou, Y. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*.

Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A. G.; Yang, M.-H.; Hao, Y.; Essa, I.; et al. 2023. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10459–10469.

Zhang, H.; Chen, Z.; Xu, H.; Hao, L.; Wu, X.; Xu, S.; Xiong, R.; and Wang, Y. 2024. Unified Cross-Structural Motion Retargeting for Humanoid Characters. *IEEE Transactions on Visualization and Computer Graphics*, 1–14.

Zhang, H.; Chen, Z.; Xu, H.; Hao, L.; Wu, X.; Xu, S.; Zhang, Z.; Wang, Y.; and Xiong, R. 2023. Semantics-aware Motion Retargeting with Vision-Language Models. *arXiv preprint arXiv:2312.01964*.

Zhao, Y.; Jiang, J.; Chen, Y.; Liu, R.; Yang, Y.; Xue, X.; and Chen, S. 2022. Metaverse: Perspectives from graphics, interactions and visualization. *Visual Informatics*, 6(1): 56–67.

Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (partial)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)

Does this paper rely on one or more datasets? (yes)

- A motivation is given for why the experiments are conducted on the selected datasets.(yes)
- All novel datasets introduced in this paper are included in a data appendix.(yes)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (yes)

Does this paper include computational experiments? (yes)

- Any code required for pre-processing data is included in the appendix(supplementary). (yes).
- All source code required for conducting and analyzing the experiments is included in a code appendix(supplementary). (yes)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
- This paper states the number of algorithm runs used to compute each reported result. (yes)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests. (yes)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes)