IMoRe: Implicit Program-Guided Reasoning for Human Motion Q&A (Supplementary Material)

Comparison with video-language models trained on large-scale data. We finetuned Qwen-2.5-VL-3B, InternVL2.5-4B-MPO and MiniCPM-V2.6 on the BabelQA dataset where the motion sequences are converted to videos consisting of skeleton images. Program and answer set information are provided in the prompt for fair comparison. As shown in Table 1, IMoRe significantly outperforms all VLM baselines. We attribute this to: (1) VLMs struggle to capture fine-grained temporal and spatial concepts from motion image sequences, and (2) their limited ability to leverage structured programs during reasoning.

	Overall	Action	Direction	Bodypart
IMoRe I	0.609	0.652	0.622	0.373
Qwen-2.5-VL-3B	0.425	0.467	0.333	0.350
InternVL2_5-4B	0.402	0.433	0.306	0.383
MiniCPM	0.384	0.410	0.306	0.367

Table 1. Comparison with VLMs

Additional ablation study. (a) Comparison between question and program for text-aware feature. We fuse the motion feature and the text feature to obtain the text-aware feature as described in Sec. 3.3 of the main paper. To verify this design, we show results using programs for text-aware motion feature (IMoRe w Pro) in Table 2. We can see that the performance drops slightly, likely because the question texts better guide token-level attention in ViT-derived motion features, improving the alignment between text and motion. (b) Question type information. We also ablate over the question type information as described in Sec 3.3 of the main paper. We can see that the performance without question type (IMoRe wo QT) in Table 2 slightly drops.

	Overall	Action	Direction	Bodypart
IMoRe I	0.609	0.652	0.622	0.373
IMoRe w Pro	0.598	0.690	0.583	0.217
IMoRe wo QT	0.603	0.655	0.596	0.254

Table 2. Additional ablation study

Visualization of feature level selection. We have compared single-level and multi-level feature selection quantitatively in the ablation study (C vs D setting in Table 3 of main paper). The improvement from C to D verifies the effectiveness of the multi-level based feature selection. We further show a qualitative attention map between program functions (y-axis) and multi-level features (x-axis) in Fig. 1. It can be seen that **filter_action()** attends to high-level feature (Feat 6) while **query_body_part()** to low-level features (Feat 0 and 1).

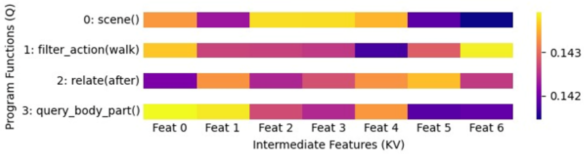

Figure 1. Visualization of feature level selection.

Visualization of concept localization. Our approach retains interpretability in two ways despite the implicit program guidance. (1) Program structure: structured programs define an explicit step-by-step reasoning path, where each function (*e.g.*, filter, relate) defines a specific operation. Although the reasoning is implemented implicitly, our model follows this program-defined sequence, making each reasoning step directly attributable to a specific symbolic instruction. (2) Intermediate traceability: As shown in Fig. 2, the attention map between program functions (y-axis) and motion segments (x-axis) reveals interpretable reasoning. After initialization at step 0, **filter_action(crawl)** at step 1 correctly attends to segment 2 (where the action occurs), thereafter attention for **relate(after)** and **query_action()** shifts towards segment 3, illustrating stepwise execution consistent with the program.

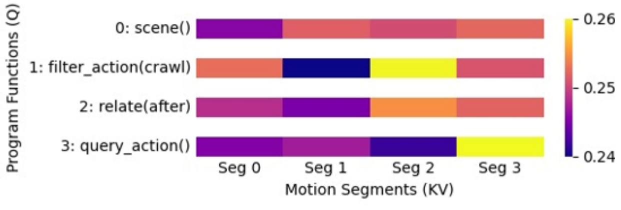

Figure 2. Visualization of concept localization.