MoRAG - Multi-Fusion Retrieval Augmented Generation for Human Motion

Sai Shashank Kalakonda Shubh Maheshwari CVIT, IIIT Hyderabad University of California San Diego sai.shashank@research.iiit.ac.in shmaheshwari@ucsd.edu Ravi Kiran Sarvadevabhatla CVIT, IIIT Hyderabad ravi.kiran@iiit.ac.in A person is balancing on A person is one foot with the other skipping in a circle leg extended behind. a) MoRAG-Diffuse (Generation) A person is A person performs doing moonwalk with yoga sun salutation arms raised b) MoRAG (Retrieval)

Figure 1. **MoRAG** is a retrieval-augmented framework for generating human motion from text. It integrates part-specific motion retrieval models with large language models to improve the quality of generation and retrieval tasks across various text descriptions. The black arrow illustrates motion translation. In the bottom figures, red, blue, and green represent the retrieved motion for the **hands**, **torso**, and **legs**. The varying transparency in the figure indicates the progression of time steps.

Abstract

We introduce **MoRAG**, a novel multi-part fusion based retrieval-augmented generation strategy for text-based human motion generation. The method enhances motion diffusion models by leveraging additional knowledge obtained through an improved motion retrieval process. By effectively prompting large language models (LLMs), we address spelling errors and rephrasing issues in motion retrieval. Our approach utilizes a multi-part retrieval strategy to improve the generalizability of motion retrieval across the language space. We create diverse samples through the spatial composition of the retrieved motions. Furthermore, by utilizing low-level, part-specific motion information, we can construct motion samples for unseen text descriptions. Our experiments demonstrate that our framework can serve as a plug-and-play module, improving the performance of motion diffusion models. Code, pretrained models and sample videos will be made available at: https://motionrag.github.io/

1. Introduction

Text-driven human motion generation has seen unprecedented growth in recent years [16, 30, 36–38, 40]. Numerous works have been proposed for this task, ranging from

Motion Retrieval Method	Te	xt Robustness	Ge	eneralizability	Diversity	Zero-shot setting
TMR [23]		×		x	X	×
TMR++ [5]		1		x	X	X
Ours		1		1	 ✓ 	1

Table 1. Comparision of text-to-motion retrieval approaches

encoder-decoder style architectures [2, 11, 12] to the recent emerging trend of diffusion-based models [7, 36, 38], which generate fine-grained, realistic motion sequences. While they can generate high-quality motion sequences for simple or familiar text descriptions similar to those in the training set, they perform poorly with complex or unseen text descriptions.

Retrieval-augmented Generation (RAG) has gained significant attention in recent years for its potential to enhance generative models by incorporating additional information through retrieval methods [9, 39]. By integrating retrievalbased techniques with generative models, RAG produces outputs that are more accurate, contextually relevant, and reliable. Moreover, this additional information helps enhance the model's generalizability across language space and also improves the stochasticity. However, the application of RAG in motion generation is underexplored.

A RAG system typically comprises two key components: the retriever and the generator. The retriever identifies relevant information from a database based on the input query, while the generator uses both the input query and the retrieved information to generate the desired content.

Recently proposed text-to-motion retrieval approaches [23, 33] aim to retrieve full-body motion sequences from the motion database using a contrastive training strategy between text and motion embeddings. However, these retrieval strategies do not perform well when text phrases contain spelling errors, rephrased text sequences, or substitution of synonymous words. (See Fig. 5)

Action-GPT [16], TMR++ [5] prompt large-language models (LLMs) to provide detailed descriptions as input. However, these approaches are limited in their ability to generate or retrieve motion sequences for text descriptions that are not present in the database, restricting the diversity of output motion sequences and reducing generalization to out-of-domain or unseen text descriptions (Fig. 6 (a))

Based on the RAG concept, ReMoDiffuse [37], adopts a hybrid retrieval approach using motion length and CLIP [26]-based text-to-text similarity, which does not incorporate any motion-specific information, which can result in inaccurate retrievals (Fig. 2) and motion generation. (Fig. 6 (b))

To overcome these shortcomings, we propose a multipart fusion-based augmented motion retrieval strategy that is capable of constructing diverse and reliable motion sequences. We train part-specific independent motion retrieval models that retrieve motion sequences with movements corresponding to each part aligned with the provided text description. The retrieved part-specific motion sequences are fused accordingly to construct full body motion sequences, allowing our method to even query unseen text descriptions.

Our experiments show that incorporating the constructed motion sequences as an additional conditioning to the diffusion based motion generation model improves the alignment with the semantic information of the text description and diversity of generated sequences.



Figure 2. MoRAG utilizes part-specific descriptions to effectively retrieve relevant samples, demonstrating robustness to variations in motion length and descriptive text. In contrast, ReMoDiffuse [37], a hybrid approach based on motion length and text similarity, fails to retrieve suitable samples when there are changes in motion length or text. Each figure of ReMoDiffuse displays the retrieved text at the top and the corresponding motion length in brackets. For MoRAG, three part-specific retrieved texts, along with their corresponding HumanML3D [12] ID, are provided using the #. tick and cross to indicate whether the motion corresponds to the input text.

In summary, our contributions are as follows:

- We propose MoRAG, a novel multi-part fusion-based retrieval augmented human motion generation framework to enhance the performance of the diffusion based motion generation model.
- We adapt *part-wise motion retrieval* approach that utilizes *generative prompts* to construct motion sequences that align with the provided text description.
- Motion sequences constructed using our retrieval strategy exhibit superior generalization and diversity, as shown by the qualitative and quantitative analysis.

2. Related Works

Text-conditioned human motion generation The early research efforts concentrated on encoder-decoder models with multimodal joint embedding space spanning both text and motion domains [1, 2]. Text2Action [1] proposed a



Figure 3. **MoRAG Overview**: Given a text description text, we generate part-specific descriptions corresponding to "Torso," "Hands," and "Legs" by prompting an LLM. These generated descriptions are used as queries to retrieve corresponding part-specific motions: R_{torso}^{i} , R_{hands}^{i} , and R_{legs}^{i} from the motion databases D_{torso} , D_{hands} , and D_{legs} , respectively. The retrieved motions are then fused to construct a full-body motion sequence C^{i} that aligns with the input text. The constructed motion samples are used as additional information in the motion generation pipeline during both training and inference, alongside the input text, to further improve model performance.

GAN based generative model constituting RNN based text encoder and action decoder. JL2P [2] focused on learning a joint embedding space for language and pose on the motion reconstruction task using text and motion embedding. Ghosh *et al.* [11] further improved the joint embedding space using a hierarchical two-stream model where two motion representations are learned, one for each lower body and upper body.

To enhance the text-to-motion generalizability, Motion-CLIP [29] incorporated image embedding of the poses into the training paradigm alongside text embedding. Both the image and text embedding are generated using CLIP. [26]. TEMOS [22] proposed a transformer based VAE encoding approach for text and motions to improve the diversity of generated motion sequences. TEACH [3] extends TEMOS [22] with an additional past motion encoder to generate long motion sequences using the text description and previous sequence. T2M-GPT [35] transforms the challenge of textconditioned motion generation into a next-index prediction task by encoding motion sequences as discrete tokens and utilizing a transformer to predict future tokens.

Unlike the above mentioned approaches, we propose a novel multi-part fusion-based augmented motion retrieval strategy to improve the diversity of retrieved motions and to enhance their generalizability for unseen text descriptions.

Motion Diffusion Models With recent advancements in diffusion models for text and image domain tasks, several works have been proposed in the area of text-to-motion generation. MotionDiffuse [36] incorporated efficient DDPM in motion generation tasks to generate diverse, variable-length, fine-grained motions. MDM [30] is a lightweight diffusion model that utilizes a transformer-encoder backbone. Instead of predicting noise, it predicts motion samples, allowing geometric losses to be used as training con-

straints. MLD [7] adapted the diffusion process on latent motion space instead of using raw motion sequences, generating better motions at a reduced computational overhead.

However, these diffusion-based models struggle to generalize across the language space, especially when dealing with complex or unusual text descriptions. Recent textto-image generation works introduced retrieval-augmented pipelines in their frameworks [6] to address such issues. ReMoDiffuse [37] extended MotionDiffuse [36] by integrating a hybrid retrieval mechanism to refine the denoising process. We improve the generalizability and diversity of ReMoDiffuse by the inclusion of large language models (LLMs) and the integration of part-specific retrieval.

Text-to-Motion retrieval Recently, significant progress has been made by multi-modal retrieval based systems in the field of text-to-image [18, 25, 26, 34] and text-to-video [8, 10]. However, it has been underexplored in the field of text-to-motion due to the lack of large and diverse annotated motion capture datasets. Recent works such as BA-BEL [24] and HumanML3D [12] have provided detailed description annotations for the large-scale motion capture collection AMASS [21]. Following, a few works have been proposed in the field of text-to-motion retrieval.

Initially, motion generation works [12] used retrieval as a performance metric for evaluation purposes. TMR [23] is the first work to showcase text-to-motion retrieval as a standalone task. To query motions, TMR [23] adopted the idea of contrastive learning from CLIP [26] and extended text-to-motion generation model TEMOS [22]. Although TMR demonstrated impressive results, there is a large scope for improvement in the generalizability of the model over language space. LAVIMO [33] integrated human-centric videos as an additional modality in the task of text-tomotion retrieval to effectively bridge the gap between text and motion. TMR++ [5] extended TMR by leveraging LLMs in the motion retrieval pipeline via label augmentations to increase the robustness and generalizability. However, a significant gap remains in utilizing these existing retrieval strategies for retrieval-based human motion generation approaches due to their lack of diversity and generalizability for complex or unseen text descriptions.

3. Proposed Method

Fig. 3 illustrates our multi-fusion retrieval-augmented human motion generation framework, MoRAG, which aims to enhance the performance of diffusion based motion generation model by leveraging additional motion information constructed using part-specific motion retrieval. Given an input text description text, we generate Ndiverse, semantically coherent human motion sequences $\{H^1, \ldots, H^n, \ldots, H^N\}$. We prompt the input text description to an LLM to generate motion descriptions specific to the "Torso," "Hands," and "Legs" (Sec.3.2). These descriptions are used for the retrieval of part-specific full-body motion sequences from pre-computed part-specific motion databases (Sec.3.3). The retrieved motion sequences are then fused to construct full-body motion sequences, which serve as additional knowledge for the diffusion model (Sec.3.4). This methodology enhances the model's ability to effectively handle both typical and complex/unseen input conditions (Sec.3.5).

3.1. Augmented Motion Retrieval Strategy

The key component of the MoRAG framework is the retrieval of diverse and semantically aligned motion samples from the database based on the given input text query. Existing text-to-motion retrieval methods [5, 23, 33] typically retrieve full-body motion samples directly from the database. However, these approaches overlook the fact that actions are frequently characterized by localized dynamics, often involving only small subsets of joint groups, such as the hands (e.g., 'eating') or legs (e.g., 'sitting'). [14, 31] This results in two significant issues: (i) limited generalizability and (ii) lack of diversity in the retrieved samples. (See Table 1 and Fig. 6 (a)) This is due to the limited availability of text-motion annotated datasets. Although BABEL [24] and HumanML3D [12] provide detailed text annotations for the large-scale motion capture collection AMASS [21], they are still insufficient to generalize across the language space. However, the AMASS dataset contains extensive low-level body parts information that holds the potential to generalize across a significantly broader language space.

Based on this observation, we design independent partspecific motion retrieval models that can retrieve full-body motion sequences with movements corresponding to specific parts aligned with the provided text description. This enables dedicated part descriptions for retrieval of actions involving specific part movement. By composing these motion samples, we can construct full-body motion sequences that are semantically coherent with the given text input. The composition also improves the expressivity of motions, since fine-grained motion details are often expressed in text in terms of body parts. The wide variety of composing combinations provides huge diversity in the constructed motion samples. Integrating these samples into a motion generation pipeline as additional information can enhance the model's performance. We also observe better generations for unseen text descriptions. (See Fig. 6 (b))

The objective is to construct a series of motion sequences $\{C^1, \ldots, C^i, \ldots, C^k\}$ from the motion database ranked from 1 to k where each motion sequence C^i is represented as a sequence of human poses $\{C_1^i, \ldots, C_t^i, \ldots, C_{f_i}^i\}$ with f_i representing the number of timesteps for motion C^i . The motion retrieval strategy in MoRAG comprises three steps: (1) generating part-specific body movement descriptions, (2) retrieving part-specific motion sequences, and (3) composing the retrieved motion sequences. Details of each are provided in the following section.

3.2. Generation of part-specific descriptions

Given the text description text, we generate part specific body movement descriptions using an LLM as a knowledge engine. We construct a suitable prompt text_{prompt}, for text using a prompt function f_{prompt} , comprising of three components:

(i) Task instructions, to specify the details of our task:

"The instructions for this task is to describe the listed body parts' position and movements in a sentence using simple language. ['Torso',' Hands', 'Legs']"

(ii) **Few-shot examples**, provides a set of examples consisting of diverse action descriptions to determine the format of the output we are expecting.

(iii) **Query**, to incorporate the input text text to generate part-specific body movement and orientation information.

"Query: Describe the below body parts position and movements involved in the action [text] in a sentence using simple language. 1) Torso 2) Hands 3) Legs".

Specifically prompting for the position provides the global orientation of the body parts which results in better retrieval.

The constructed prompt is then passed through LLM to generate descriptions of the positions and movements for the specified body parts denoted as $text_{torso}$, $text_{hands}$ and $text_{legs}$. Training a retrieval model on these body part descriptions enables the retrieval of motions for rephrased and spell-error text phrases, thereby having a better generalization over language space. (Fig.5)



Figure 4. **MoRAG Training**: Our objective is to construct three independent part-specific motion databases. The training paradigm includes three motion retrieval models: $MoRAG_{torso}$, $MoRAG_{hands}$, and $MoRAG_{legs}$, each corresponding to a specific body part. We train these three models independently using part-specific body movement descriptions generated by LLMs for text phrases $text_i$ and their corresponding full-body motion sequences $motion_i$. We adopt a contrastive training objective between part-specific text embeddings $(Z_{p,i}^T)$ generated by text encoders (T_p^{Enc}) and motion embeddings $(Z_{p,i}^M)$ generated by the corresponding part-specific motion encoder (M_p^{Enc}) . The diagonal elements, representing positive pairs (green), are maximized, while the off-diagonal elements, representing negative pairs with text similarity below a threshold (red), are minimized. For simplicity, we do not visualize the motion decoder, but we follow a similar training procedure as described in [23].

3.3. Multi-part motion retrieval

As shown in Fig. 4, MoRAG uses 3 independently trained TMR [23] models, $MoRAG = \{MoRAG_{torso}, MoRAG_{hands}, MoRAG_{legs}\}$ corresponding to the respective body part. We do not train separate left and right body parts models to avoid asynchronous movements in the composed motion. For a part $p \in \{torso, hands, legs\}$, the retrieval model $MoRAG_p = \{T_p^{Enc}, M_p^{Enc}, M_p^{Dec}, \mathcal{D}_p^{Enc}\}$ consists of a text encoder, motion encoder, motion decoder, and motion database respectively. The model architecture for the encoder and decoder are based on TEMOS [22].

Text Encoders (T_p^{Enc}) : The LLM-generated partspecific motion descriptions of the text sequence text are first passed through a pre-trained and frozen DistilBERT [27] to generate features \mathcal{F}_p^T for each part-wise text description text_p. Along with the features \mathcal{F}_p^T , two learnable distribution tokens are passed as input to the text encoders. The outputs corresponding to the distribution tokens passed are considered as Gaussian distribution parameters $(\mu_p^T \text{ and } \sigma_p^T)$ from which the latent vector Z_p^T is sampled using reparametrization trick [17].

$$(\mu_p^T, \sigma_p^T) = T_p^{Enc}(\mathcal{F}_p^T) \tag{1}$$

$$Z_p^T \sim \mathcal{N}(\mu^T, \sigma^T) \tag{2}$$

Motion Encoders (M_p^{Enc}) : Similarly, Z_p^M is obtained from the motion encoders by inputting the corresponding full-body motion sequence $M_{1:f}$ associated with the text description text and duration f.

$$(\mu_p^M, \sigma_p^M) = M_p^{Enc}(M_{1:f})$$
 (3)

$$Z_p^M \sim \mathcal{N}(\mu^M, \sigma^M) \tag{4}$$

During retrieval, instead of sampling, we directly use the embedding corresponding to the mean parameter (i.e $Z_p^T = \mu_p^T$ and $Z_p^M = \mu_p^M$)

To enhance the effectiveness of motion retrieval, we train all three motion encoders (M_p^{Enc}) using full-body motion sequences instead of just the respective body parts. This approach is based on the observation that LLM-generated part-specific descriptions contain information about the queried body part about other body parts. Utilizing fullbody motion sequences allows us to leverage intra-joint information, resulting in more coherent and semantically accurate motion retrieval.

Motion Decoders (M_p^{Dec}) : Motion decoders input a latent vector Z and sinusoidal positional encoding of the duration f and output a full body motion sequence $\hat{M}_{1:f}$ non-autoregressively. The input latent vector Z is obtained from one of the two encoders during training. However, since our task is motion retrieval, the decoder is not used during inference.

$$\hat{M}_{1:f} = M_p^{Dec}(Z) \tag{5}$$

$$Z \in \{Z_n^T, Z_n^M\} \tag{6}$$

Loss : Each retrieval model, $MoRAG_p$ is trained with the loss \mathcal{L}^p [23]:

$$\mathcal{L}^{p} = \mathcal{L}_{R} + \lambda_{KL} \mathcal{L}_{KL} + \lambda_{E} \mathcal{L}_{E} + \lambda_{NCE} \mathcal{L}_{NCE}$$
(7)

 \mathcal{L}_R is the motion reconstruction loss given motion and text embeddings to the decoder. \mathcal{L}_{KL} is the Kullback-Leibler(KL) divergence loss composed of four losses. The first two are for the text and motion distributions with normal distribution and the other two are between text and motion distributions. \mathcal{L}_E is the cross-modal embedding similarity loss between both text and motion latent embeddings Z_p^T and Z_p^M .



Figure 5. **LLM Importance**: Incorporating part-wise descriptions generated by LLMs into text-to-motion retrieval improves generalization over the language space. (a) **Spell Error** - MoRAG successfully retrieves and constructs the correct motion sequence when 'sit-ups' is replaced with 'situps', unlike TMR [23]. (b) **Rephrasing** - MoRAG effectively retrieves the correct motion sequence even when the voice is changed from active to passive. (c) **Substitution** - MoRAG accurately retrieves the correct motion sequence when 'chest' is replaced with its synonym 'heart'.

 \mathcal{L}_{NCE} is the contrastive loss which is based on InfoNCE [32] formulation, used to better structure the crossmodel latent space. The text and its corresponding motion embedding are considered positive pairs $(Z_{p,i}^T \text{ and } Z_{p,i}^M)$, whereas all other combinations are considered to be negative $(Z_{p,i}^T \text{ and } Z_{p,j}^M)$ with $i \neq j$. Similarity matrix Scomputes the pairwise cosine similarities for all the pairs, $S_{ij} = \cos(Z_{p,i}^T, Z_{p,j}^M)$. However, not all negative pairs are involved in the loss computation. Text-motion pairs with text description similarities above a certain threshold, referred to as 'wrong negatives', are filtered out from the loss computation. The threshold to filter negatives is set to 0.8. These text similarities are computed using MPNet [28].

$$\mathcal{L}_{NCE} = -\frac{1}{2N} \sum_{i} \left(\log \frac{e^{S_{ii}/\tau}}{\sum_{j} e^{S_{ij}/\tau}} + \log \frac{e^{S_{ii}/\tau}}{\sum_{j} e^{S_{ji}/\tau}} \right)$$
(8)

Motion database (\mathcal{D}_p) : Post training, we create a database consisting of three key-value tables for every body part where each key is a unique identifier for a motion sample from the AMASS [21] database. The corresponding value is a vector inferred from the motion encoder. During retrieval, the LLM-generated part description is encoded into a query vector for every text encoder, T_p^{Enc} . We use this query vector to search the corresponding vector indexes, finding the k-nearest neighbors in the embedding space using cosine similarity. The corresponding k full-body motion sequences $\{R_p^1, \ldots, R_p^i, \ldots, R_p^k\}$ are retrieved for each body part p.

3.4. Spatial motion composition

The retrieved motion sequences $\{R_p^1, \ldots, R_p^i, \ldots, R_p^k\}$ are composed such that the i_{th} sequence corresponding to each part p is used to construct the i_{th} full-body motion sequence C^i . This results in k full-body motion sequences $\{C^1, \ldots, C^i, \ldots, C^k\}$, which are used as additional guidance for the motion diffusion model. We followed a rankby-rank combination approach to generate these top-k sequences. However, alternative combination methods could be employed to create a significantly larger number of sequences. Our composition approach is similar to SINC [4] but we do not require the use of an LLM for mapping joints from the retrieved sequences.

To construct the composed motion sequence C^i using the corresponding retrieved full-body motion sequences, $\{R^i_{torso}, R^i_{hands}, R^i_{legs}\}$, we follow these steps: (1) Trimming all three retrieved sequences to the length of the shortest one; $f_{min} = \min_p(f_p)$, (2) Selecting the respective body part's joint information from the corresponding retrieved motions. We follow the SMPL [19] skeleton structure with the first 22 joints and partition it into three disjoint sets of joints: $J = \{j_{torso} \cup j_{hands} \cup j_{legs}\}$.

$$C_{f}^{i}[j_{p}] = R_{p,f}^{i}[j_{p}]$$

$$p \in \{torso, hands, legs\}, f \in [1, f_{min}]$$

$$(9)$$

(3) Choosing the global orientation and translation from R_{legs}^{i} , as leg motion is closely associated with changes in global translation and orientation.

3.5. MoRAG-Diffuse

For generation, we extend ReMoDiffuse [37], a diffusion-based model, by incorporating our retrieval mech-



Figure 6. **Qualitative Results**: Comparison of motion retrieval and generation using our multi-part fusion approach: retrieval is compared with TMR++ [5], a state-of-the-art motion retrieval method and generation is compared with ReMoDiffuse [37]. **Top**: Our method demonstrates superior generalization capabilities. **Middle**: Our approach generates accurate motion sequences for unseen text descriptions. **Bottom**: Our setup exhibits increased diversity.

Methods		R Precision \uparrow		$FID\downarrow$	MM Dist \downarrow	$ $ Diversity \rightarrow	MultiModality \uparrow
	Top 1	Top 2	Top 3	[
Real motions	$\mid 0.511^{\pm 0.003}$	$0.703^{\pm 0.003}$	$0.797^{\pm 0.002}$	$0.002^{\pm 0.000}$	$2.974^{\pm 0.008}$	$9.503^{\pm 0.065}$	-
MDM [30]	$0.320^{\pm 0.005}$	$0.498^{\pm 0.004}$	$0.611^{\pm 0.007}$	$0.544^{\pm 0.044}$	$5.566^{\pm0.027}$	$9.559^{\pm0.086}$	$2.799^{\pm 0.72}$
MotionDiffuse [36]	$0.491^{\pm 0.001}$	$0.681^{\pm 0.001}$	$0.782^{\pm 0.001}$	$0.630^{\pm 0.001}$	$3.113^{\pm0.001}$	$9.410^{\pm0.049}$	$1.553^{\pm 0.042}$
MLD [7]	$0.481^{\pm 0.003}$	$0.673^{\pm 0.003}$	$0.772^{\pm 0.002}$	$0.473^{\pm 0.013}$	$3.196^{\pm0.010}$	$9.724^{\pm0.082}$	$2.413^{\pm 0.079}$
ReMoDiffuse [37]	$0.510^{\pm 0.005}$	$0.698^{\pm 0.006}$	$0.795^{\pm 0.004}$	$0.103^{\pm 0.004}$	$2.974^{\pm 0.016}$	$9.018^{\pm0.075}$	$1.795^{\pm 0.043}$
FineMoGen [38]	$0.504^{\pm 0.002}$	$0.690^{\pm 0.002}$	$0.784^{\pm0.002}$	$0.151^{\pm 0.008}$	$2.998^{\pm0.008}$	$9.263^{\pm0.094}$	$2.696^{\pm 0.079}$
MoRAG-Diffuse	$0.511^{\pm 0.003}$	$0.699^{\pm 0.003}$	$0.792^{\pm 0.002}$	$0.270^{\pm 0.010}$	$2.950^{\pm 0.012}$	$9.536^{\pm0.104}$	$2.773^{\pm 0.114}$

Table 2. Quantitative Results: We compare the results of text-to-motion generation between ours and the state-of-the-art diffusionbased methods on HumanML3D [12] dataset. Our method achieves better semantic relevance, diversity, and multimodality performances.Indicate best results, indicates second best results.

anism within the motion generation pipeline. Unlike Re-MoDiffuse [37], we adapt our multi-part composed motion, rather than their motion length and text based similarity retrieval approach.

For the top-k retrieved motion sequence C^i , we follow the 263 dimension motion representation as in [12], where each human pose C_t^i is represented by $(\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{c}^f)$, where $\dot{r}^a \in \mathbb{R}$ is root angular velocity along Y-axis; $\dot{r}^x \in \mathbb{R}, \dot{r}^z \in \mathbb{R}$ are global root velocities in X-Z plane; $\dot{r}^y \in \mathbb{R}$ is root height; $\mathbf{j}^p \in \mathbb{R}^{3 \times n(J)}$,

 $\mathbf{j}^v \in \mathbb{R}^{3 \times n(J)}$, $\mathbf{j}^r \in \mathbb{R}^{6 \times n(J)}$ are the local pose positions, velocity and rotation respectively. $\mathbf{c}^f \in \mathbb{R}^4$ is the foot contact features calculated by the heel and toe joint velocity. n(J) is the number of joints and T_i represents the number of timesteps for motion C^i .

To effectively utilize information from retrieved motion samples, we use the Semantics-Modulated Transformer (SMT) introduced in ReMoDiffuse [37]. It comprises of N identical decoder layers, each featuring a Semantics-Modulated Attention (SMA) layer and a feed-forward network (FFN) layer. The SMA layer integrates information from the input description and the retrieved samples, refining the noised motion sequence throughout the denoising process. The SMA layer consists of a cross-attention mechanism where the noised motion sequence serves as the query vector Q. The key vector K and value vector V are derived from three sources of data: (1) the noised motion sequence itself; (2) CLIP's text features of the input description text which are further processed by two learnable transformer encoder layers; and (3) the retrieved motion and text features R^m , R^t , extracted using transformer-based encoders from the retrieved samples. As our composed motion samples do not have associated text sequences for generating text features R^t , we use the features of the input text description text.

4. Experiments & Results

First, we describe the dataset (Sec. 4.1) and implementation details (Sec.4.2) used in our experiments. Following that, we analyze the effectiveness of the MoRAG framework, comparing it to previous research works by providing an analysis of both retrieval and generation results. (Sec.4.3).

4.1. Dataset

We chose HumanML3D [12] 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 [21] and Human-Act12 [13] 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 [23] and ReMoDiffuse [37] to train the retrieval and generation models respectively.

4.2. Implementation Details

We use OpenAI's GPT-3.5-turbo-instruct as the large language model (LLM) due to its efficiency in understanding and executing specific instructions and its ability to provide direct answers to questions. The proposed prompting strategy consumes a maximum of 256 tokens together for prompt and generation. We use the completions API endpoint with the default parameters to generate the desired part-specific descriptions.

For part-specific retrieval models $(MoRAG_p)$, we use AdamW [20] 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 [23]. For MoRAG-Diffuse, we use similar settings as that of ReMoDiffuse [37] used for HumanML3D [12]. 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 [37] for 50k steps.

4.3. Results

Fig. 6 presents qualitative comparisons of MoRAG for both retrieval and generation tasks. We compare our retrieval results with TMR++ [5], a state-of-the-art motion retrieval model, and our generation results with ReMoDiffuse [37], focusing on generalizability, zero-shot performance, and diversity.

We observe that MoRAG constructed samples, utilizing part-specific retrieved motions, exhibit superior generalizability across the language space and effectively adapt to low-level changes. The richer and dedicated part-specific descriptions generated by the LLM helped retrieve precise part-specific motion sequences corresponding to the input text. Multi-part fusion has improved the construction of motion samples for unseen text descriptions, achieving better semantic alignment with the input text. Current motion retrieval methods treat the input skeleton in a monolithic manner, processing all joints in the pose tree as a whole. However, these approaches overlook the significance of subparts, which could enhance generalizability for unseen text descriptions. By dividing each action into sub-actions corresponding to specific subsets of joints, retrieval from the database can be made more effective. For example, in Fig. 6 (a), the text phrase, "A person is eating while seated on the ground", doesn't exist in the database which leads to the retrieval of the closest matched sample by TMR++ [5], "this person is sitting on the floor and reaches to his head with his right arm.". However, MoRAG searches for the partspecific sub-actions "eating" and "sitting on the ground," which can be easily retrieved from the database. When composed, these sub-actions construct a relevant motion sequence. Additionally, the extensive range of composition combinations contributes to significant diversity in the constructed sequences.

Similarly, MoRAG-Diffuse demonstrates improved generalization to the language space and generation of motion sequences for unseen text conditions, leveraging low-level information captured through retrieval conditioning.

Table. 2 summarizes the results of using our framework in comparison with the existing diffusion-based motion generation models. Incorporating part-specific motion retrieval models as additional knowledge in the motion generation pipeline shows an improvement over Diversity, Multi-Modal Distance, and MultiModality metrics. As observed in MUGL [15], quality scores such as FID based on feature representations often fail to capture the key action dynamics of the motion sequences. We empirically observed that these scores correlate poorly with the visual quality of motion generations.

5. Conclusion

In this paper, we enhance the performance of the diffusion-based motion generation model using a multi-part fusion-based retrieval augmented motion generation framework, MoRAG. Our method incorporates additional guidance into the motion generation pipeline using the diverse motion sequences constructed from the samples retrieved from part-specific motion databases. We propose a simple solution to construct multiple diverse, semantically coherent motion samples from the database, even for unseen descriptions, which is not feasible with existing full-body motion retrieval approaches. This makes MoRAG a more viable alternative for text-to-human motion generation by combining the strengths of both retrieval models — rapid motion sample construction and generative models — the ability to create novel outputs.

Future work could extend our approach to other architecture-based generative models. Incorporating additional body part information, such as fingers, head, and lip movements from respective part-specific databases, would enhance the realism of generated samples and better handle more complex language descriptions. Furthermore, our approach can be applied to create new data samples, useful for both training and guiding motion generation models, thereby expanding its potential to handle unusual inputs.

References

- Hyemin Ahn, Timothy Ha, et al. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, pages 5915–5920, 2018. 2
- [2] C. Ahuja and L. Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, 2019. 2, 3
- [3] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Güll Varol. TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*, 2022. 3
- [4] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. *ICCV*, 2023. 6
- [5] Léore Bensabath, Mathis Petrovich, and Gül Varol. Tmr++: A cross-dataset study for text-based 3d human motion retrieval. 2024. 2, 4, 7, 8
- [6] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator, 2022. 3
- [7] 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. 2, 3, 7

- [8] Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, and Qi Wu. Prompt switch: Efficient clip adaptation for text-video retrieval, 2023. 3
- [9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. 2
- [10] Zijian Gao, Jingyu Liu, Weiqi Sun, Sheng Chen, Dedan Chang, and Lili Zhao. Clip2tv: Align, match and distill for video-text retrieval, 2022. 3
- [11] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Oct. 2021. 2, 3
- [12] 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. 2, 3, 4, 7, 8
- [13] 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.
- [14] Debtanu Gupta, Shubh Maheshwari, Sai Shashank Kalakonda, Manasvi, and Ravi Kiran Sarvadevabhatla. Dsag: A scalable deep framework for action-conditioned multi-actor full body motion synthesis. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023. 4
- [15] Debtanu Gupta, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Mugl: Large scale multi person conditional action generation with locomotion. In WACV, 2022. 9
- [16] Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Action-gpt: Leveraging large-scale language models for improved and generalized action generation. In *IEEE International Conference on Multimedia* and Expo (ICME), 2023. 1, 2
- [17] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. arXiv, 2019. 5
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 3
- [19] Matthew Loper, Naureen Mahmood, et al. Smpl: A skinned multi-person linear model. *SIGGRAPH Asia*, 34(6):248:1– 248:16, 2015. 6
- [20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 8
- [21] 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. 3, 4, 6, 8

- [22] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.
 3, 5
- [23] 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. 2, 3, 4, 5, 6, 8
- [24] 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 Vi*sion and Pattern Recognition (CVPR), pages 722–731, June 2021. 3, 4
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 2, 3
- [27] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. 5
- [28] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020. 6
- [29] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV* 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII, pages 358–374. Springer, 2022. 3
- [30] 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. 1, 3, 7
- [31] Neel Trivedi and Ravi Kiran Sarvadevabhatla. Psumnet: Unified modality part streams are all you need for efficient pose-based action recognition. In *European Conference on Computer Vision*, pages 211–227. Springer, 2022. 4
- [32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.
 6
- [33] Kangning Yin, Shihao Zou, Yuxuan Ge, and Zheng Tian. Trimodal motion retrieval by learning a joint embedding space, 2024. 2, 3, 4
- [34] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 3

- [35] 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. 3
- [36] 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. 1, 2, 3, 7
- [37] 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, 6, 7, 8
- [38] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatiotemporal motion generation and editing. *NeurIPS*, 2023. 1, 2,7
- [39] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey, 2024. 2
- [40] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey, 2023. 1