VimoRAG: Video-based Retrieval-augmented 3D Motion Generation for Motion Language Models

Project: https://walkermitty.github.io/VimoRAG/

Paper: https://arxiv.org/abs/2508.12081

Code: https://github.com/WalkerMitty/VimoRAG

**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. HITSZ 2. UM. 3. Surrey
4. NJU 5. NTU 6. NUS

Motivation

How to Address OOD/OOV Challenges in Motion Generation?

- Video-based retrieval-augmented generation for motion LLMs
  - Why videos? Massive, scalable, and rich in human motion—far beyond small 3D datasets
  - 2D videos and 3D motions share semantic and structural cues (e.g., pose, dynamics)

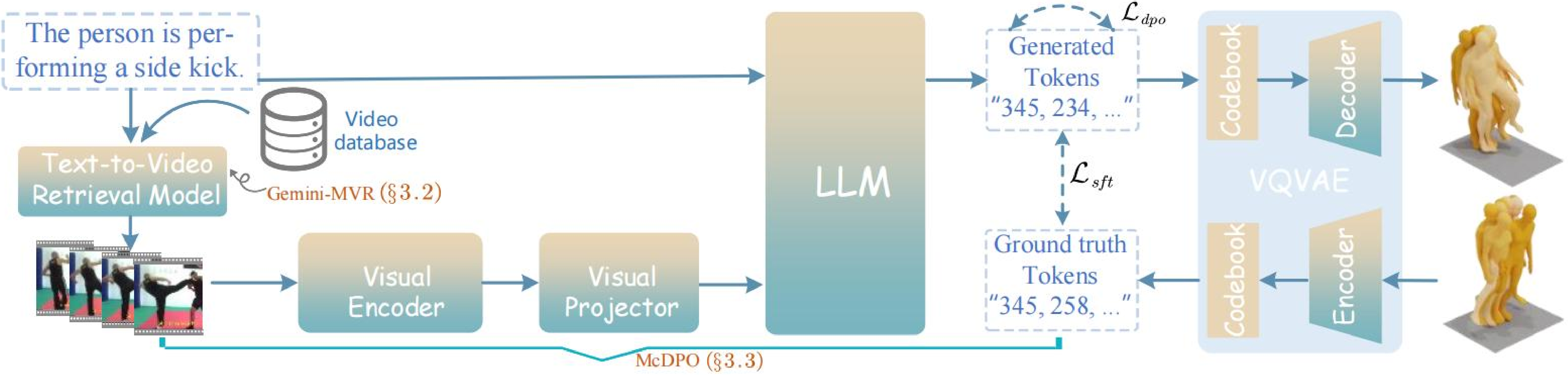
Challenges

- Pretrained VFMs perform poorly on action-level retrieval
- Noisy or misaligned videos cause error propagation in generation

Method

Overview of VimoRAG

- Gemini-MVR: An effective model for text-to-motion video retrieval
- McDPO: A training strategy to mitigate error propagation in retrieval-augmented generation



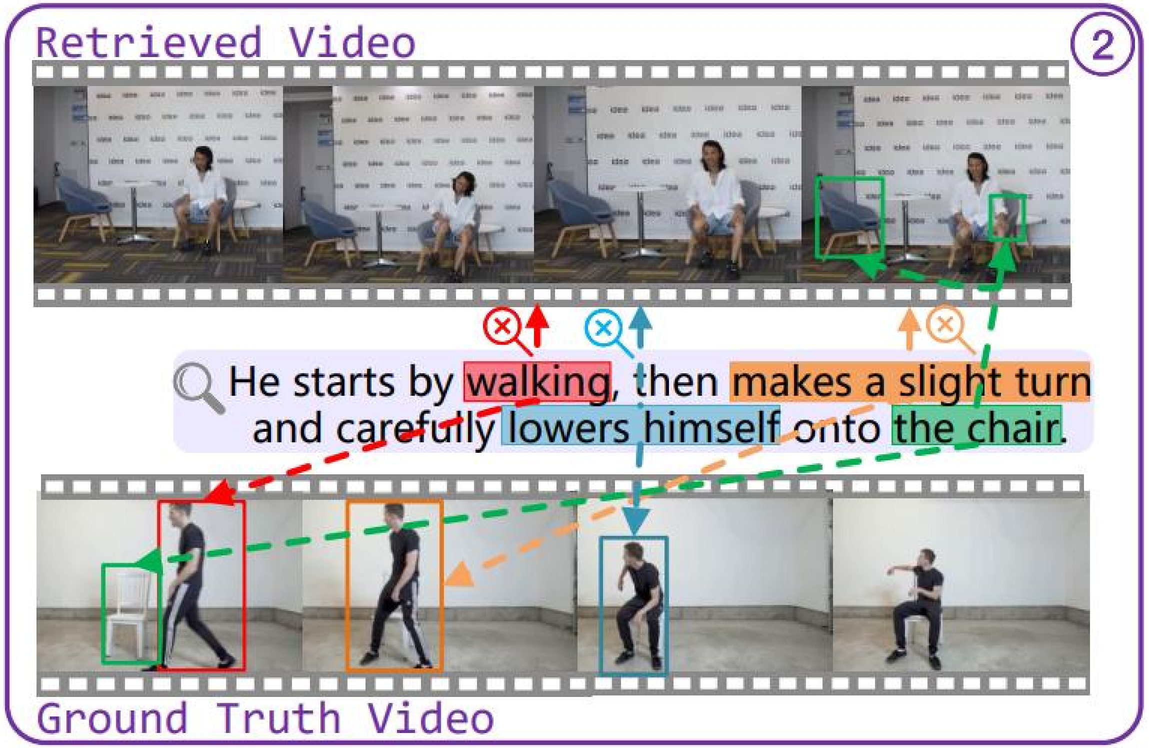
Method

Gemini-MVR

- Motivation

Pretrained VFM (e.g., InternVideo):

Weak at distinguishing human poses, but sensitive to objects/entities.


Retrieved Video
2
He starts by walking, then makes a slight turn and carefully lowers himself onto the chair.
Ground Truth Video

Method

Gemini-MVR

- ➢ Dual retrieval branches:

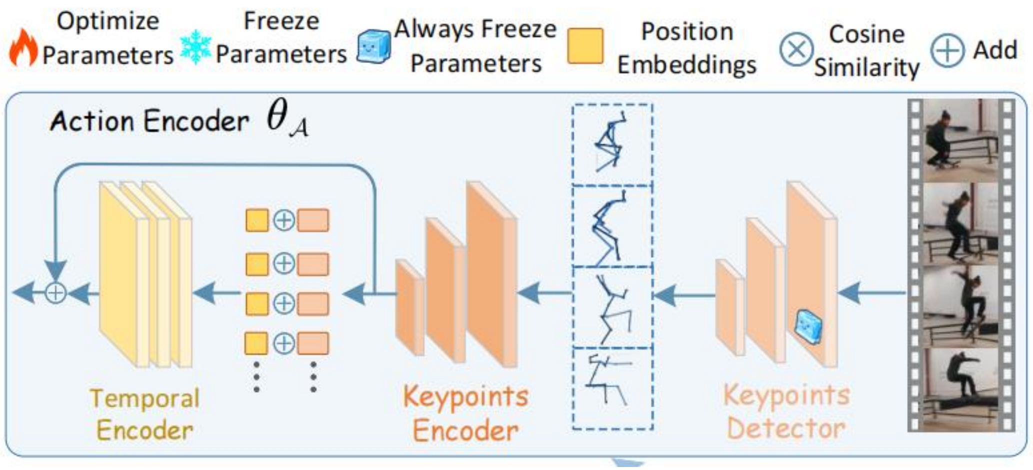
  Action-level + object-level

- ➢ Lightweight router:

  Dynamically fuses the two streams

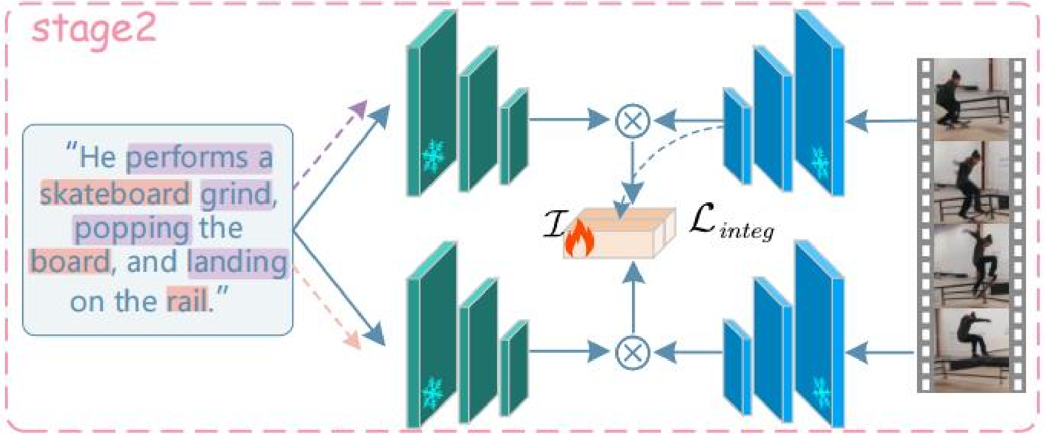
- ➢ Action encoder:

  Pose-aware design for fine-grained motion



predicate semantic extractor

action encoder

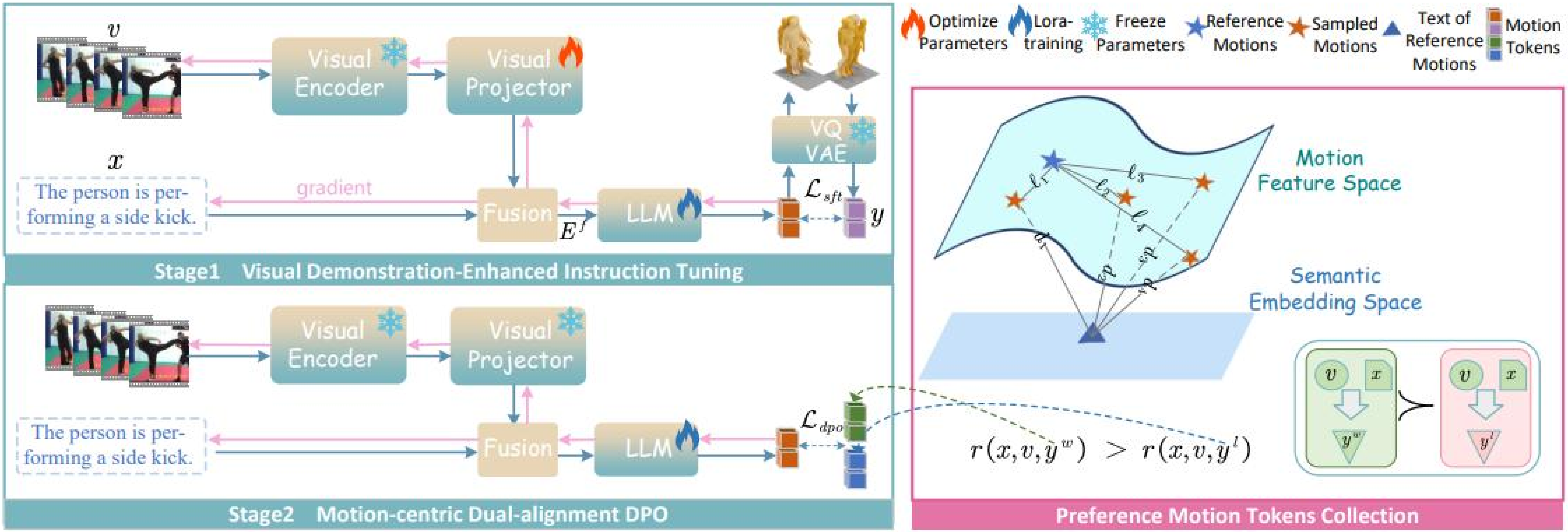

stage2
"He performs a skateboard grind, popping the board, and landing on the rail."
$\mathcal{I}$
$\mathcal{L}_{integ}$

argument semantic extractor

object encoder

Method

Motion centric dual-alignment DPO (McDPO)



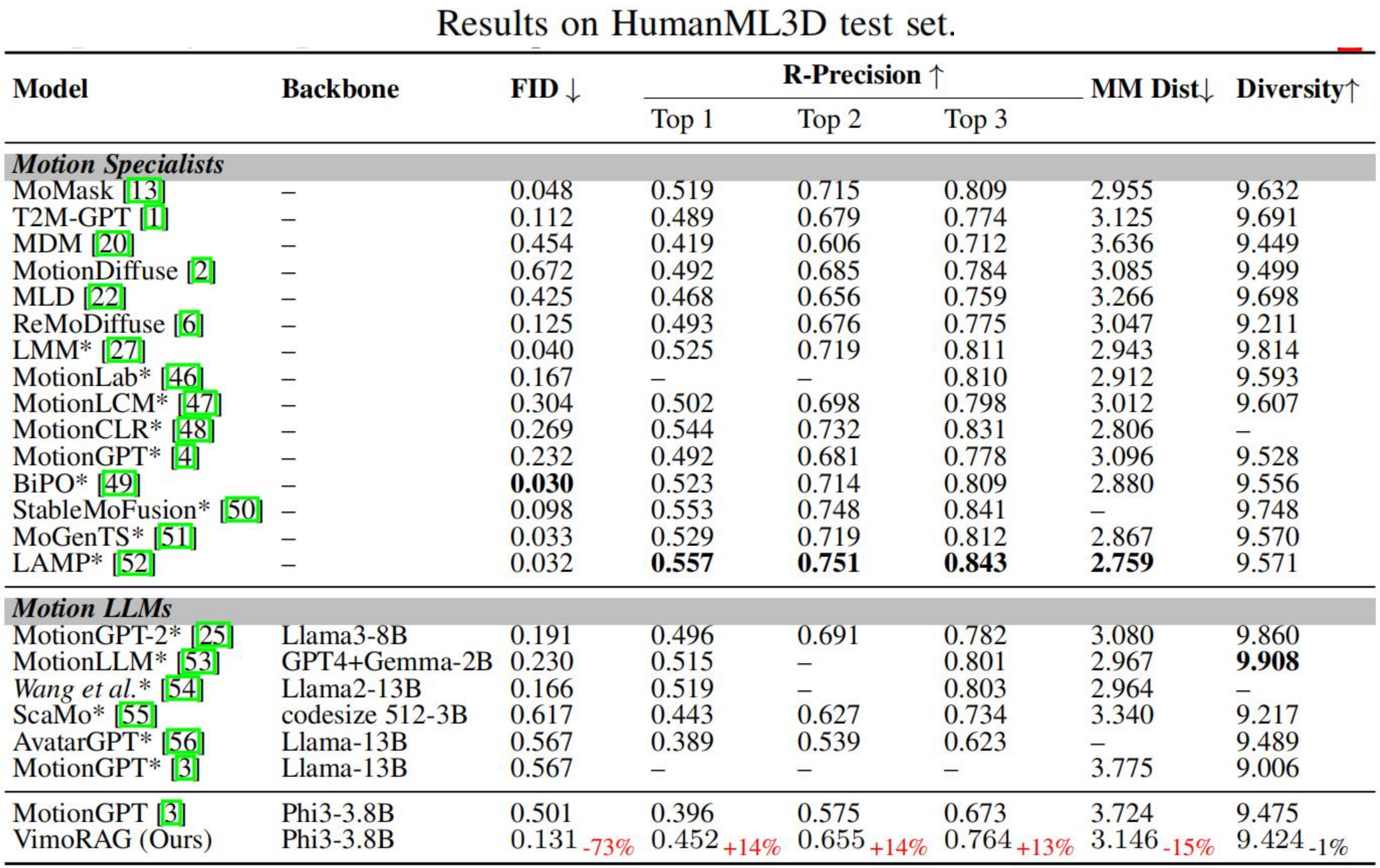
Reward Model:

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

Experiment

In-domain

- Outperforms MotionGPT by a large margin under the same backbone
- achieves the best FID score among motion LLMs



Results on HumanML3D test set.

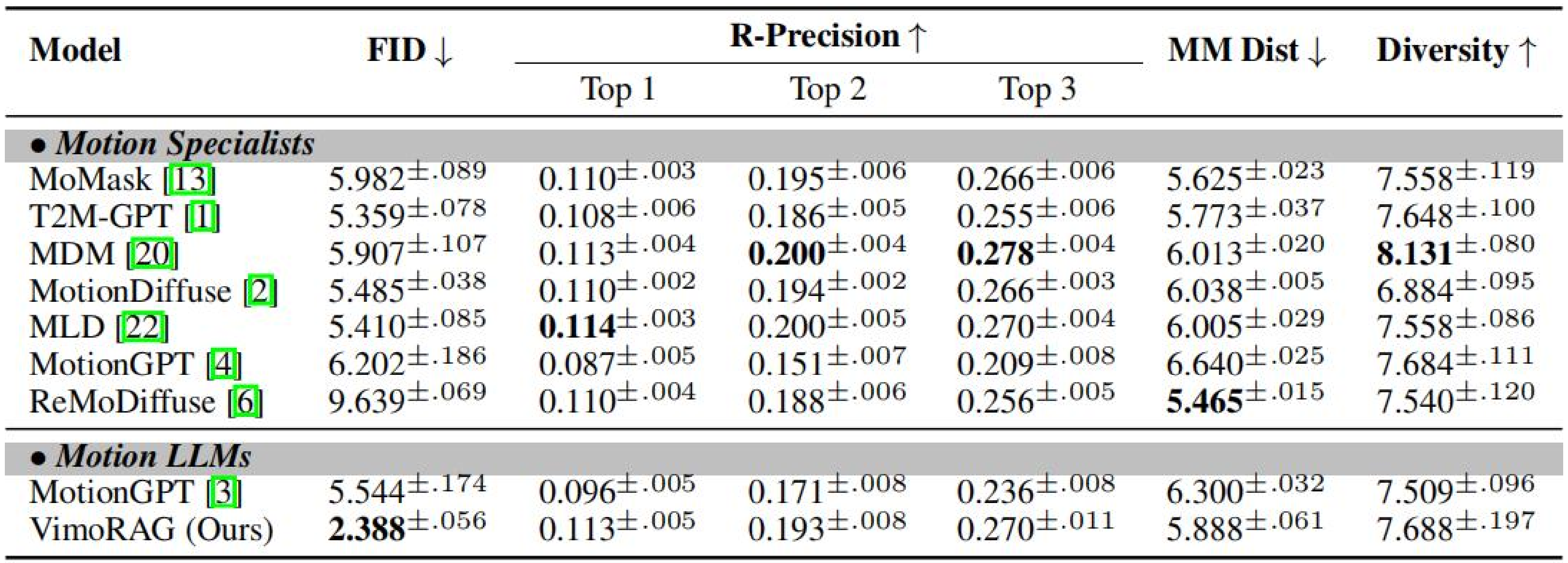
Model	Backbone	FID ↓	R-Precision ↑			MM Dist↓	Diversity↑
			Top 1	Top 2	Top 3		
Motion Specialists							
MoMask [13]	–	0.048	0.519	0.715	0.809	2.955	9.632
T2M-GPT [1]	–	0.112	0.489	0.679	0.774	3.125	9.691
MDM [20]	–	0.454	0.419	0.606	0.712	3.636	9.449
MotionDiffuse [2]	–	0.672	0.492	0.685	0.784	3.085	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%

Experiment

Out-of-domain

- VimoRAG achieves the best FID score, with other metrics closely matching SoTA

Zero-shot results on IDEA400 test set.

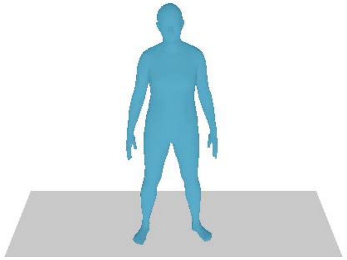


Model	FID ↓	R-Precision ↑			MM Dist ↓	Diversity ↑
		Top 1	Top 2	Top 3		
● ***Motion Specialists***						
MoMask [13]	5.982$^{\pm.089}$	0.110$^{\pm.003}$	0.195$^{\pm.006}$	0.266$^{\pm.006}$	5.625$^{\pm.023}$	7.558$^{\pm.119}$
T2M-GPT [1]	5.359$^{\pm.078}$	0.108$^{\pm.006}$	0.186$^{\pm.005}$	0.255$^{\pm.006}$	5.773$^{\pm.037}$	7.648$^{\pm.100}$
MDM [20]	5.907$^{\pm.107}$	0.113$^{\pm.004}$	**0.200**$^{\pm.004}$	**0.278**$^{\pm.004}$	6.013$^{\pm.020}$	**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$^{\pm.085}$	**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$^{\pm.005}$	**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)	**2.388**$^{\pm.056}$	0.113$^{\pm.005}$	0.193$^{\pm.008}$	0.270$^{\pm.011}$	5.888$^{\pm.061}$	7.688$^{\pm.197}$

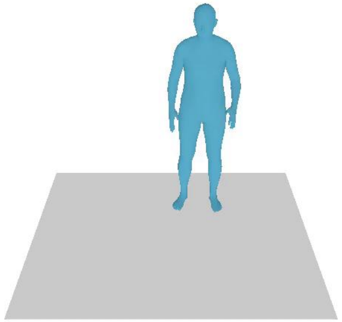
Experiment

Visualization (zero-shot)

The person is standing upright with a rapid sequence of **raising both fists from waist level to above the head** and then **lowering them back down** in a cheering motion.



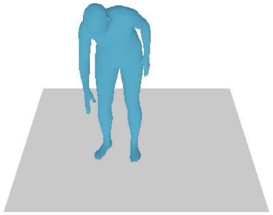
The person appears to be **mimicking the action of riding a bicycle** while standing up; alternating raising knees as if pedaling, and swinging arms as though holding handlebars.



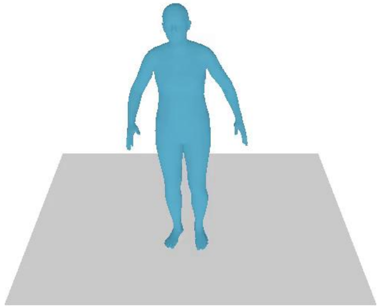
Experiment

Visualization (zero-shot)

The person is bending over to **put food on the floor for a pet**, then straightening up and stepping back to standing position.



The person is **preparing to throw a frisbee**. Starting with a stance where the weight is on the back foot, they shift the weight forward, bringing the arm with the frisbee back for momentum. Then, they step forward with the opposite leg, rotating the torso and extending the arm to release the frisbee.

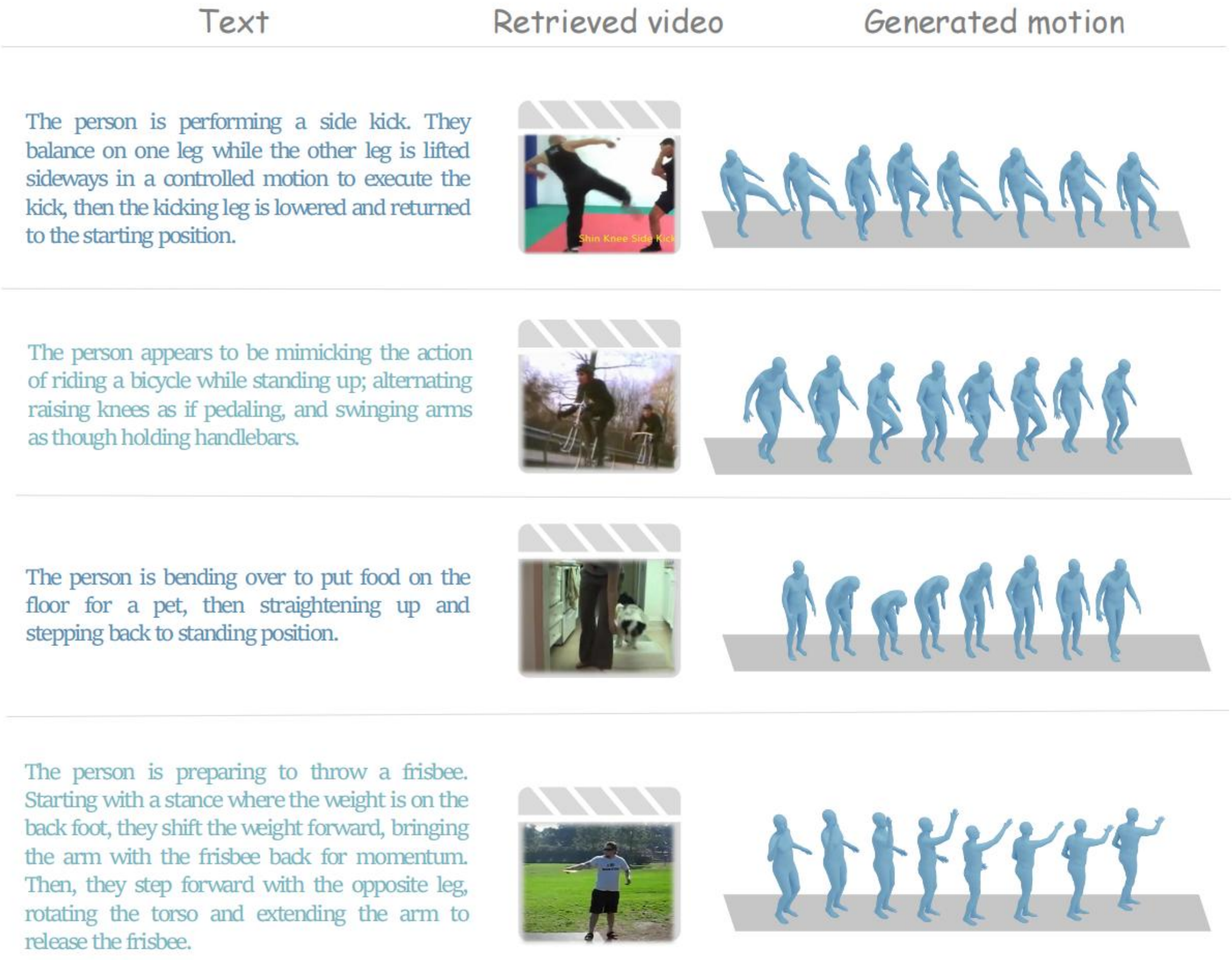


Experiment

NEURAL INFORMATION PROCESSING SYSTEMS

Visualization (zero-shot)

More results can be found:
https://walkermitty.github.io/VimoRAG/



The person is performing a side kick. They balance on one leg while the other leg is lifted sideways in a controlled motion to execute the kick, then the kicking leg is lowered and returned to the starting position.

The person appears to be mimicking the action of riding a bicycle while standing up; alternating raising knees as if pedaling, and swinging arms as though holding handlebars.

The person is bending over to put food on the floor for a pet, then straightening up and stepping back to standing position.

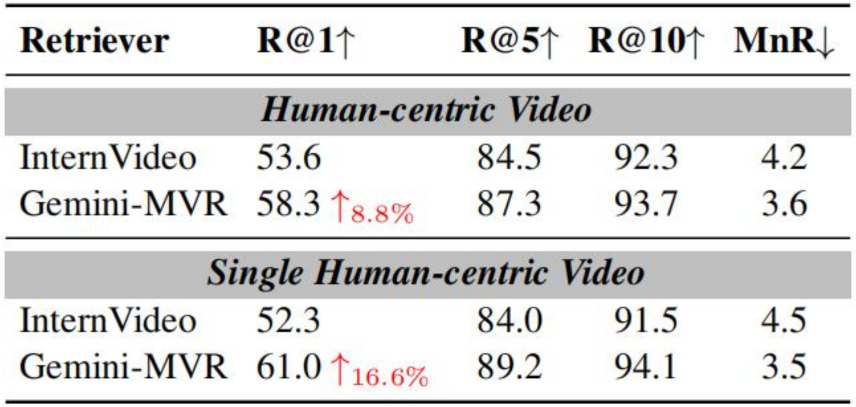
The person is preparing to throw a frisbee. Starting with a stance where the weight is on the back foot, they shift the weight forward, bringing the arm with the frisbee back for momentum. Then, they step forward with the opposite leg, rotating the torso and extending the arm to release the frisbee.

Experiment

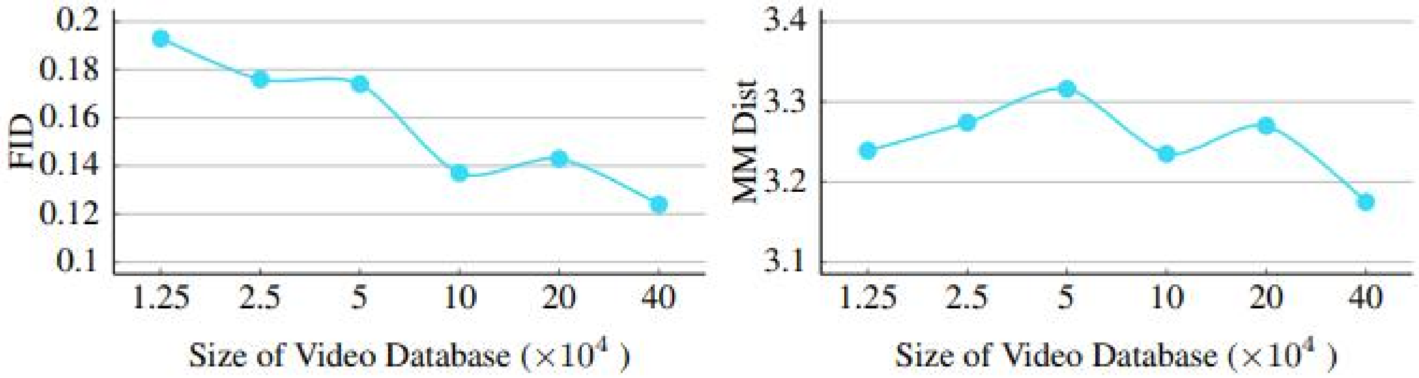
Discussion

- ➢ Compared to the object-level VFM, Gemini-MVR achieves a significant improvement in the Recall@1 metric

- ➢ As the video retrieval database grows, VimoRAG shows steadily improving performance



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



Ending

NEURAL INFORMATION PROCESSING SYSTEMS

Project: https://walkermitty.github.io/VimoRAG/

Paper: https://arxiv.org/abs/2508.12081

Code: https://github.com/WalkerMitty/VimoRAG

Contact us： 182haidong@gmail.com

Thanks for your time