

1 LLM Rewriting

pseudo supervision

Standard Caption



LLM Rewrite

Concise Caption

Fine-grained Caption



Frozen



Trainable



Qwen2.5-7B-Instruct



Caption

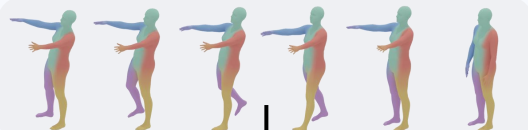


Embedding

2 Granularity-Aware Training



A person raises both hands while walking forward.



Text Encoder

ViT-based Motion Encoder

Standard Projection

Concise Adapter

Concise Extractor

Standard Projection

Fine Extractor

Fine Adapter

Standard Projection

Concise Adapter

Concise Extractor

Standard Projection

Fine Extractor

t_{fine}

t_{std}

t_{con}

m_{con}

m_{std}

m_{fine}

InfoNCE (t_{con}, m_{con})

(anchor)

InfoNCE (t_{fine}, m_{fine})

3 Granularity-Aware Inference



A person walks awkwardly across space with their arms held out to the sides for balance.

select branch by granularity

Concise

standard + concise

Standard

standard branch

Fine

standard + fine



Score Calibration

Calibrated Ranking

Calibrated Ranking

Top-ranked Motions

