$\mathcal{V}imoRAG$: Video-based Retrieval-augmented 3D Motion Generation for Motion Language Models

Haidong Xu[1], **Guangwei Xu**, **Zhedong Zheng**[2], **Xiatian Zhu**[3], **Wei Ji**[4], **Xiangtai Li**[5], **Ruijie Guo**, **Meishan Zhang**[1], **Min Zhang**[1], **Hao Fei**[6*]
[1] Harbin Institute of Technology (Shenzhen) [2] University of Macau [3] University of Surrey
[4] Nanjing University [5] Nanyang Technological University [6] National University of Singapore
182haidong@gmail.com, haofei7419@gmail.com

Abstract

This paper introduces VimoRAG, a novel video-based retrieval-augmented motion generation framework for motion large language models (LLMs). As motion LLMs face severe out-of-domain/out-of-vocabulary issues due to limited annotated data, VimoRAG leverages large-scale in-the-wild video databases to enhance 3D motion generation by retrieving relevant 2D human motion signals. While video-based motion RAG is nontrivial, we address two key bottlenecks: (1) developing an effective motion-centered video retrieval model that distinguishes human poses and actions, and (2) mitigating the issue of error propagation caused by suboptimal retrieval results. We design the Gemini Motion Video Retriever mechanism and the Motion-centric Dual-alignment DPO Trainer, enabling effective retrieval and generation processes. Experimental results show that VimoRAG significantly boosts the performance of motion LLMs constrained to text-only input. All the resources (https://walkermitty.github.io/VimoRAG/) are available.

1 Introduction

Generating diverse and realistic human motions from free-form text prompts has significant practical applications, including video gaming, robotic assistance, and virtual reality. Previous advancements, ranging from transformer-based VAEs [1] to recent diffusion-based generative models [2], have led to an increasingly promising generative performance. With the emergence of LLMs, motion-language models (aka. motion LLMs) have been proposed [3, 4]. These unified architectures not only understand various motions but also support motion generation. To achieve competitive capabilities, motion LLMs require training on extremely large-scale datasets to ensure sufficient capacity. Particularly for motion generation, a substantial amount of labeled data (i.e., text-motion pairs) is essential, without which, models face severe out-of-domain (OOD) and out-of-vocabulary (OOV) issues, making it difficult to generalize to the vast variety of dynamic human motions.

However, existing datasets of text-motion pairs are severely limited in scale, comprising only approximately 14k samples [5], while the cost of large-scale annotation is prohibitively high. To address this issue, *Zhang et al.* introduce ReMoDiffuse [6], a retrieval-augmented generation (RAG) method that retrieves relevant supplementary supervision signals from 3D motion databases. While this method provides a promising direction to address the scarcity of labeled data, their effectiveness might still remain constrained by the size of existing 3D motion databases, i.e., totaling only 14k samples from datasets such as HumanML3D [5].

To address this, this paper explores an innovative RAG paradigm: *retrieving information from larger-scale in-the-wild videos to supplement abundant motion signals that can enhance motion generation*. Although videos represent a 2D visual modality, intuitively, the 2D human motions depicted in videos

*Corresponding Author: Hao Fei

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

inherently share similar characteristics with 3D human motions [7], which can be utilized to guide the learning process of motion LLMs. Most importantly, existing video data is highly accessible and virtually unlimited in scale, and in-the-wild videos capture diverse and unconstrained human motions, offering strong potential to address OOD/V challenges. To this end, we introduce a simple but effective framework, named `VimoRAG` (cf. Fig. 2). VimoRAG first retrieves a video from an unlabeled video database based on the input text, then inputs both the text and the retrieved videos into an LLM to generate motion tokens, which are finally decoded into a motion sequence using VQ-VAE [8].

Yet this framework is nontrivial, facing at least two major challenges within such a new paradigm ►**Challenge-I:** ***Current video foundation models (VFMs) perform poorly in retrieving human-centric videos.*** Our preliminary experiments indicate that existing VFMs, despite performing well on general-purpose video retrieval tasks (e.g., excelling at recognizing objects and attributes), struggle to distinguish human poses, actions, and behaviors in videos (cf. Fig. 1). ►**Challenge-II:** ***The error propagation caused by inaccurate or low-quality retrieved videos.*** When the retrieval quality is poor, it significantly affects the quality of the generated content. Unfortunately, this issue is not thoroughly investigated in ReMoDiffuse [6].

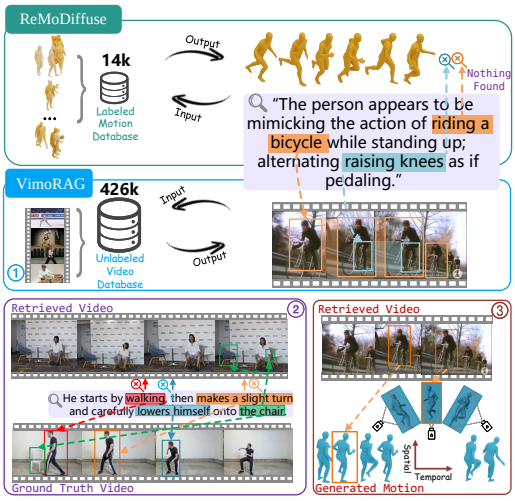

Figure 1: ReMoDiffuse is a RAG-based motion generation method, which is limited by the small scale of motion data and its reliance on annotated captions. We propose `VimoRAG`, which advances in ① enabling retrieval from large-scale, **in-the-wild video databases without text captions**. ② Identifying and overcoming key challenges in human-centric text-to-video retrieval. ③ Ensuring alignment between retrieved videos and generated motions while mitigating error propagation.

To address the first challenge, we design a Gemini Motion Video Retriever (termed *Gemini-MVR*). Gemini-MVR incorporates dual fine-grained retrieval channels, at the action level and object level respectively, where a keypoints-aware router assigns weights to these two retrievers, allowing the system to simultaneously focus on human pose features and object information in complex videos, thereby improving the accuracy of human-centric video retrieval.

To address the second challenge, we propose a Motion-centric Dual-alignment DPO training strategy (named *McDPO*). McDPO is to guide the LLM on how to utilize the prior information from the retrieval video (i.e., when to use it, when not to use it, and how much to rely on it) by performing self-correction.

We leverage a lightweight LLM Phi3-3.8B [9] following *Maaz et al.* [10] to evaluate the performance of this framework. To evaluate its effectiveness, we conduct both cross-domain and in-domain experiments. To explore its potential, we conduct scaling experiments with varying retrieval corpus sizes. Specifically, in OOD scenarios, we conduct zero-shot experiments on the challenging IDEA400 [11] test set, where we verify that VimoRAG exhibits strong generalization capabilities. We evaluate in-domain performance on the representative HumanML3D benchmark and observe that VimoRAG consistently improves most of the metrics compared to existing motion LLMs that operate with text-only input, and further exhibit sustained performance gains as the retrieval corpus expands. In summary, our contributions are as follows:

- To our knowledge, this paper is the first to propose a novel paradigm of **video-based 3D motion RAG**, which significantly alleviates the motion data scarcity bottleneck in existing methods.
- We present the `VimoRAG` framework, which integrates two plug-and-play modules—*Gemini-MVR* retriever and *McDPO* trainer—to address two key bottlenecks: human-centric video retrieval and error propagation in cross-modal motion RAG pipelines.
- Experimental results demonstrate that VimoRAG achieves substantial performance improvements in OOD settings and consistently enhances vanilla motion LLMs in in-domain scenarios. Furthermore, it exhibits clear potential for further gains as the retrieval corpus expands.

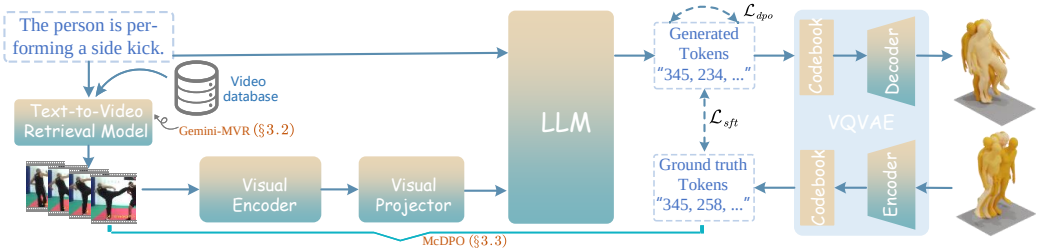

Figure 2: Overview of the VimoRAG pipeline: (1) text-to-video retrieval via Gemini-MVR, and (2) video-augmented motion generation guided by text and retrieved video. Gemini-MVR (Sec. 3.2) is designed to improve cross-modal human-centric video retrieval, while the McDPO training strategy (Sec. 3.3) mitigates error propagation caused by noisy retrievals.

2 Related Work

Motion generation [12] has long been a hot research topic in the related community, aiming to generate human-like 3D motion based on a given prompt, such as text, action, or incomplete motion. Text-to-motion is among the most significant task settings and has attracted substantial research attention [13, 6, 14–18]. For instance, T2M-GPT [1] explores a generative framework utilizing VQ-VAE [8] and Transformer [19] for motion generation. MDM [20] introduces a diffusion-based [21] generative model trained across multiple motion tasks. MLD [22] enhances the latent diffusion model to produce motions conditioned on various inputs. These motion generation specialists, following in-house training, deliver high performance in motion generation.

In recent years, the emergence of LLMs [23] has demonstrated unprecedented levels of intelligence, driving the evolution from specialists to generalists. In the motion domain, motion language models [3, 4, 24, 25] have been proposed, where motion-aware encoders are connected to a central LLM, leading to motion generalists (also named motion LLMs) capable of perceiving various motions. Further, motion generators (e.g., diffusion or VQ-VAE models) are integrated to achieve unified motion LLMs for both comprehension and generation [3]. To achieve robust motion manipulation capabilities, these motion LLMs require fine-tuning on large annotated datasets. However, compared to comprehension tasks, motion generation is more reliant on data, but motion annotation datasets are often quite limited due to the high cost of annotation.

In a recent study, Remodiffuse [6] introduces a motion generation method based on the retrieval-augmented generation (RAG) paradigm. It performs text-to-text retrieval from a labeled 3D motion database to fetch related motion signals and enhance generation quality. However, as previously mentioned, existing motion databases are typically limited in scale. In contrast, large-scale video databases are more accessible and diverse. Motivated by this, we explore a human motion-centric video retrieval framework to support 3D motion generation, where motion-consistent 2D features extracted from videos are effectively transferred to guide the 3D motion synthesis. Compared to Remodiffuse, our approach introduces two key innovations. First, we leverage cross-modal text-to-video retrieval to eliminate the reliance on motion databases that require manually written textual descriptions. Second, we are the first to identify the issue of error propagation in motion-RAG frameworks, and propose a novel algorithm, McDPO, to address it.

Notably, several motion LLM studies [26, 27, 24] have also explored human-related video tasks. Inspired by MotionBank, we construct our video corpus from multiple action-centric datasets. Unlike previous works that focus on building high-quality video collections, this work instead centers on validating the potential and robustness of VimoRAG framework when retrieving from a wild video corpus, and on addressing the potential inconsistency between video input and the generated motion.

3 Our Approach: `VimoRAG`

VimoRAG is a **Vi**deo-based **R**etrieval-**a**ugmented **Mo**tion **G**eneration framework. We first introduce the overall architecture and our collected video database in Section 3.1. Then we describe the details of two key components (Gimi-MVR and McDPO) in Section 3.2 and Section 3.3 respectively.

3.1 Preliminaries

Overall Architecture. As depicted in Figure 2, VimoRAG is a pipeline composed of two essential steps. The initial step involves text-to-video retrieval, in which a motion description text is used to retrieve semantically relevant videos (the rank-1 video is used in this paper) from an unlabeled wild video database with our Gimi-MVR model. The subsequent step involves video-augmented motion generation, where both the text and retrieved videos are fed into the generation model to produce the motion sequence. Leveraging our novel McDPO trainer, we facilitate the contextually aligned motion generation process.

Human-centric Video Database. In theory, it is possible to retrieve the most semantically relevant videos from all available short videos on the internet to overcome the challenges of open-vocabulary text descriptions in the industry. However, in academia, in order to efficiently verify the feasibility of our method and to advance this research direction, we gather and filter a vast human-centric video database (**HcVD**) consisting of 425,988 videos, sourced from [11, 28–33].

To train a better retrieval model on human-centric videos, we leverage the widely used Qwen2-VL [34] model to synthesize textual descriptions following *Zhao et al* [35]. We clarify that the synthetic captions are used solely to train the retriever, and are not involved in the motion RAG pipeline. VimoRAG is a ready-to-use framework that requires neither large-scale annotated videos nor text-to-text retrieval. Additionally, we enhance the dataset quality by employing AlphaPose [36] to filter out videos without human detection. More details are presented in the Appendix C.

3.2 Gemini Motion Video Retriever

Figure 3 illustrates the overall architecture of our method. The model builds upon the CLIP [37] framework, where the text and video branches are encoded separately, and their similarity is computed for subsequent ranking.

Our model consists of two independent retrievers—object-level and action-level—whose outputs are fused by a lightweight router to produce the final similarity score. The object-level retriever captures visual entities (objects) and their textual arguments, while the action-level retriever targets motion and predicate-level semantics.

This design is driven by two main motivations, as discussed in the introduction: **First**, from our analysis of failure cases in existing VFMs, we observe that when a query describes only human actions without clear environmental or object cues, VFMs often struggle to retrieve the correct video. To address this, we introduce a dedicated action-level retriever. **Second**, we observe that many motions involve interaction with the environment, making object-level cues essential for understanding actions. While existing VFMs may struggle with behavior-only queries, they encode rich general knowledge from large-scale pretraining. To leverage this strength, we retain the VFM itself as the object-level retriever ($\theta_{\mathcal{G}}$ and $\theta_{\mathcal{O}}$ in Figure 3).

Specifically, for the action-level retriever, we carefully design an action encoder $\theta_{\mathcal{A}}$ and predicate semantic extractor $\theta_{\mathcal{P}}$. In the action encoder, we first extract the 2D human keypoints from the m frames of an input video v. Then we

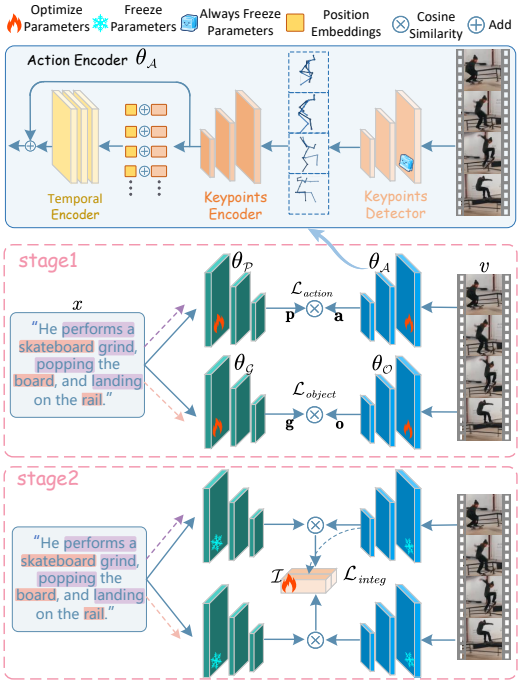

Figure 3: The architecture of the **Gemini-MVR** model. $\theta_{\mathcal{P}}$ and $\theta_{\mathcal{G}}$ represent the predicate semantic extractor and the argument semantic extractor, respectively. $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{O}}$ denote the action encoder and object encoder, respectively. We simply introduce a lightweight action-level retriever and a routing module $\mathcal{I}$, while keeping the architecture of VFMs unchanged. This twin-module design provides strong extensibility.

encode each frame of keypoints to the feature space separately. After this, we add learnable position embeddings frame by frame and feed them into a temporal encoder. In order to avoid information loss in the sequential encoding process, the residual operation is adopted here. To obtain a good initial weight, we adopt the pretrained AlphaPose [36] and the pretrained MotionBERT [38] as the keypoints detector and encoder accordingly. The temporal encoder is implemented by a transformer block [19] and is initialized randomly. $\theta_{\mathcal{P}}$ is initialized from the text encoder in InternVideo [39].

We train the action-level retriever using the contrastive learning loss [40] $\mathcal{L}_{action} = \mathcal{L}_{p2a} + \mathcal{L}_{a2p}$ ($\mathcal{L}_{a2p}$ is provided in the Appendix D due to space limitations):

$$\mathcal{L}_{p2a} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(s(\mathbf{p}_i, \mathbf{a}_i))}{\sum_{j=1}^{B}\exp(s(\mathbf{p}_i, \mathbf{a}_j))}\,, \tag{1}$$

where $\mathbf{p}$ and $\mathbf{a}$ are embedding vectors encoded by $\theta_{\mathcal{P}}$ and $\theta_{\mathcal{A}}$ for the input value t and v, respectively. B denotes the number of text-video pairs in a training batch. $s(,)$ denotes the cosine similarity. For the object-level retriever, we directly adopt InternVideo [39], one of the most widely used VFMs, owing to its extensive common knowledge. The argument semantic extractor $\theta_{\mathcal{G}}$ and object encoder $\theta_{\mathcal{O}}$ are initialized using the text encoder and video encoder of InternVideo. During the fine-tuning stage, we utilize a loss function $\mathcal{L}_{object}$ similar to $\mathcal{L}_{action}$. Let $\mathbf{g}$ and $\mathbf{o}$ denote the embedding vectors encoded by $\theta_{\mathcal{G}}$ and $\theta_{\mathcal{O}}$ for the input value t and v, respectively. We just replace $\mathbf{p}$ with $\mathbf{g}$ and replace $\mathbf{a}$ with $\mathbf{o}$ in $\mathcal{L}_{action}$, and then we can get the symmetric loss function $\mathcal{L}_{object}$. It is worth emphasizing that the two semantic extractors do not explicitly extract the semantics of predicates and arguments. We hypothesize that, through contrastive learning with different video features, each semantic extractor implicitly captures its respective focus—predicates and arguments.

Following the independent training of the two retrievers in stage 1, the subsequent step involves determining an effective approach to integrate the two retrievers. A key consideration is that the allocation of weights should adapt to the characteristics of different motion videos. Building on this foundation, we propose an action-aware similarity integrator model, denoted as $\mathcal{I}$. Considering the two-level retrieval models can be significantly large, the optimal $\mathcal{I}$ should be sufficiently lightweight to minimize retrieval delay. To achieve this, we employ a simple linear method. The cosine similarity $s(t, v)$ is calculated as follows:

$$s(t, v) = \frac{\mathcal{I}_0(\mathbf{a})s(\mathbf{p}, \mathbf{a})}{\mathcal{I}_0(\mathbf{a}) + \mathcal{I}_1(\mathbf{a})} + \frac{\mathcal{I}_1(\mathbf{a})s(\mathbf{g}, \mathbf{o})}{\mathcal{I}_0(\mathbf{a}) + \mathcal{I}_1(\mathbf{a})}\,, \tag{2}$$

where $\mathcal{I}_0$ and $\mathcal{I}_1$ denote two output channels of $\mathcal{I}$. In stage 2, the training loss function is $\mathcal{L}_{integ} = \mathcal{L}_{t2v} + \mathcal{L}_{v2t}$, where $\mathcal{L}_{t2v}$ is calculated as follows:

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(s(t_i, v_i))}{\sum_{j=1}^{B}\exp(s(t_i, v_j))}\,. \tag{3}$$

Notably, $\mathcal{L}_{v2t}$ and $\mathcal{L}_{t2v}$ exhibit a symmetric structure.

Training and Inference. Since the action-level retriever is trained from scratch, while the object-level retriever has already been pretrained on a large corpus of text-video pairs, we first pretrain the action-level retriever using a subset of the HcVD In Stage 1, we fine-tune both pretrained retrieval models in parallel, allowing for the optimization of all modules. In Stage 2, we freeze the two pretrained models and optimize only the similarity integrator. During the retrieval stage, we compute the similarity between the query text in the benchmark datasets for motion generation and each video in the HcVD using Equation 2. More implementation details can be found in the Appendix D.

3.3 Motion-centric Dual-alignment DPO Trainer

As illustrated in Figure 4, to fully leverage the descriptive information in the text and the rich 2D visual prior in the retrieved videos, we utilize LLM to project all the information from different modalities into the language space in Stage 1. However, as mentioned in the introduction, there is an inherent gap between the 2D visual prior and the target 3D motion, as the motion prior in the retrieved videos represents only a sample of the full target motion space. Additionally, the 2D visual priors do not always align semantically with the text. To guide the LLM to learn the appropriate direction for generation when such gaps arise, we construct the McDPO training set using a dual-alignment reward

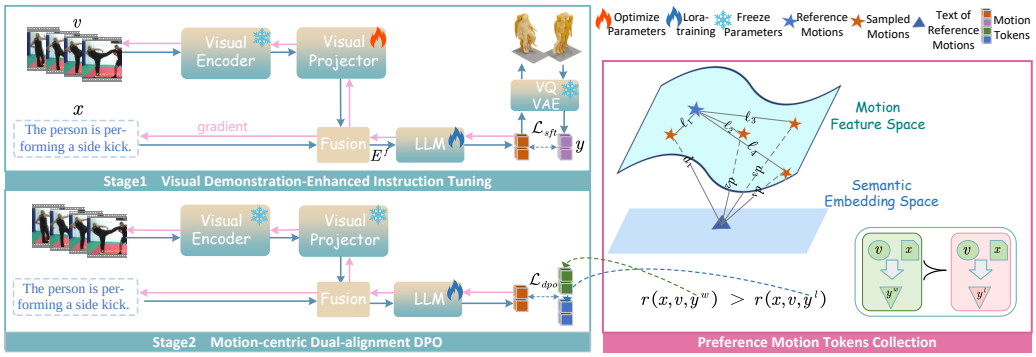

Figure 4: The **McDPO** training strategy. Given a text t and a retrieved video v, we first perform visual demonstration-enhanced instruction tuning to establish a base reference model π_{ref}. Then, based on the motion-centric dual-alignment reward model, we construct a preference dataset and apply DPO training. The reward model jointly measures motion similarity in the feature space and semantic consistency with the text, guiding the model to learn informative motion priors and maximize preference rewards through self-improvement.

Table 1: Zero-shot results on IDEA400 test set. All motions are generated by the models trained on HumanML3D training set. All results are reproduced using the officially released models (codes). VimoRAG achieves the best FID score, with other metrics closely matching SoTA.

Model	FID ↓	R-Precision ↑			MM Dist ↓	Diversity ↑
		Top 1	Top 2	Top 3		
• *Motion Specialists*						
MoMask [13]	$5.982^{\pm.089}$	$0.110^{\pm.003}$	$0.195^{\pm.006}$	$0.266^{\pm.006}$	$5.625^{\pm.023}$	$7.558^{\pm.119}$
T2M-GPT [1]	$5.359^{\pm.078}$	$0.108^{\pm.006}$	$0.186^{\pm.005}$	$0.255^{\pm.006}$	$5.773^{\pm.037}$	$7.648^{\pm.100}$
MDM [20]	$5.907^{\pm.107}$	$0.113^{\pm.004}$	$\mathbf{0.200}^{\pm.004}$	$\mathbf{0.278}^{\pm.004}$	$6.013^{\pm.020}$	$\mathbf{8.131}^{\pm.080}$
MotionDiffuse [2]	$5.485^{\pm.038}$	$0.110^{\pm.002}$	$0.194^{\pm.002}$	$0.266^{\pm.003}$	$6.038^{\pm.005}$	$6.884^{\pm.095}$
MLD [22]	$5.410^{\pm.085}$	$\mathbf{0.114}^{\pm.003}$	$0.200^{\pm.005}$	$0.270^{\pm.004}$	$6.005^{\pm.029}$	$7.558^{\pm.086}$
MotionGPT [4]	$6.202^{\pm.186}$	$0.087^{\pm.005}$	$0.151^{\pm.007}$	$0.209^{\pm.008}$	$6.640^{\pm.025}$	$7.684^{\pm.111}$
ReMoDiffuse [6]	$9.639^{\pm.069}$	$0.110^{\pm.004}$	$0.188^{\pm.006}$	$0.256^{\pm.005}$	$\mathbf{5.465}^{\pm.015}$	$7.540^{\pm.120}$
• *Motion LLMs*						
MotionGPT [3]	$5.544^{\pm.174}$	$0.096^{\pm.005}$	$0.171^{\pm.008}$	$0.236^{\pm.008}$	$6.300^{\pm.032}$	$7.509^{\pm.096}$
VimoRAG (Ours)	$\mathbf{2.388}^{\pm.056}$	$0.113^{\pm.005}$	$0.193^{\pm.008}$	$0.270^{\pm.011}$	$5.888^{\pm.061}$	$7.688^{\pm.197}$

model, **allowing LLM to learn the most informative priors to maximize preference rewards by distinguishing its own struggle cases.** We describe the details in the following parts.

Visual Demonstration-Enhanced Instruction Tuning. Given an input text $x = \{x_1, x_2, ...x_{n_x}\}$ of n_x tokens, a system prompt $\mathcal{S}$ of the LLM, an instruction template $\mathcal{T}$, and a retrieved video $v = \{v_1, v_2, ..., v_{n_v}\}$ of n_v frames, we first embed v into k segment-wise embeddings $E^v = \{E_1^v, E_2^v, ..., E_k^v\}$ following *Maaz et al.* [10] We then concatenate these elements to obtain the final input embeddings $E^f = [emb(\mathcal{P}), \mathcal{T}(emb(x), E^v)]$ where $emb(\cdot)$ denotes the embedding layers of the LLM. For the target motion, we leverage the widely used VQ-VAE [8] to encode the continuous motion sequence into discrete motion tokens $y = \{y_1, y_2, ..., y_{n_y}\}$ following MotionGPT [3]. Inspired by them, we fine-tune the LLM using the following instruction-tuning format in Stage 1:

> **System Prompt $\mathcal{P}$:** You are a helpful AI assistant.
> $\%<|user|>$
> **Instruction Template $\mathcal{T}$:** Generate a sequence of motion tokens matching the following human motion description. You can use the video as a reference. Video information: { Retrieved Video v} Motion description: {Input Text x }
> $\%<|assistant|>$
> **Answer y:** {Sequence of Motion Tokens}

The loss function $\mathcal{L}_{sft}$ in Stage 1 is as followed: $\mathcal{L}_{sft} = -\sum_n \log p_\theta(y_n | y_{<n}, E^f)$.

Motion-centric Dual-alignment DPO. To empower the motion LLM with the ability to autonomously adapt to video priors of differing quality during the generation process, we introduce

Table 2: Results on HumanML3D test set. "*" denotes results from original papers; others are reproduced using official code. **VimoRAG achieves the best FID and competitive performance across metrics among existing motion LLMs**. This framework significantly improves five metrics (highlighted in red) over MotionGPT [3] (Phi3-3.8B), **demonstrating the substantial advantage of incorporating video priors**. The complete results with confidence intervals are shown in Table 7.

Model	Backbone	FID ↓	R-Precision ↑			MM Dist↓	Diversity↑
			Top 1	Top 2	Top 3		
Motion Specialists							
MoMask [13]	–	0.048	0.519	0.715	0.809	2.955	9.632
T2M-GPT [1]	–	0.112	0.489	0.679	0.774	3.125	9.691
MDM [20]	–	0.454	0.419	0.606	0.712	3.636	9.449
MotionDiffuse [2]	–	0.672	0.492	0.685	0.784	3.085	9.499
MLD [22]	–	0.425	0.468	0.656	0.759	3.266	9.698
ReMoDiffuse [6]	–	0.125	0.493	0.676	0.775	3.047	9.211
LMM* [27]	–	0.040	0.525	0.719	0.811	2.943	9.814
MotionLab* [46]	–	0.167	–	–	0.810	2.912	9.593
MotionLCM* [47]	–	0.304	0.502	0.698	0.798	3.012	9.607
MotionCLR* [48]	–	0.269	0.544	0.732	0.831	2.806	–
MotionGPT* [4]	–	0.232	0.492	0.681	0.778	3.096	9.528
BiPO* [49]	–	**0.030**	0.523	0.714	0.809	2.880	9.556
StableMoFusion* [50]	–	0.098	0.553	0.748	0.841	–	9.748
MoGenTS* [51]	–	0.033	0.529	0.719	0.812	2.867	9.570
LAMP* [52]	–	0.032	**0.557**	**0.751**	**0.843**	**2.759**	9.571
Motion LLMs							
MotionGPT-2* [25]	Llama3-8B	0.191	0.496	0.691	0.782	3.080	9.860
MotionLLM* [53]	GPT4+Gemma-2B	0.230	0.515	–	0.801	2.967	**9.908**
*Wang et al.** [54]	Llama2-13B	0.166	0.519	–	0.803	2.964	–
ScaMo* [55]	codesize 512-3B	0.617	0.443	0.627	0.734	3.340	9.217
AvatarGPT* [56]	Llama-13B	0.567	0.389	0.539	0.623	–	9.489
MotionGPT* [3]	Llama-13B	0.567	–	–	–	3.775	9.006
MotionGPT [3]	Phi3-3.8B	0.501	0.396	0.575	0.673	3.724	9.475
VimoRAG (Ours)	Phi3-3.8B	0.131 -73%	0.452 +14%	0.655 +14%	0.764 +13%	3.146 -15%	9.424 -1%

motion-centric dual-alignment DPO training strategy. As illustrated in Figure 4, after the completion of Stage 1 training, we obtain a base reference model π_{ref}. The videos used in Stage 1 are the retrieval results of the retrieval model, not ground truth, meaning that π_{ref} has been learning to handle the potential gap between text and video during training. However, π_{ref} struggles to learn this aspect. In fact, we discover early in our experimens that π_{ref}'s performance on the training set is not stable, particularly when the gap is large. The experimental results shown in Figure 6 support this observation (`NMC-R1` is worse than `MC-R1`).

To construct such a training set, we first use π_{ref} to randomly sample κ times to generate a motions' candidate set. In contrast to recent works [41, 42], which rely on human or AI-based proxies as reward models, we devise a more efficient dual-alignment reward model. This model outputs a reward score for a sampled motion sequence $\hat{y}_i$ as follows:

$$r(x, v, \hat{y}_i) = -(w_\ell \frac{\ell(\hat{y}_i, y)}{\sum_{j\in\kappa}\ell(\hat{y}_j, y)} + w_d \frac{d(\hat{y}_i, x)}{\sum_{j\in\kappa} d(\hat{y}_j, x)}), \tag{4}$$

where $\ell(\cdot)$ and $d(\cdot)$ denote the distribution distance and Euclidean distance computed based on the features of the two inputs, respectively. w_ℓ and w_d are hyper-parameters to respectively control the degree of alignment within the motion modality and between the text-motion modalities. The reward model encourages the preferred motions to be closer to the reference motions in the feature space and more semantically aligned with the paired text in the semantic space. By leveraging the reward model, we collect the chosen motions y_w and the rejected motions y_l from the existing motions' candidate set, thereby constructing a DPO dataset $\mathcal{D}_{dpo} = \{(x, v, y^w, y^l)\}$. Finally, we adopt the following DPO training objective [43, 44]. Here π_θ and σ denote the policy model and the logistic function, respectively. γ here denotes the weighting coefficient.

$$\mathcal{L}_{dpo} = -\mathbb{E}_{(x,v,y^w,y^l)\sim\mathcal{D}_{dpo}}\left[\log\sigma\Big(\gamma\log\frac{\pi_\theta(y^w|x,v)}{\pi_{ref}(y^w|x,v)} - \gamma\log\frac{\pi_\theta(y^l|x,v)}{\pi_{ref}(y^l|x,v)}\Big)\right]. \tag{5}$$

Training Strategy. We employ phi-3-mini [9] as the backbone LLM and utilize the LoRA [45] tuning method in both training stages. In Stage 1, we additionally tune the visual adapter while freezing the remaining modules. In Stage 2, all modules except for the LLM are kept frozen. Due to space constraints, further training configurations are detailed in the Appendix D.

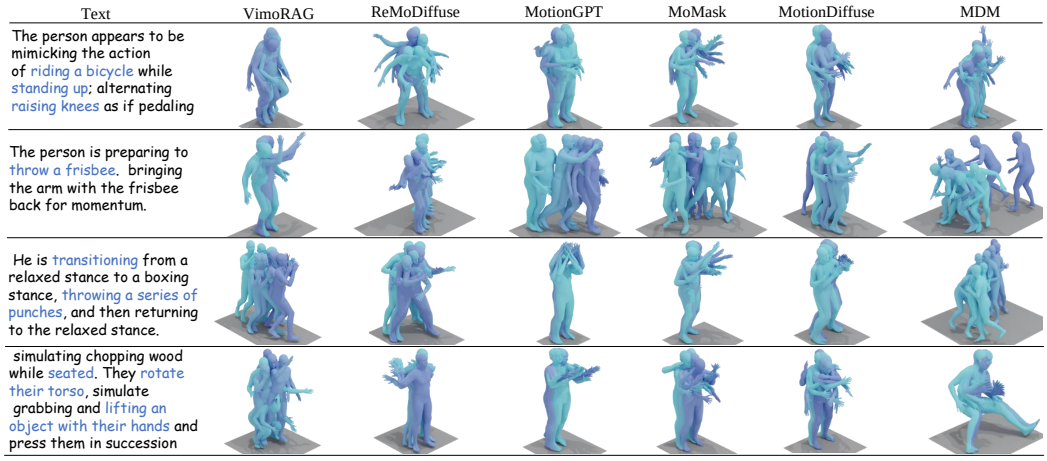

Figure 5: Zero-shot qualitative results on IDEA400 test set. All motions are directly generated by the models trained on HumanML3D training set. The text presented here only includes words related to motion due to space constraints. The full text and more results are available in Figure 12 and 13.

4 Experiment

4.1 Datasets, Metrics and Baselines

We conduct extensive experiments on two widely used large-scale datasets following the existing works [16]. The first is the **IDEA400** dataset, a high-quality whole-body motion dataset composed of 12.5K clips and 2.6M frames in MotionX [11], which is utilized to assess OOD performance. The other dataset, **HumanML3D** [5], comprising 14,616 motion clips and 44,970 text descriptions, is utilized to evaluate in-domain performance. For evaluation metrics, we adopt several widely recognized measures: *Frechet Inception Distance* (FID), *R-recall*, *MultiModal Distance* (MM Dist), *Diversity*. More details can be found in Appendix A due to the space limit.

4.2 Implementation Details

We implement VimoRAG with PyTorch. We use the same Gemini-MVR for all the text-to-motion experiments because the retrieval model is decoupled with the generation phrase in our framework. In Stage 1 of McDPO, we train 2 epochs with a learning rate 2e-4 for LoRA parameters ($rank = 128, \alpha = 256$), with a learning rate 2e-5 for the visual adapter's parameters. In Stage 2 of McDPO, we train 1 epoch with a learning rate 2e-4. Inference is conducted using a single NVIDIA A800 GPU, while training is accelerated using 8 GPUs to enhance efficiency. Further details regarding the model configurations, training settings, and pose representation are provided in the Appendix A.

4.3 Main Results

Quantitative Results. Tables 1 and 2 present a quantitative comparison between VimoRAG and SoTA techniques. For a fair comparison, each experiment is conducted ten times, and we report the results with a 95% confidence interval. As illustrated in Table 1, VimoRAG achieves the best FID score, indicating its strong generalization capability in generating high-fidelity motions in OOD scenarios. According to Table 2, VimoRAG **outperforms MotionGPT** [3] **by a large margin** across all metrics when using the same backbone. It achieves the best FID score and competitive performance on other metrics among motion LLMs based on the Phi-3 3.8B backbone. As shown in Figure 7, **performance improves steadily with larger retrieval sets, highlighting VimoRAG's potential to enhance motion LLMs.**

Qualitative Comparison. Figure 5 demonstrates the qualitative comparison results on the IDEA400 test set. Among the results, those from VimoRAG appear to align more closely with the intended meaning of the given text. The text descriptions in IDEA400 are quite intricate and differ significantly from those in the HumanML3D dataset. In the fourth demonstration, the text entails multiple changes in action, such as *"seated, rotate, grabbing, lifting"*. Among all the motions showcased, the results from VimoRAG encompass more action types related to the text. Similar phenomena can be observed in other cases as well. More results are shown in Appendix (Section B) and our anonymous GitHub.

Table 3: Ablation study on HumanML3D validation set. `Gem` denotes Gemini-MVR retriever, `Mc` denotes McDPO, `Ran` denotes random retriever, `Int` denotes InternVideo retriever. The FID drop of `Gem+Mc` relative to each setting is highlighted in blue.

Settings	FID↓	R-Precision↑		
		Top 1	Top 2	Top 3
Gem+Mc	0.148	0.429	0.625	0.756
Ran+Mc	0.544 ↓72.8%	0.420	0.644	0.750
Int+Mc	0.205 ↓27.8%	0.433	0.638	0.736
Gem	0.260 ↓43.1%	0.403	0.582	0.682

Table 4: Text-to-video retrieval performance on HcVD test sets. Compared to the object-level VFM, Gemini-MVR achieves a significant improvement in the Recall@1 metric.

Retriever	R@1↑	R@5↑	R@10↑	MnR↓
Human-centric Video				
InternVideo	53.6	84.5	92.3	4.2
Gemini-MVR	58.3 ↑8.8%	87.3	93.7	3.6
Single Human-centric Video				
InternVideo	52.3	84.0	91.5	4.5
Gemini-MVR	61.0 ↑16.6%	89.2	94.1	3.5

4.4 Ablation Study

Impact of Motion Video Retriever. To study the the influence of the video retriever, we replace the Gemini-MVR with two other retrievers: one being the fine-tuned InternVideo [39] model, and the other being random retrieval from the HcVD. The results in Table 3 show that random retrieval leads to a notable increase in the FID score, indicating the importance of accurate video priors in generating high-fidelity results. Moreover, replacing the retriever with InternVideo also leads to a rise in FID score, further confirming the effectiveness of Gemini-MVR. It is important to note that the performance of the retrieval model in RAG is influenced by the generation module, hence we also conduct text-to-video retrieval experiments. As shown in Table 4, Gemini-MVR achieves an 8.8% increase in R@1 for the human-centric video set (pool size is 1990), and a 16.6% increase in R@1 for the single human-centric video set. These results further validate the effectiveness of Gemini-MVR.

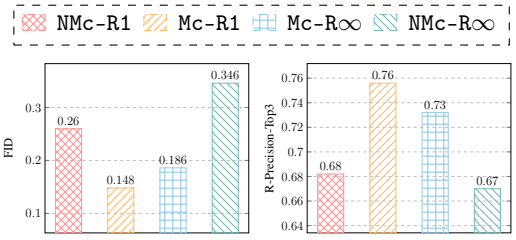

Figure 6: In-depth exploration of McDPO. `Mc` stands for McDPO setting, and `NMc` stands for No McDPO setting. `R1` indicates the use of rank-1 video during inference, while `R∞` indicates the use of random video during inference.

Impact of McDPO. Table 3 demonstrates that removing McDPO leads to a substantial overall performance drop, as evidenced by the comparison between the `Gem+Mc` and `Gem` settings. These findings indicate that the McDPO trainer effectively mitigates the error propagation issue, a point we further analyze in detail in the discussion section 4.5.

4.5 In-depth Discussion

The Effect of Retrieval Database Size. Figure 7 illustrates the changes in the FID and MM-Dist metrics as the size of the database increases. It can be observed that as the database size increases, both metrics show a decreasing trend, demonstrating VimoRAG's scalability potential with larger retrieval corpora — a promising property given that wild video datasets can be easily scaled in real-world applications.

The Role of McDPO. To further analyze the role of McDPO in VimoRAG, we conduct a crossover experiment involving different video priors. As illustrated in Figure 6, the FID score in the `Mc-R∞` setting is lower than that in the `NMc-R∞` setting, indicating that McDPO achieves a significantly lower FID score, **even when random video priors are utilized**. This further suggests that McDPO effectively enables the model to disregard non-informative video priors. Moreover, we observe that McDPO also achieves a significantly lower FID score and higher R-Precision score when provided with a rank-1 video. In essence, the model appears to possess the ability to distinguish between informative and non-informative video priors, effectively utilizing relevant information

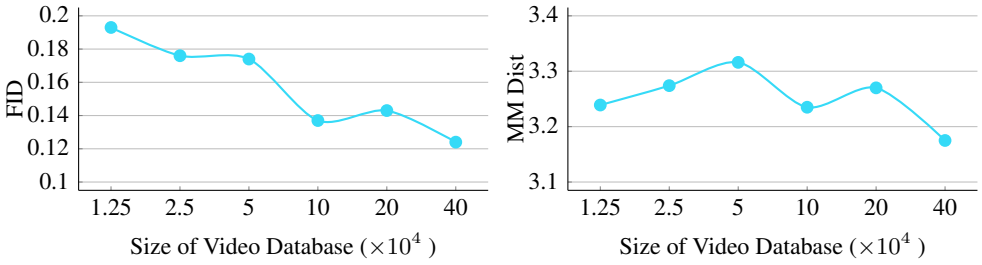

Figure 7: As the video retrieval database grows, VimoRAG shows steadily improving performance, demonstrating strong potential for real-world applications.

while disregarding noise. We hypothesize that this ability arises from the model's implicit alignment with these two perspectives.

5 Discussion

Limitations. A limitation of our work is that VimoRAG is designed for LLMs, which results in longer processing times than those of existing smaller models (motion specialists). We evaluate the latency of our framework, with detailed results shown in Table 6. We acknowledge this constraint and intend to explore methods for reducing latency in future works.

Impact Statement. While the capabilities of our framework present significant opportunities for motion generation, they also raise ethical concerns regarding its potential misuse. Malicious users could exploit the system to create content that promotes violence or other harmful behaviors, posing risks to societal well-being. To mitigate these potential impacts, we will implement a strict licensing mechanism upon the release of our method. This licensing will govern the academic research and applications of our model, ensuring that its deployment is aligned with ethical standards and responsible use.

6 Conclusion and Future Work

In this paper, we propose `VimoRAG`, a novel framework that integrates large-scale in-the-wild video databases to enhance motion generation for motion LLMs. We tackle two key challenges—human-centric video retrieval and error propagation—through the proposed Gemini-MVR model and the McDPO training strategy. Our experiments show that `VimoRAG` further boosts motion LLMs with substantial performance gains in both OOD and in-domain settings, and its performance steadily improves with larger retrieval corpora, showing strong scalability potential.

In future work, we aim to advance along two directions. First, we will explore which types of LLMs are most suitable as the backbone of `VimoRAG`, and how to define appropriate metrics for automatically selecting them. While numerous LLMs are available today, our focus in this paper is not on identifying the best-performing LLM, but rather on addressing the core challenges within the RAG system. Second, building upon the success of video-based RAG, we plan to incorporate video, 3D data, and potentially even image data as priors to develop a unified RAG framework and investigating whether this multimodal integration can lead to further performance gains.

Acknowledgments

This work was supported in part by Shenzhen Science and Technology Program (JCYJ20241202123503005). We also acknowledge supports from Guangdong Basic and Applied Basic Research Foundation 2025A1515012281, Nanjing Municipal Science and Technology Bureau 202401035 and University of Macau MYRG-GRG2024-00077-FST-UMDF.

References

[1] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the CVPR*, pages 14730–14740, 2023.

[2] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024.

[3] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI*, pages 7368–7376, 2024.

[4] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Proceedings of the NeurIPS*, pages 20067–20079, 2023.

[5] 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 CVPR*, pages 5152–5161, 2022.

[6] 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 ICCV*, pages 364–373, 2023.

[7] Liangdong Qiu, Chengxing Yu, Yanran Li, Zhao Wang, Haibin Huang, Chongyang Ma, Di Zhang, Pengfei Wan, and Xiaoguang Han. Vimo: Generating motions from casual videos. *arXiv preprint arXiv:2408.06614*, 2024.

[8] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Proceedings of the NeurIPS*, 30, 2017.

[9] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[10] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024.

[11] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: a large-scale 3d expressive whole-body human motion dataset. In *Proceedings of the NeurIPS*, pages 25268–25280, 2023.

[12] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449, 2024.

[13] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the CVPR*, pages 1900–1910, 2024.

[14] Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibei Yang, Xin Chen, Jingyi Yu, and Lan Xu. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of the CVPR*, pages 482–493, 2024.

[15] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Proceedings of the ECCV*, pages 358–374, 2022.

[16] Jinpeng Liu, Wenxun Dai, Chunyu Wang, Yiji Cheng, Yansong Tang, and Xin Tong. Plan, posture and go: Towards open-vocabulary text-to-motion generation. In *Proceedings of the ECCV*, pages 445–463, 2025.

[17] Ke Fan, Jiangning Zhang, Ran Yi, Jingyu Gong, Yabiao Wang, Yating Wang, Xin Tan, Chengjie Wang, and Lizhuang Ma. Textual decomposition then sub-motion-space scattering for open-vocabulary motion generation. *arXiv preprint arXiv:2411.04079*, 2024.

[18] Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-wen Chen. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In *Proceedings of the CVPR*, pages 23222–23231, 2023.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the NeurIPS*, 2017.

[20] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *Proceedings of the ICLR*, 2023.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Proceedings of the NeurIPS*, pages 6840–6851, 2020.

[22] 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 CVPR*, pages 18000–18010, 2023.

[23] Yingqiang Ge, Wenyue Hua, Kai Mei, jianchao ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. Openagi: When llm meets domain experts. In *Proceedings of the NeurIPS*, pages 5539–5568, 2023.

[24] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024.

[25] Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*, 2024.

[26] Liang Xu, Shaoyang Hua, Zili Lin, Yifan Liu, Feipeng Ma, Yichao Yan, Xin Jin, Xiaokang Yang, and Wenjun Zeng. Motionbank: A large-scale video motion benchmark with disentangled rule-based annotations. *arXiv preprint arXiv:2410.13790*, 2024.

[27] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. In *Proceedings of the ECCV*, pages 397–421, 2024.

[28] Orit Kliper-Gross, Tal Hassner, and Lior Wolf. The action similarity labeling challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):615–621, 2011.

[29] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *Proceedings of the ICCV*, pages 2556–2563, 2011.

[30] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[31] Weiyu Zhang, Menglong Zhu, and Derpanis G. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the ICCV*, pages 2248–2255, 2013.

[32] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[33] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the CVPR*, pages 1010–1019, 2016.

[34] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[35] Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025.

[36] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, 2022.

[37] 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 *Proceedings of the ICML*, pages 8748–8763, 2021.

[38] Wentao Zhu, Xiaoxuan Ma, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the ICCV*, pages 15085–15099, 2023.

[39] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[40] Huaishao Luo, Lei Ji, Ming Zhong, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[41] Jenny Sheng, Matthieu Lin, Andrew Zhao, Kevin Pruvost, Yu-Hui Wen, Yangguang Li, Gao Huang, and Yong-Jin Liu. Exploring text-to-motion generation with human preference. In *Proceedings of the CVPR*, pages 1888–1899, 2024.

[42] Massimiliano Pappa, Luca Collorone, Indro Spinelli, and Fabio Galasso. Modipo: text-to-motion alignment via ai-feedback-driven direct preference optimization. *arXiv preprint arXiv:2405.03803*, 2024.

[43] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the NeurIPS*, pages 53728–53741, 2023.

[44] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.

[45] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[46] Ziyan Guo, Zeyu Hu, Na Zhao, and De Wen Soh. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. *arXiv preprint arXiv:2502.02358*, 2025.

[47] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *Proceedings of the ECCV*, pages 390–408, 2024.

[48] Ling-Hao Chen, Wenxun Dai, Xuan Ju, Shunlin Lu, and Lei Zhang. Motionclr: Motion generation and training-free editing via understanding attention mechanisms. *arXiv e-prints*, pages arXiv–2410, 2024.

[49] Seong-Eun Hong, Soobin Lim, Juyeong Hwang, Minwook Chang, and Hyeongyeop Kang. Bipo: Bidirectional partial occlusion network for text-to-motion synthesis. *arXiv preprint arXiv:2412.00112*, 2024.

[50] Yiheng Huang, Hui Yang, Chuanchen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. In *Proceedings of the MM*, pages 224–232, 2024.

[51] Weihao Yuan, Yisheng He, Weichao Shen, Yuan Dong, Xiaodong Gu, Zilong Dong, Liefeng Bo, and Qixing Huang. Mogents: Motion generation based on spatial-temporal joint modeling. *Proceedings of the NeurIPS*, pages 130739–130763, 2024.

[52] Zhe Li, Weihao Yuan, Yisheng HE, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence Tianruo Yang. LaMP: Language-motion pretraining for motion generation, retrieval, and captioning. In *Proceedings of the ICLR*, 2025.

[53] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with llms. *arXiv preprint arXiv:2405.17013*, 2024.

[54] Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Qin Jin, and Zongqing Lu. Quo vadis, motion generation? from large language models to large motion models. *arXiv preprint arXiv:2410.03311*, 2024.

[55] Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. *arXiv preprint arXiv:2412.14559*, 2024.

[56] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the CVPR*, pages 1357–1366, 2024.

[57] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *Proceedings of the ECCV*, pages 396–416, 2025.

[58] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction parts.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical results are here.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide full details of the methodology, datasets, and experiments in Section 3, Section 4.2, and the Appendix (Section A, Section C, and Section D).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All resources are available at out github page.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide full details of the methodology, datasets, and experiments in Section 3, Section 4.2, and the Appendix (Section A, Section C, and Section D).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In Table 7 and Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Section 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that we have reviewed and adhered to the NeurIPS Code of Ethics throughout the research process.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: In the Appendix Section 5.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used in this work are properly cited in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide full details of the datasets in Appendix (Section C). A github link is also provided for demo access.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core methodology of this research does not involve the use of LLMs as an important, original, or non-standard component.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

In the **appendix**, we present more experimental settings and results (Section A), more qualitative results (Section B), details of human-centric video database (Section C), more implementation details of VimoRAG (Section D).

A More Experimental Settings and Results

A.1 Experimental Settings

In the training configurations for Gemini-MVR, we set the maximum number of video frames to 16. We employ the Adam optimizer with parameters b1=0.9, b2=0.98, epsilon=1e-6, and a weight decay of 0.2 throughout all training stages. For the action-level retriever, we conduct training for 10 epochs with a batch size of 2048 and a learning rate of 1e-4. In the case of the object-level retriever, we train for 5 epochs using a batch size of 128, the learning rate is set to 4e-6 for the CLIP-related modules and 1e-3 for the remaining modules. Regarding the similarity integrator model, we also train for 5 epochs with a batch size of 128 and a learning rate of 1e-3. Additional details are available in our code.

For the training configurations of McDPO, during Stage 1, we set the batch size to 64, weight decay to 0.0, and the maximum context length to 4096, employing a bf16 precision format. We conduct training for 2 epochs on the HumanML3D training set. In Stage 2, we train for 1 epoch with a learning rate of 2e-4, weight decay of 0.0, batch size of 8, and $\gamma = 0.1$ as defined in Equation 5. We set the temperature of the LLM to 0.9 across all experiments. The maximum number of video frames is set to 16 during the generation stage. Most hyper-parameters are selected through grid search techniques applied to the validation sets.

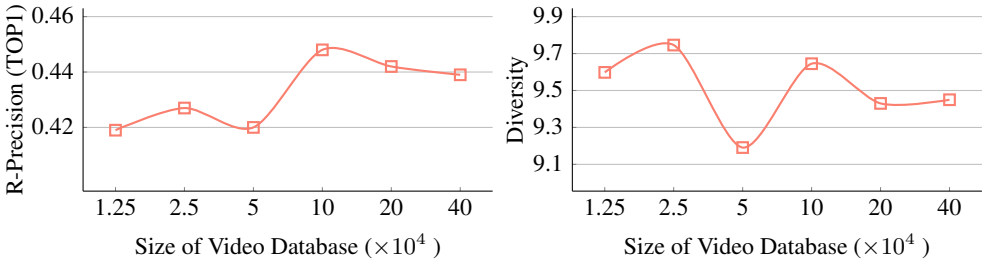

Figure 8: Variation of *R-Precision* (TOP1) and *Diversity* metrics as a function of video database size. A larger database, exceeding 100,000 entries, enhances R-precision. There appears to be no evident correlation between the diversity metric and the size of the database.

Table 5: Performance of **various LoRA parameters** in McDPO on the HumanML3D validation set. The best results are highlighted **in bold** for each setting.

Rank	α	**FID** ↓	**MultiModal Dist** ↓	**R-Precision** ↑			**Diversity** ↑
				Top-1	Top-2	Top-3	
• $\alpha/rank = 2$							
8	16	0.417	3.216	0.437	0.625	0.745	9.636
16	32	0.249	3.146	0.426	0.642	0.764	**9.936**
32	64	0.221	3.109	0.449	**0.670**	0.759	9.718
64	128	0.192	3.089	**0.454**	0.664	0.767	9.560
128	256	**0.179**	**3.046**	0.447	0.667	**0.772**	9.567
• $rank = 128$							
128	64	0.235	3.110	0.452	0.656	0.770	9.591
128	128	**0.156**	3.114	0.445	0.637	0.754	**9.728**
128	256	0.179	**3.046**	**0.447**	**0.667**	**0.772**	9.567
• $\alpha = 256$							
32	256	0.270	3.184	0.437	0.638	0.742	9.294
64	256	**0.152**	3.168	0.443	0.654	0.761	9.486
128	256	0.179	**3.046**	**0.447**	**0.667**	**0.772**	**9.567**

Table 6: Analysis of average latency per instance during the inference phase (in seconds). *Retrieval* refers to the process of transforming text into video, *Generation* indicates the phase of generating motion tokens, and *Decoding* pertains to the conversion of tokens into features using VQ-VAE. It is evident that the retrieval phase does not represent a bottleneck within the entire pipeline, rather, the generation process of LLM constitutes the limiting factor. Notably, the parameter size of the LLM utilized in VimoRAG is 3.8 billion.

	Retrieval Time (s)	**Generation Time (s)**	**Decoding Time (s)**	**Total Time (s)**	**Tokens/s**
MotionGPT-13B	0	14.89	0.12	15.01	13.44
VimoRAG	0.48	7.02	0.12	7.62	27.92

Table 7: Complete results with confidence intervals on the HumanML3D test set.

Model	**Backbone**	**FID ↓**	**R-Precision ↑**			**MM Dist ↓**	**Diversity ↑**
			Top 1	Top 2	Top 3		
• ***Motion Specialists***							
MoMask [13]	–	$0.048^{\pm.004}$	$0.519^{\pm.005}$	$0.715^{\pm.005}$	$0.809^{\pm.004}$	$2.955^{\pm.011}$	$9.632^{\pm.094}$
T2M-GPT [1]	–	$0.112^{\pm.006}$	$0.489^{\pm.006}$	$0.679^{\pm.006}$	$0.774^{\pm.004}$	$3.125^{\pm.015}$	$9.691^{\pm.062}$
MDM [20]	–	$0.454^{\pm.012}$	$0.419^{\pm.004}$	$0.606^{\pm.004}$	$0.712^{\pm.004}$	$3.636^{\pm.015}$	$9.449^{\pm.136}$
MotionDiffuse [2]	–	$0.672^{\pm.025}$	$0.492^{\pm.004}$	$0.685^{\pm.003}$	$0.784^{\pm.003}$	$3.085^{\pm.134}$	$9.499^{\pm.184}$
MLD [22]	–	$0.425^{\pm.145}$	$0.468^{\pm.005}$	$0.656^{\pm.003}$	$0.759^{\pm.004}$	$3.266^{\pm.019}$	$9.698^{\pm.094}$
ReMoDiffuse [6]	–	$0.125^{\pm.142}$	$0.493^{\pm.004}$	$0.676^{\pm.003}$	$0.775^{\pm.003}$	$3.047^{\pm.007}$	$9.211^{\pm.129}$
LMM* [27]	–	$0.040^{\pm.002}$	$0.525^{\pm.002}$	$0.719^{\pm.002}$	$0.811^{\pm.002}$	$2.943^{\pm.012}$	$9.814^{\pm.076}$
MotionLab* [46]	–	$0.167^{\pm-}$	–	–	$0.810^{\pm-}$	$2.912^{\pm-}$	$9.593^{\pm-}$
MotionLCM* [47]	–	$0.304^{\pm.012}$	$0.502^{\pm.003}$	$0.698^{\pm.002}$	$0.798^{\pm.002}$	$3.012^{\pm.007}$	$9.607^{\pm.006}$
MotionCLR* [48]	–	$0.269^{\pm.001}$	$0.544^{\pm.001}$	$0.732^{\pm.001}$	$0.831^{\pm.002}$	$2.806^{\pm.014}$	–
MotionGPT* [4]	–	$0.232^{\pm.008}$	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.002}$	$3.096^{\pm.008}$	$9.528^{\pm.071}$
BiPO* [49]	–	$\mathbf{0.030}^{\pm.002}$	$0.523^{\pm.003}$	$0.714^{\pm.002}$	$0.809^{\pm.002}$	$2.880^{\pm.009}$	$9.556^{\pm.076}$
StableMoFusion* [50]	–	$0.098^{\pm.003}$	$0.553^{\pm.003}$	$0.748^{\pm.002}$	$0.841^{\pm.002}$	–	$9.748^{\pm.092}$
MoGenTS* [51]	–	$0.033^{\pm.001}$	$0.529^{\pm.003}$	$0.719^{\pm.002}$	$0.812^{\pm.002}$	$2.867^{\pm.006}$	$9.570^{\pm.077}$
LAMP* [52]	–	$0.032^{\pm.002}$	$\mathbf{0.557}^{\pm.003}$	$\mathbf{0.751}^{\pm.002}$	$\mathbf{0.843}^{\pm.001}$	$\mathbf{2.759}^{\pm.007}$	$9.571^{\pm.069}$
• ***Motion LLMs***							
MotionGPT-2* [25]	Llama3-8B	$0.191^{\pm.004}$	$0.496^{\pm.002}$	$0.691^{\pm.003}$	$0.782^{\pm.004}$	$3.080^{\pm.013}$	$9.860^{\pm.026}$
MotionLLM* [53]	GPT4+Gemma-2B	$0.230^{\pm.009}$	$0.515^{\pm.004}$	–	$0.801^{\pm.004}$	$2.967^{\pm.020}$	$\mathbf{9.908}^{\pm.102}$
*Wang et al.** [54]	Llama2-13B	$0.166^{\pm-}$	$0.519^{\pm-}$	–	$0.803^{\pm-}$	$2.964^{\pm-}$	–
ScaMo* [55]	codesize 512-3B	$0.617^{\pm-}$	$0.443^{\pm-}$	$0.627^{\pm-}$	$0.734^{\pm-}$	$3.340^{\pm-}$	$9.217^{\pm-}$
AvatarGPT* [56]	Llama-13B	$0.567^{\pm-}$	$0.389^{\pm-}$	$0.539^{\pm-}$	$0.623^{\pm-}$	–	$9.489^{\pm-}$
MotionGPT* [3]	Llama-13B	$0.567^{\pm-}$	–	–	–	$3.775^{\pm-}$	$9.006^{\pm-}$
MotionGPT [3]	Phi3-3.8B	$0.501^{\pm.005}$	$0.396^{\pm.002}$	$0.575^{\pm.005}$	$0.673^{\pm.004}$	$3.724^{\pm.012}$	$9.475^{\pm.110}$
VimoRAG (Ours)	Phi3-3.8B	$0.131^{\pm.007}$	$0.452^{\pm.002}$	$0.655^{\pm.006}$	$0.764^{\pm.005}$	$3.146^{\pm.011}$	$9.424^{\pm.149}$

A.2 Metrics

For motion generation evaluation metrics, we adopt several widely recognized measures: *Frechet Inception Distance* (FID), which quantifies generation fidelity by measuring the distributional distance between the generated motions and reference motions in feature space. *R-recall*, *MultiModal Distance* (MM Dist), which evaluate the semantic consistency between text and motions. *Diversity*. which assesses the diversity of the generated motions corresponding to a given textual input. For text-to-video retrieval metrics, we adopt the widely used metrics retrieval *recall* (R@1, R@5, R@10 are adopted in this paper), *Median Rank* (MdR) and *Mean Rank* (MnR). Recall measures the proportion of relevant results returned by the retrieval model within the top-k results. Median Rank represents the middle value of the ranks at which the correct results appear in the retrieval list. Mean Rank calculates the average position of the correct results in the retrieval lists.

A.3 More Details of Datasets

IDEA400 represents a high-quality subset of Motion-X [11], consisting of a large-scale whole-body motion dataset composed of 12.5K clips and 2.6M frames. This dataset encompasses a diverse array of gestures and detailed pose descriptions. Unlike existing works [16], which meticulously select a

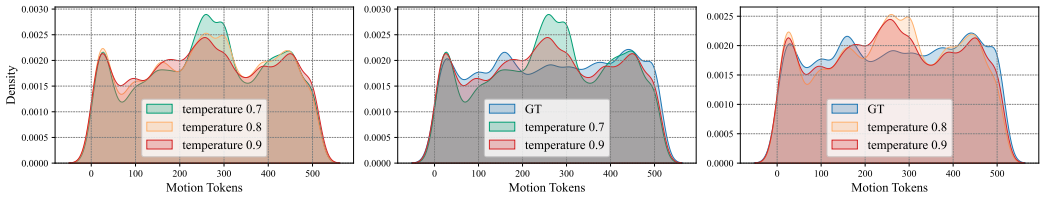

Figure 9: The impact of varying temperature values on the distribution of generated motion tokens. Notably, when the temperature is set to 0.9, the distribution of the generated motion tokens closely resembles the ground truth (GT) distribution.

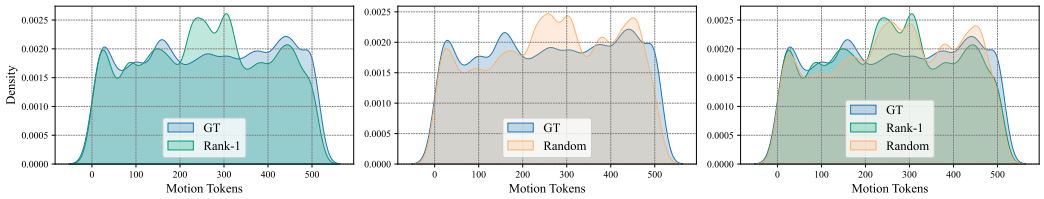

Figure 10: The influence of different retrieval conditions on the distribution of generated motion tokens is depicted.

test set with text descriptions similar to those found in HumanML3D, we adopt a different approach by randomly sampling 10% of the clips without any specific selection criteria. We argue that this methodology more closely resembles an out-of-distribution (OOD) scenario, wherein the test set features a distribution distinctly different from that of the HumanML3D training set.

HumanML3D [5] constitutes the largest dataset available, providing text descriptions alongside body-only motions. It includes a total of 14,616 motion clips and 44,970 text descriptions, with 5,371 unique words present across these descriptions. The dataset is partitioned into a training set (80%), a validation set (5%), and a test set (15%).

Regarding pose representation, we adhere to the same specifications as outlined in HumanML3D [5]. It is important to note that the number of joints (J) is set at 22 for both HumanML3D and IDEA400. We process the motion features for IDEA400 using the same settings as HumanML3D, resulting in a total dimension of 263. These features include root angular velocity, root linear velocities, root height, local joint positions (velocities), and 6D rotations.

A.4 More Experimental Results

The Effect of Video Database Size. To investigate the impact of video database size on performance, we conduct experiments at six different scales using the validation set from the HumanML3D dataset. Figure 7 illustrates the changes in the *Frechet Inception Distance* (FID) and *Multi-Modal Distance* (MM-Dist) metrics as the size of the database increases, as discussed in Section 4.3.

We also report the changes in R-Precision and diversity metrics under the same experimental settings. As illustrated in Figure 8, we observe that a larger database contributes to an improvement in R-Precision, particularly when the database size exceeds 100,000 entries, highlighting the advantages of utilizing large-scale video retrieval repositories. Interestingly, we find that the diversity metric does not increase in tandem with the growth of the video database size. We hypothesize that when the retrieved videos lack informativeness, the noise introduced enhances the diversity of the model's outputs. Consequently, smaller video databases can still yield substantial diversity.

Through this experiment, we demonstrate the significant potential of retrieval-enhanced methods based on large-scale video databases. However, resource constraints limit us to conduct further experiments. We hope to explore the effects of even larger databases on performance in future work.

The Analysis of Latency. We analyze the latency of the VimoRAG framework, as depicted in Table 6, where we report the average time taken per instance during the inference phase. The entire pipeline is divided into three stages: the retrieval phase (conducted by the text-to-video model), the

Figure 11: Words and phrases that frequently appear in text descriptions in HcVD database.

generation phase (performed by the LLM), and the decoding phase (executed by the VQ-VAE). The results indicate that the generation phase exhibits the highest latency, accounting for 92% of the total processing time. This is primarily due to the need for the LLM to load a substantial number of parameters (3.8 billion in VimoRAG). In comparison, the retrieval and decoding phases account for only 6% and 2% of the total time, respectively. This finding highlights that, despite the generation of high-fidelity motion leveraging the world knowledge of the LLM, it remains the latency bottleneck in the overall generation framework. Consequently, this drives us to explore more efficient generative models in our future work. To illustrate the impact of different LLM sizes on latency, we also test the latency of the MotionGPT-13B model [3] on the same computational hardware. As shown in Table 6, the total time for the MotionGPT-13B is nearly twice that of VimoRAG.

Hyper-parameters of LoRA in McDPO. In order to investigate the impact of key hyper-parameters on the performance of McDPO training, we conduct comparative experiments focusing on the LoRA fine-tuning parameters $rank$ and α. As illustrated in Table 1, varying $rank$ and α significantly influence the outcomes, a phenomenon that aligns with previous findings in the works [3, 25]. We also observe that when maintaining a constant ratio between $rank$ and α, both the FID and MultiModal Dist metrics gradually decrease as $rank$ and α increase. The increase in $rank$ and α corresponds to a greater number of parameters available for optimization within the model, thereby enhancing its capability to fit the dataset more effectively. However, we note a corresponding decline in diversity, indicating that a larger parameter scale adversely affects the model's ability to generate diverse outputs.

The Impact of the Temperature of LLM. As shown in the Figure 9, we visualize the impact of different temperatures on the distribution of generated motion tokens. We examine the effects of three commonly used temperatures (0.7, 0.8, and 0.9) on the results. Firstly, it is evident that the output distributions generated by different temperatures exhibit differences, indicating that the model is sensitive to this hyperparameter. Additionally, we observe that the distribution at a temperature of 0.9 is closest to the ground truth distribution. Generally, as the temperature parameter increases, the output diversity of the model also rises. We believe that a larger temperature in this study aids in enhancing the model's generalization capability.

Analysis of Distribution Resulting from the Retrieval Process. As shown in Figure 10, we visualize the distribution of generated motion tokens under two conditions: one utilizing rank-1 videos during inference and the other using random videos. The figure demonstrates that the distributions of motion tokens obtained from random videos and rank-1 videos are generally comparable on a macroscopic level. This indicates that even when the retrieval system fails, the distribution of our model's generated results remains relatively close to the distributions obtained from rank-1 retrieval and the ground truth distribution from a macroscopic perspective. This demonstrates the model's strong fault tolerance and robustness. Notably, when the motion token index is less than 200, the

Table 8: Details of the resources utilized in the construction of the HcVD database.

Dataset	Number of Videos	Tasks
UCF101 [32]	13,320	Action Recognition
NTU RGB+D [33]	114,480	Action Recognition
ASLAN [28]	3,697	Action Similarity Labeling
HMDB51 [29]	6,849	Human Motion Recognition
Kinetics-400 [30]	306,245	Human Action Classification
PennAction [31]	2,326	Action Classification, Action Detection
MotionX [11]	32,500	Human Mesh Recovery, Human Mesh Generation

generated distribution under the rank-1 condition closely aligns with the ground truth distribution, indicating the effectiveness of this retrieval strategy in capturing motion characteristics.

B More Qualitative Results

To comprehensively evaluate the generation performance of VimoRAG in out-of-distribution (OOD) scenarios, we present more visualization results on the IDEA400 dataset in Figures 12 and 13. It is noteworthy that the model utilized for generation is trained solely on the HumanML3D training set. We also present the retrieved videos on the right side of each case. The cases depicted in Figure 5 are further illustrated in these two figures. In contrast, we provide full-text descriptions in these figures. It is evident that the generated motions align with the text descriptions, aided by the retrieved videos. Taking a complex description, *"The person is simulating chopping wood while seated. They rotate their torso, simulate grabbing and lifting an object with their hands, bring it overhead, and then perform a striking motion downward as if impacting a log between their legs. This action is repeated, emulating the motion of splitting wood with an axe."* in Figure 12, as an example, VimoRAG is capable of generating such uncommon actions and seamless transitions between movements. As a supplement, we have showcased some of the original videos in our anonymous repository.

C Details of Human-centric Video Database

To train our retrieval models, we annotate a text description for each video using the widely utilized LMM, Qwen2-VL-7B-Instruct. In total, we synthesize 425,988 captions. It is important to note that we use these text captions exclusively during the training phase of Gemini-MVR and do not employ them in any retrieval processes within VimoRAG. This means that VimoRAG also performs effectively with another large video database during the inference stage. As illustrated in Figure 11, the word cloud presents the diverse types of actions included in the HcVD database. This richness and diversity also elucidate why VimoRAG achieves exceptional performance from a different perspective. We present more details of the resources that are used for the construction of the HcVD in Table 8. Qwen2-VL supports dynamic frame selection (set FPS as 2.0). The prompt we used for data synthesis is "Please describe the person's actions in the video using a single sentence that contains a series of verbs.".

D More Implementation Details of VimoRAG

To enhance the reproducibility of VimoRAG, this section presents additional model details, which are also available in our code repository.

More Details of the Gemini-MVR. The temporal encoder consists of 4 transformer layers, each featuring 12 attention heads and a width of 768. We utilize learnable position embeddings with a context length of 77 in the action encoder. For the keypoints encoder, a projection layer is placed atop MotionBERT [38], with the input and output channels of the projection layer set at 8704 and 768, respectively. To derive the final representation of the keypoints, we implement mean pooling over all the encoded frames, in accordance with existing works [40, 39]. The similarity integrator is implemented as a linear transformation, with an input channel of 768 and an output channel of 2. The

$\mathcal{L}_{a2p}$ is defined as follows:

$$\mathcal{L}_{a2p} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(s(\mathbf{a}_i,\mathbf{p}_i))}{\sum_{j=1}^{B}\exp(s(\mathbf{a}_i,\mathbf{p}_j))}\,. \tag{6}$$

More Details of the Generation Model. In our framework, we employ the VFM InternVideo2 [57] as the video encoder and CLIP-large [37] as the image encoder following *Maaz et al.* [10]. Specifically, we adopt the `InternVideo2-Stage2_1B-224p-f4` variant of InternVideo2 and the `CLIP-ViT-L/14-336` model. The visual projector is a two-layer MLP with a hidden size of 1024, where a GELU [58] activation function is applied after the first linear layer. For the LoRA parameters, we configure the rank to 128 and set alpha to 256. For the model details and training configurations of VQ-VAE, we adopt the same settings as those used in existing works [1, 3]. Specifically, the codebook size is set to 512x512. The temporal downsampling rate is set to 3 in the encoder of VQ-VAE.

More Details of the McDPO Dataset. Firstly, we use the π_{ref} model obtained from Stage 1 to sample k times (where $k = 3$) on a random 25% subset of the training set. The reason for extracting this subset is to reduce inference costs (which is a similar approach used by *Zhang et al.* [44]). For each input sample, we obtain k different outputs as candidate samples. To identify positive and negative samples from these multiple outputs, we utilize Equation 4 to calculate the reward scores and select the cases with the highest and lowest scores as the positive and negative samples, respectively. We configure $w_\ell = 0.9$ and $w_d = 0.1$ in Equation 4 in our experiments.

Figure 12: Additional visualization results on the IDEA400 dataset (**Part** I).

Figure 13: Additional visualization results on the IDEA400 dataset (**Part** II).