MoRAG - Multi-Fusion Retrieval Augmented Generation for Human Motion

Contents

1. Dataset

We chose HumanML3D [3] to evaluate our framework due to its extensive and diverse collection of motions paired with a wide range of text annotations. It provides annotations for the motions in the AMASS [6] and HumanAct12 [4] datasets. On average, each motion is annotated three times with different texts, and each annotation contains approximately 12 words. Overall, HumanML3D consists of 14,616 motions and 44,970 descriptions. The data is augmented by mirroring left and right. We follow the same splits as TMR [7] and ReMoDiffuse [10] to train the retrieval and generation models respectively.

2. Implementation Details

We use OpenAI's GPT-3.5-turbo-instruct for its efficiency in executing specific instructions and providing direct answers. Our prompting strategy allows a maximum of 256 tokens for both the prompt and the generation. The completions API, with default parameters, is used to generate the desired part-specific descriptions.

For part-specific retrieval models ($MoRAG_p$), we use AdamW [5] optimizer with a learning rate of 0.0001 and a batch size of 32. The latent dimensionality of the embeddings is 256. We set the temperature τ to 0.1, and the weight of the contrastive loss term λ_{NCE} to 0.1. Other hyperparameter values are used similarly to those in TMR [7].

For MoRAG-Diffuse, we use similar settings as that of ReMoDiffuse [10] used for HumanML3D [3]. For the diffusion model, the variances β_t are spread linearly from 0.0001 to 0.02, and the total number of diffusion steps is 1000. Adam optimizer with a learning rate of 0.0002 is used to train the model. MoRAG-Diffuse was trained on an NVIDIA GeForce RTX 2080 Ti, with a batch size of 64, using initial weights of ReMoDiffuse [10] for 50k steps.

3. Prompt strategy

Table 1 presents the LLM outputs for various text descriptions generated using our prompt function, demonstrating its effectiveness in motion retrieval. We observe that the generated descriptions are often correlated with other body parts, which led us to train part-specific retrieval models on full-body motion sequences.

3.1. Significance of "position"

To train independent part-specific retrieval models, $MoRAG_p$, for $p \in \{torso, hands, legs\}$, it is essential to obtain movement information specific to each part. However, since the framework includes a composition step that combines the retrieved motions into a single motion sequence, relying solely on movement information is insufficient. The composition also requires positional information. For example, as shown in `Videos/Position-Significance.mp4` for the text `text` = *”A person is swimming”*, explicitly prompting for the positional information allows the leg description to reflect its relative position to the ground, thereby retrieving the correct motion sample. The global orientation of the leg corresponding retrieved sample is important as it will determine the global orientation of the composed motion sequence.

3.2. Spatial Composition

`Videos/Spatial-Composition` contains videos that illustrate the composition workflow for combining part-specific motions, R_{part}, retrieved from corresponding

text	LLM Output			MoRAG part-specific retrieval		
	$\texttt{text}_{torso}$	$\texttt{text}_{hands}$	$\texttt{text}_{legs}$	R_{torso}	R_{hands}	R_{legs}
A person is standing and raising both hands	The person's torso is upright and still, while standing tall and balanced.	The person's hands are being lifted up towards the sky, using their arms to extend upwards.	The legs are steady and stable, acting as a strong foundation to support the body as the arms raise up.	a person uses their hands to clap (#002573)	a person raises his hands above his head. (#002315)	a person raises his hands above his head. (#002315)
A person is standing and raising single hand	The person's torso is upright and still, while their arm raises up.	One hand is lifted up from the side of the body, extended upwards and reaching toward the ceiling.	The legs are supporting the person's weight, standing firmly on the ground.	(person stands still and lifts right hand to face and mouth area#000813)	a person raises his right arm and then lowers it. (#001179)	a figure claps around shoulder height (#007523)
A person is standing on one leg and raising both hands	The person's upper body is straight and centered while standing on one leg.	Both hands are lifted above the head, reaching towards the sky.	One leg is holding the person's weight while the other is slightly lifted off the ground, balancing the body.	a person grabs their right foot and places it on their left thigh, and balances on one foot and then does the same with the other foot. (#006123)	raising and lowering arms. (#011583)	a person balances on their left leg and then their right. (#009917)
A person is standing on one leg and raising single hand	The person's body is upright and balanced on one leg, with the other leg lifted off the ground.	One hand is raised up in the air, reaching toward the ceiling or sky.	The standing leg is firmly planted on the ground, while the other leg is lifted up and may be slightly bent or straight.	the person raises their left foot up to their knee and then kicks their foot out, then returns their foot to their knee. (#004012)	a person raises his right arm and then lowers it. (#001179)	person balances on left leg then with arms up high has arms fully extended keeping balance on left leg (#004180)
A person is running with their arms crossed.	The person's upper body is straight and upright, while their chest and shoulders may be slightly forward as they run.	The hands are crossed over the chest, alternating in front of the body as the person runs.	The legs are moving back and forth in a rhythmic motion, propelling the person forward as they run.	a person jogs forward with arms moving to his side. (#002755)	a figure stands in place, crossing its arms (#003973)	a person slowly jogs to the left to right and the jogs back into place (#004273)
A person is running with their arms stretched out to the sides.	The person's torso is upright and slightly leaning forward as they jog, with their chest and shoulders in a relaxed position.	The person's hands are held out to the sides at shoulder level, moving rhythmically with each step.	The legs are moving in a bouncing motion, alternating between the left and right sides as the person jogs forward.	jogging forward in medium pace. (#000972)	the person is flying like a airplane. (#006433)	a person slowly jogs to the left to right and the jogs back into place (#004273)
A person is walking with their arms circling around.	As the person runs, their torso is upright and facing forward, with their chest and stomach comfortably relaxed.	The person is moving their arms back and forth in a swinging motion, with their arms stretched out to the sides.	The legs are taking turns lifting off the ground and propelling the person forward in a steady, rhythmic motion.	running from side to side. (#014305)	this person moves both arms out to his sides in a large circular motion then walks forward. (#005375)	a person doing a casual walk (#004183)
A person is walking with their arms crossed.	The person's upper body remains upright while they walk with their arms crossed, with their spine and shoulders in a stable position.	The person's hands are placed on opposite shoulders, crossing in front of their torso as they walk.	The legs are alternately lifting and stepping forward, propelling the person forward while they walk.	a person walks on uneven ground whilst holding on to handrail. (#001029)	a person crosses their arms in front of their chest, then drops them back at their sides. (#001014)	a person walks forward while holding out their arms for balance (#002087)

Table 1. **Prompt Examples :** We illustrate the LLM-generated part-specific outputs for text descriptions alongside their corresponding top-1 retrieval results to demonstrate the effectiveness of our prompt strategy. The HumanML3D [3] ID for the retrieved motions is indicated with the # symbol.

part-specific databases, $\mathcal{D}_{part}$, using the LLM-generated part-specific descriptions, $\texttt{text}_{part}$, where $part \in \{torso, hands, legs\}$.

3.3. Issue with left/right retrieval strategy

In MoRAG, we avoided the strategy of retrieving the left and right hands and legs separately, as their movements can become asynchronous when composed into a single motion sequence. This issue is illustrated in `Videos/Issue-with-left-and-right-Retrieval-Strategy`, which demonstrates asynchronous motion composition for the texts *"A person is clapping his hands"* and *"A person is taking large strides while walking."*

4. Qualitative Analysis

We present qualitative analysis across three key aspects: (1) Generalizability, (2) Diversity, and (3) Zero-shot performance, for both MoRAG (retrieval model) and MoRAG-Diffuse (generative model). We also provide baseline comparisons for generalizability and zero-shot capabilities of MoRAG against TMR++ [1], a state-of-the-art motion retrieval model, and ReMoDiffuse [10], which serves as the basis for MoRAG-Diffuse. Related videos can be found at `Videos/MoRAG` and `Videos/MoRAG-Diffuse`.

4.1. Generalizability

The incorporation of part-specific descriptions $\texttt{text}_{torso}$, $\texttt{text}_{hands}$ and $\texttt{text}_{legs}$, generated by LLMs enables MoRAG to construct motion samples that accurately capture subtle variations in text, allowing for finer detail representation. By utilizing these detailed descriptions, the model retrieves more precise and nuanced motions, reflecting minor differences in body part movements or positions within the input text. This results in more realistic and contextually appropriate motion sequences. Example videos can be found at `Videos/MoRAG/Generalizability`.

Conditioning these samples in the motion generation pipeline improves the model's comprehension of the language space, allowing it to effectively capture variations in text descriptions. This improvement is reflected in the Multi-Modal Distance metric. (Tab. 2) Example videos can be found at `Videos/MoRAG-Diffuse/Generalizability`.

4.2. Diversity

MoRAG constructs diverse set of motion samples for a given input text `text` by utilizing both, (1) LLMs' ability to generate diverse text descriptions for a prompt and (2) various combinations of retrieved part-specific motion samples. Examples of these diverse samples can be found at

Methods	R Precision ↑			FID ↓	MM Dist ↓	Diversity →	MultiModality ↑
	Top 1	Top 2	Top 3				
Real motions	$0.511^{\pm0.003}$	$0.703^{\pm0.003}$	$0.797^{\pm0.002}$	$0.002^{\pm0.000}$	$2.974^{\pm0.008}$	$9.503^{\pm0.065}$	-
MDM [8]	$0.320^{\pm0.005}$	$0.498^{\pm0.004}$	$0.611^{\pm0.007}$	$0.544^{\pm0.044}$	$5.566^{\pm0.027}$	$9.559^{\pm0.086}$	$2.799^{\pm0.72}$
MotionDiffuse [9]	$0.491^{\pm0.001}$	$0.681^{\pm0.001}$	$0.782^{\pm0.001}$	$0.630^{\pm0.001}$	$3.113^{\pm0.001}$	$9.410^{\pm0.049}$	$1.553^{\pm0.042}$
MLD [2]	$0.481^{\pm0.003}$	$0.673^{\pm0.003}$	$0.772^{\pm0.002}$	$0.473^{\pm0.013}$	$3.196^{\pm0.010}$	$9.724^{\pm0.082}$	$2.413^{\pm0.079}$
ReMoDiffuse [10]	$0.510^{\pm0.005}$	$0.698^{\pm0.006}$	$0.795^{\pm0.004}$	$0.103^{\pm0.004}$	$2.974^{\pm0.016}$	$9.018^{\pm0.075}$	$1.795^{\pm0.043}$
FineMoGen [11]	$0.504^{\pm0.002}$	$0.690^{\pm0.002}$	$0.784^{\pm0.002}$	$0.151^{\pm0.008}$	$2.998^{\pm0.008}$	$9.263^{\pm0.094}$	$2.696^{\pm0.079}$
MoRAG-Diffuse	$0.511^{\pm0.003}$	$0.699^{\pm0.003}$	$0.792^{\pm0.002}$	$0.270^{\pm0.010}$	$2.950^{\pm0.012}$	$9.536^{\pm0.104}$	$2.773^{\pm0.114}$

Table 2. **Quantitative Results:** We compare the results of text-to-motion generation between ours and the state-of-the-art diffusion based methods on HumanML3D [3] dataset. Our method achieves better semantic relevance, diversity, and multimodality performances. Indicate best results , indicates second best results.

`Videos/MoRAG/Diversity`.

The diverse samples produced by MoRAG improve the diversity of MoRAG-Diffuse, as indicated by the Diversity metric in Tab. 2. Related videos can be found at `Videos/MoRAG-Diffuse/Diversity`.

4.3. Zero-Shot Performance

Our approach facilitates the construction of motion sequences for unseen text phrases (zero-shot). This capability arises from two key aspects of the MoRAG framework: (1) it utilizes LLM-generated descriptions rather than the input text directly, and (2) it employs part-wise spatial composition. While the input text may be novel, the LLM-generated part-specific descriptions are not entirely unknown; relevant samples exist where the desired action is performed by specific body parts. Our method retrieves such samples independently for each body part and integrates them to construct motion sequences for unseen text descriptions. Example videos can be found at `Videos/MoRAG/Zero-shot`.

Conditioning the samples constructed by MoRAG facilitates the generation of motion sequences for previously unseen text phrases in MoRAG-Diffuse. Example videos demonstrating this capability are available at `Videos/MoRAG-Diffuse/Zero-shot`.

5. Metrics

For quantitative evaluations, we adopt the performance metrics used in ReMoDiffuse [10], which include R Precision, Frechet Inception Distance (FID), Multi-Modal Distance, Diversity and Multimodality. For R Precision and MultiModality, higher scores indicate superior performance. Conversely, lower scores are preferred for FID and Multi-Modal Distance. For Diversity, performance improves as the score more closely aligns with the real motions.

(1) R Precision evaluates how well the generated motion sequences semantically match the text descriptions. We calculate Top-1, Top-2, and Top-3 accuracy by measuring the Euclidean distance between each motion sequence and 32 text descriptions (one ground truth and 31 randomly selected descriptions). (2) FID calculates the distance between features extracted from real and generated motion sequences. (3) Multi-modal distance (MM Dist for short) measures the average Euclidean distance between the feature vectors of text and generated motions. (4) Diversity measures the variability and richness of the generated sequences. Average euclidean distance is calculated between the two-equal sized subsets which are randomly sampled from the generated motions from all test texts. (5) Multimodality measures the diversity of generated motions for a given text. We generate 10 pairs of motions for each text and compute average distance between the each pair feature vectors.

5.1. Challenges in applying motion retrieval metrics to spatially composed sequences

To evaluate text-to-motion retrieval methods such as TMR [7] and TMR++ [1], the similarity between the text corresponding to the retrieved motion sample and the input text is computed. However, for our approach, which involves the spatial composition of motion sequences, there is no single corresponding text for the entire composed sequence. As a result, these metrics cannot be computed in our case.

6. Code

The details of our MoRAG python implementation can be found in the folder `code`.

- `utils.py`: Consists util functions which will be used in MoRAG.
- `prompt.py`: Consists prompt function and openAI call to generate part-specific descriptions for given text.
- `morag.py`: Constructs part-wise fused motion sequences for a given text using the generated descrip-

tions from `prompt.py`

- `morag-diffuse.py`: Conditions the diffusion model using the composed motion sequences constructed using `morag.py`. Here we provided only RetrievalDatabase code, not the complete motion generation code.

References

[1] Léore Bensabath, Mathis Petrovich, and Gül Varol. Tmr++: A cross-dataset study for text-based 3d human motion retrieval. 2024. 2, 3

[2] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 3

[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. 1, 2, 3

[4] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 1

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 1

[6] 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, Oct. 2019. 1

[7] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 3

[8] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[9] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3

[10] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 364–373, October 2023. 1, 2, 3

[11] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *NeurIPS*, 2023. 3