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

Supplementary Material

In this document, we provide the following supplementary content:

- Improvement on Automatic Captioning Pipeline
- Semantic Label Extraction
- Implementation Details
- Additional Results

A. Improvement on Automatic Captioning Pipeline

The original automatic captioning pipeline proposed by the Posescript [4] relies on the extraction, selection and aggregation of elementary pieces of pose information, called posecodes which are eventually converted into sentences to produce a description. The original posecodes consist of various types: Angle posecodes, which describe the bending of a body part at a joint; Distance posecodes, which categorize the space between two keypoints; Special posecodes, assessing the vertical or horizontal orientation of a body part; Relative position posecodes, calculating the positional relationship between different body parts; Ground-contact posecodes, indicating whether a body part is in contact with the ground; and Higher-level posecodes, which extract more complex pose concepts. To produce more accurate and natural descriptions, we have implemented several enhancements to the captioning pipeline.

- **Simplify Posecodes**. Due to the diverse nature of posecodes, there is often redundancy in the captions generated. Consequently, we have merged the relative posecodes with the distance posecodes. Moreover, relative position posecodes are computed only when two body parts are in close proximity.
- **Vertices involved**. When computing the relative position between different body parts, we utilize the bounding box of the vertices associated with each body part rather than treating each joint as a point. This approach allows for a more accurate representation of the spatial relationships between body parts such as *"hands putting on the hips"*, *"arms crossing in front of the chest"*.
- **Simplify Captions**. Since a single body part can be involved in multiple posecodes and multiple body parts can be described with the same description, we have streamlined the captioning process by using the term "Both" and incorporating adverbial adjuncts to simplify and clarify the descriptions.
- **Vision Language Model Relabel**. A description comprises both the descriptions of individual body parts and an overall action label. Originally, the action labels for Posescript were sourced from the BABEL dataset[5], which primarily offered labels depicting motion sequences rather than specific poses, thus lacking diversity. To generate more diverse and accurate pose labels that better describe the specific poses, we have leveraged a vision language model[2] to relabel the poses effectively.
- **Language Model Post Process**. After converting posecodes into sentences to create descriptions, we further utilize a language model [1] to refine and polish these generated descriptions, enhancing their naturalness and readability.

The comparision between origin caption and improved caption are listed in the Fig 1. The improvement on the automatic captioning pipeline make the generated caption more natural and clear for the pose generation task.

B. Semantic Label Extraction

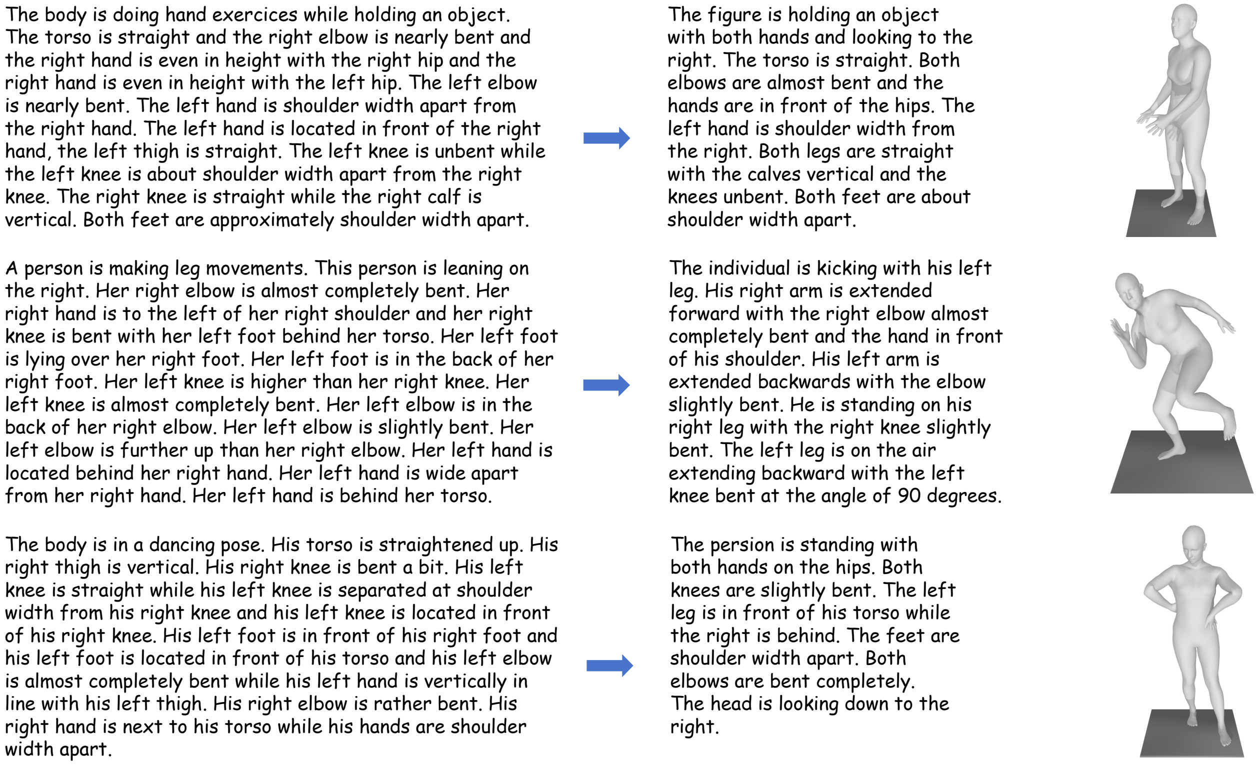
Fig 3 illustrates the complete prompt used in semantic label extraction. We first define the overall objective and task requirements. Next, we define the output format of the LLM and provide examples of pose descriptions for the LLM to reference. These examples are manually labelled. We also attempt joint level label extraction with a modified prompt that explains the SMPL human topology to the LLM in Fig **??**. We test the LLM-based labelling and rule-based labelling with 200 manually labelled captions. The rule-based method is a simple template matching with joint names. The metrics precision and recall are computed as follows:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

where the tp refers to the true positive, fp represents the false positive and fn is the false negative. The GPT 3.5 [3] and GPT 4 [2] may return joints that are not listed in the candidates. And we drop them from the final labels.

The output labels from the rule-based method, utilizing a template matching approach, achieve 100% precision but encounter a low recall rate when faced with stochastic overall descriptions. This limitation restricts the model's capacity to assign descriptions to all pertinent body parts. Conversely, while the LLM-based method might occasionally involve unrelated body parts in the descriptions, it is more effective in assigning descriptions to the majority of relevant body parts.



The body is doing hand exercices while holding an object. The torso is straight and the right elbow is nearly bent and the right hand is even in height with the right hip and the right hand is even in height with the left hip. The left elbow is nearly bent. The left hand is shoulder width apart from the right hand. The left hand is located in front of the right hand, the left thigh is straight. The left knee is unbent while the left knee is about shoulder width apart from the right knee. The right knee is straight while the right calf is vertical. Both feet are approximately shoulder width apart.

The figure is holding an object with both hands and looking to the right. The torso is straight. Both elbows are almost bent and the hands are in front of the hips. The left hand is shoulder width from the right. Both legs are straight with the calves vertical and the knees unbent. Both feet are about shoulder width apart.

A person is making leg movements. This person is leaning on the right. Her right elbow is almost completely bent. Her right hand is to the left of her right shoulder and her right knee is bent with her left foot behind her torso. Her left foot is lying over her right foot. Her left foot is in the back of her right foot. Her left knee is higher than her right knee. Her left knee is almost completely bent. Her left elbow is in the back of her right elbow. Her left elbow is slightly bent. Her left elbow is further up than her right elbow. Her left hand is located behind her right hand. Her left hand is wide apart from her right hand. Her left hand is behind her torso.

The individual is kicking with his left leg. His right arm is extended forward with the right elbow almost completely bent and the hand in front of his shoulder. His left arm is extended backwards with the elbow slightly bent. He is standing on his right leg with the right knee slightly bent. The left leg is on the air extending backward with the left knee bent at the angle of 90 degrees.

The body is in a dancing pose. His torso is straightened up. His right thigh is vertical. His right knee is bent a bit. His left knee is straight while his left knee is separated at shoulder width from his right knee and his left knee is located in front of his right knee. His left foot is in front of his right foot and his left foot is located in front of his torso and his left elbow is almost completely bent while his left hand is vertically in line with his left thigh. His right elbow is rather bent. His right hand is next to his torso while his hands are shoulder width apart.

The persion is standing with both hands on the hips. Both knees are slightly bent. The left leg is in front of his torso while the right is behind. The feet are shoulder width apart. Both elbows are bent completely. The head is looking down to the right.

Figure 1. Comparison between origin and improved captions. From the first column to the last column are the original captions from the automatic pipeline, the improved automatically generated captions and corresponding pose.

Objective	BP mprecision ↑	BP mrecall ↑	JT mprecision ↑	JT mrecall ↑
GPT 3.5	97.8%	98.3%	75.4%	77.7%
GPT 4	98.9%	**99.1**%	84.6%	**85.5**%
Rule based	**100%**	88.3%	**100%**	71.2%

Table 1. We evaluate the LLM-based method and the rule-based template matching method on 200 manually labeled pose descriptions. The BP mprecision means body part level mean precision. The BP mrecall is the body part level mean recall rate. The JT mprecision refers to the joint level mean precision and JT mrecall represents the joint level mean recall rate.

C. Implementation Details

Our models are implemented using PyTorch. Specifically, for the Body Part Group Residual VQ-VAE, we utilize a graph encoder-decoder architecture along with a linear fuzzy module. This group residual VQ-VAE is structured into 6 groups, each comprising 4 quantization layers. Each layer's codebook contains 256 32-dimensional codes. The quantization dropout ratio (q) is set at 0.2. For training the RVQ-VAE, the mini-batch size is uniformly maintained at 512.

The Mask Transformer is trained with a batch size of 256, featuring 8 layers and a latent dimension of 2048. We conduct the training on four A100 GPUs (40G each). Initially, the model is pretrained using automatically generated captions, followed by finetuning with manually labeled captions. The learning rate for pretraining is set at 1e-4, which is reduced by 50% whenever the improvement metrics plateau. For finetuning, the learning rate is reduced further to 1e-5.

D. Additional Results

We have included additional cases in Fig 4 to validate the effectiveness of our proposed method in the task of pose generation from detailed text descriptions. The quantitative results demonstrate the superiority of our method compared to other generation pipelines.

You are a helpful assistant, helping me to extract body parts from the pose description. I will give you a pose description. You must label each sentence in the description with the related body part. There are several main rules you need to follow:

Rule 1. Format of answer.

1.1 Your answer should follow the format:

{'S1':[related body parts], 'S2':[related body parts], ...} where each 'S' represent the sentence in sequence of the description.

Rule 2. Available body parts.

2.1 The available body parts labels are 'left arm', 'right arm', 'left leg', 'right leg', 'torso', 'head'. You must only select body parts from the available body parts label. Neck are included in the 'head' body part. The 'torso' involves waist, stomach and chest.

Example 1:

The individual is kicking with his left leg. His right arm is extended forward with the right elbow almost completely bent and the hand in front of his shoulder. His left arm is extended backwards with the elbow slightly bent. He is standing on his right leg with the right knee slightly bent. The left leg is on the air extending backward with the left knee bent at the angle of 90 degrees.

Expected answers:

{'S1':['left leg'], 'S2':['right arm'], 'S3':['left arm'], 'S4':['right leg'], 'S5':['left leg']}

Explanation:

The description is consist of 5 sentences. Therefore, the final answer should contain 5 keys.

Example 2:

This person is in a semi-lunge position throwing a bowling ball. The left leg is forward and slightly bent at the knee. The right leg is extended back with the toe of the right foot touching the ground. The torso is slightly leaned forward. The right arm is extended forward. The right elbow is slightly bent. The left arm is extended out to the side.

Expected answers:

{'S1':['left leg', 'right leg', 'left arm', 'right arm'], 'S2':['left leg'], 'S3':['right leg'], 'S4':['torso'], 'S5':['right arm'], 'S6':['right arm'], 'S7':['left arm']}

Explanation:

The semi-lunge position requires both legs. The throwing ball action uses both arms.

Figure 2. Prompt for body part level semantic label extraction with LLM.

References

[1] Better writing with deepl write. https://www.deepl.com/write.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, pages 346–362. Springer, 2022.

[5] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021.

You are a helpful assistant, helping me to extract body parts from the pose description. I will give you a pose description. You must label each sentence in the description with the related body joints. There are several main rules you need to follow:

Rule 1. Format of answer.

1.2 Your answer should follow the format:
{'S1':[related body joints], 'S2':[related body joints], ...} where each 'S' represent the sentence in sequence of the description.

Rule 2. Available body parts.

2.1 The available body joints labels are 'left shoulder', 'right shoulder', 'left elbow', 'right elbow', 'left wrist', 'right wrist', 'left hip', 'right hip', 'left knee', 'right knee', 'left ankle', 'right ankle', 'left foot', 'right foot', 'torso', 'spine', 'neck' and 'head'. You must only select body joints from the available body joints labels.

Rule 3. SMPL topology.

3.1 The human topology can be summarized as follows: The 'torso' represents the center of the human body. The 'spine' is above the 'torso' and connect 'torso' with 'neck', 'left shoulder' and 'right shoulder'. The 'left wrist', 'right wrist' represent the left hand and right hand. The 'left foot', 'right foot' represent the left toe and right toe which are connected to the 'left ankle' and 'right ankle'.

Example 1:

The individual is kicking with his left leg. His right arm is extended forward with the right elbow almost completely bent and the hand in front of his shoulder. His left arm is extended backwards with the elbow slightly bent. He is standing on his right leg with the right knee slightly bent. The left leg is on the air extending backward with the left knee bent at the angle of 90 degrees.

Expected answers:

{'S1':['left hip', 'left knee', 'left ankle', 'left foot'], 'S2':['right shoulder', 'right elbow', 'right wrist'], 'S3':['left shoulder', 'left elbow'], 'S4':['right hip', 'right knee'], 'S5':['left hip', 'left knee']}

Example 2:

This person is in a semi-lunge position throwing a bowling ball. The left leg is forward and slightly bent at the knee. The right leg is extended back with the toe of the right foot touching the ground. The torso is slightly leaned forward. The right arm is extended forward. The right elbow is slightly bent. The left arm is extended out to the side.

Expected answers:

{'S1':['left shoulder', 'right shoulder', 'left elbow', 'right elbow', 'left wrist', 'right wrist', 'left hip', 'right hip', 'left knee', 'right knee], 'S2':['left hip', 'left knee'], 'S3':['right hip', 'right knee', 'right ankle', 'right foot'], 'S4':['torso', 'spline'], 'S5':['right shoulder], 'S6':['right elbow'], 'S7':['left shoulder']}

Figure 3. Prompt for joint level semantic label extraction with LLM.

This person is in a semi-lunge position throwing a bowling ball. The right leg is forward and bent at the knee. The left leg is extended back with left knee bent at 90 degrees and the toe of the right foot touching the ground. The torso is slightly leaned forward. The left arm is straight and extended forward. The right arm is stretched back with the elbow hardly bent. The head is looking to the left hand.

He is standing upright with both feet shoulder width apart. Both elbows are bent at right angles and both forearms are crossed almost horizon-tally in front of the torso, with the left hand in front of the right. Both knees are almost bent.

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.

Someone is standing with one arm raised vertically above their head with the head facing forward. Their torso is vertical and their right elbow is straight. Their right hand is above their head. Their left elbow is slightly bent and their left hand is at the side of their left hip. Both thighs are straight and the left knee is separated from the right knee by shoulder width. The left knee is unbent as well as the right knee. The feet are about shoulder width apart.

The person is in a dance pose. He is standing on his left leg, with the other leg bent at the knee and the foot lifted off the ground and placed behind. Both arms are raised above the head and slightly bent at the elbows, forming a 'C' shape. The left elbow is behind the right. The left hand is spread out behind the right hand. The head is looking to the left.

The figure is leaning forward. Their torso is inclined forward. Both elbows are slightly bent and the forearms are vertical with hands on the sides of the body. Their left knee is approximately shoulder width apart from their right knee. Both knees are slightly bent, their feet are approximately shoulder width apart. The head is facing forward.

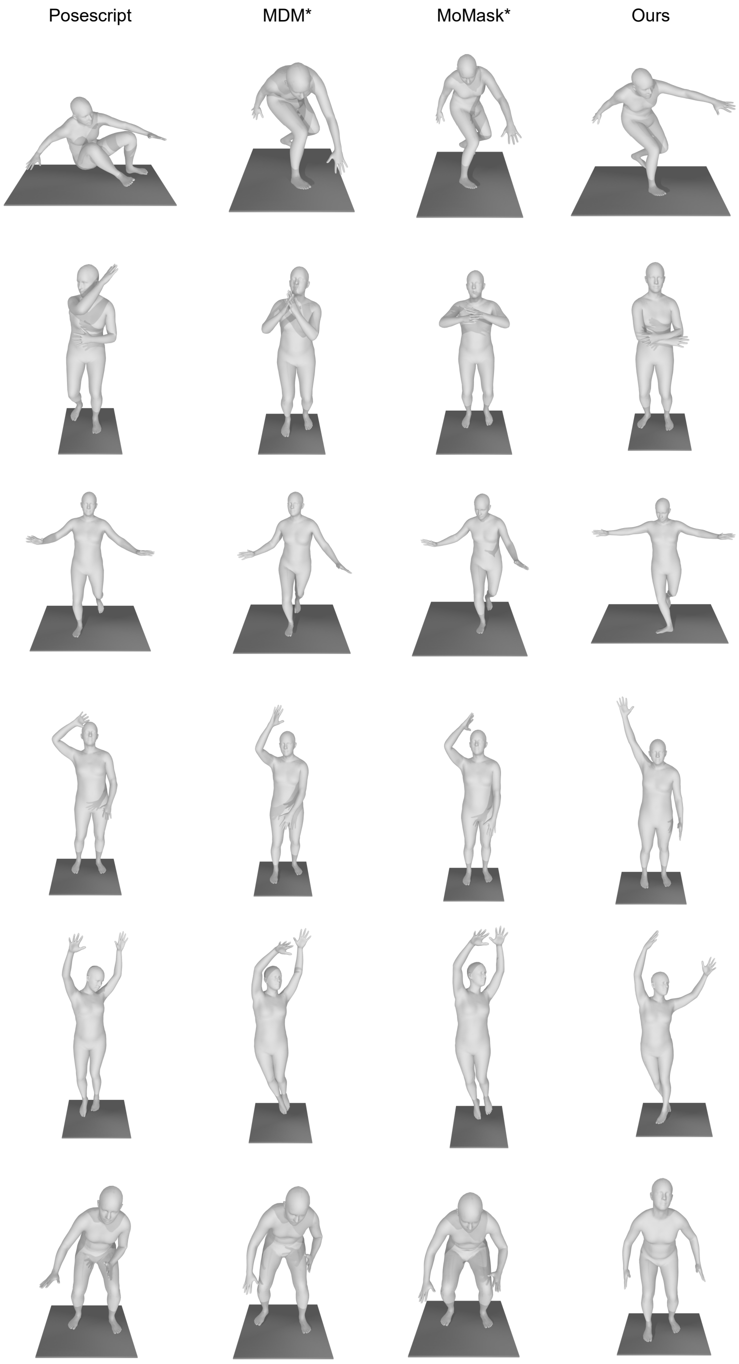


Figure 4. Additional quantitative results.