

MotionHalluc

Diagnosing Kinematic Hallucinations in Fine-Grained Motion Reasoning

Weile Guo¹ Shenghong He¹ Danying Mo¹
Chengdong Xu¹ Xuexun Liu² Chao Yu^{1*}

¹ Sun Yat-Sen University ² Shenzhen University

* Corresponding author

Motion feedback DO NOT match the video pair

A query video shows the motion to correct. A reference video provides the target.

DH Q) Is correction accurate to improve motion in Video1 based on Video2?
C) "Keep knees slightly bent. "
A. Yes B. No

Video1
Query



Video2
Reference



Gemini: In Video 1, the person keeps his knees very straight. In Video 2, the person maintains a slight, soft bend in the knees. Therefore, answer is A.

Motion feedback must match the actual kinematic difference

Models can confuse the query and reference motions, giving corrections in the wrong direction.

Can models verify the actual motion difference?

Three diagnostics for motion hallucinations

Directional: 624 QA Attributional: 600 QA Temporal: 316 QA

Directional Hallucination

Question: Is the following correction accurate and necessary to improve the query motion in Video1 based on the reference motion in Video2?

Correction: "Keep your torso more upright."

Original



A. Yes B. No

Reversed (Exchange V1 and V2 in the Origin)

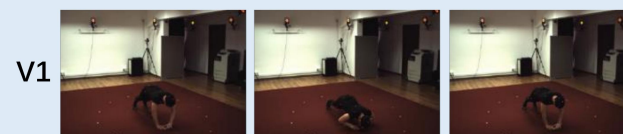
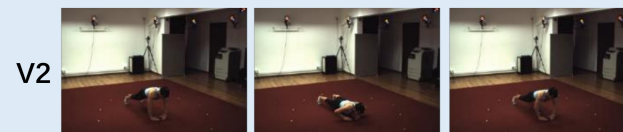


A. Yes B. No

Attributional Hallucination

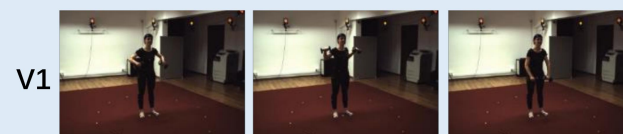
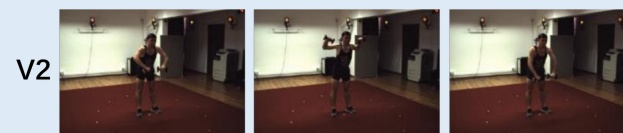
Question: Which of the following correction accurate and necessary to improve the query motion in Video1 based on the reference motion in Video2?

Same Joint



- A. "Make body closer to the ground."
B. "Keep your body farther from the ground."

Different Joints

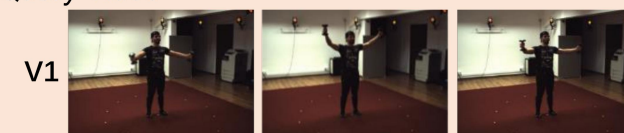


- A. "Lean the body slightly backward."
B. "Ensure upper arms parallel to the floor."

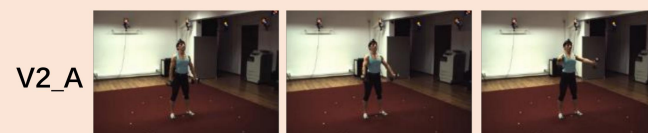
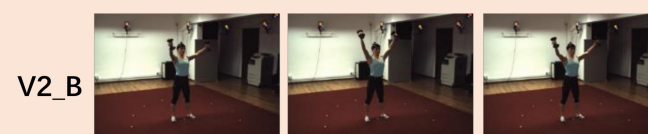
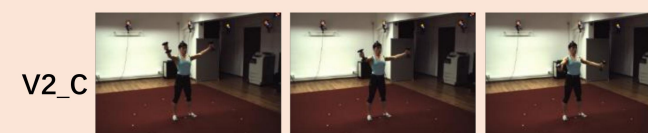
Temporal Hallucination

Question: You are given a clip from Video1 that captures a specific phase of an action. Video2 provides three candidate clips. Which clip from Video2 corresponds to the same action phase as Video1?

Query Video



Candidate Videos



A. Clip A B. Clip B C. Clip C

MotionHalluc benchmark construction

553

video pairs

32

action categories

1,540

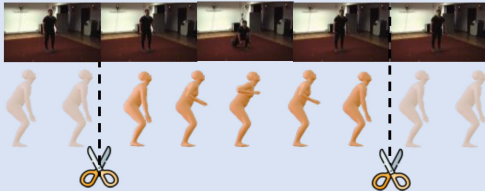
diagnostic questions

1. Data Collection

Fit3D Dataset

- ✓ 47 exercise categories
- ✓ Multi-view videos
- ✓ Motion capture

Automatic segment

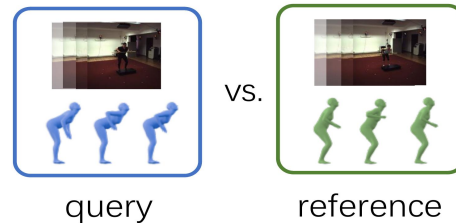


Selection & Filtering

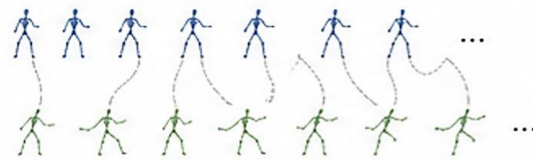
- ✓ Select 32 actions
- ⚡ Keep 1 repetition per video
- ⚙️ Diverse clips w/o near-duplicate repetitions

2. Video Pairing

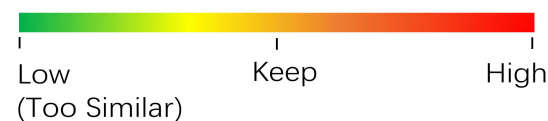
Automatic pairing



DTW alignment on motion

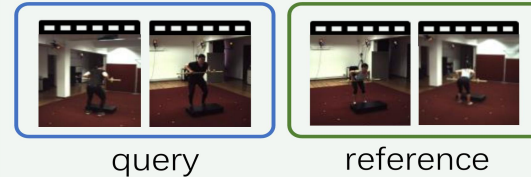


Motion Similarity



3. Human Annotation

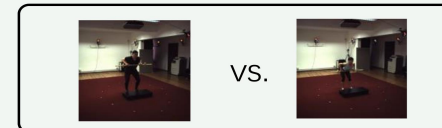
Annotation on multi-view videos



- 👁️ Grounded on observable physical discrepancy
- 🚫 Exclude speculative or non-observable factors

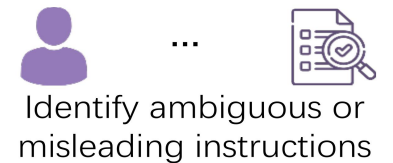
Instructions & Timestamps

- Keep back straight
- ...



4. Data Curation

Stage1: Full review



Stage2: Revision

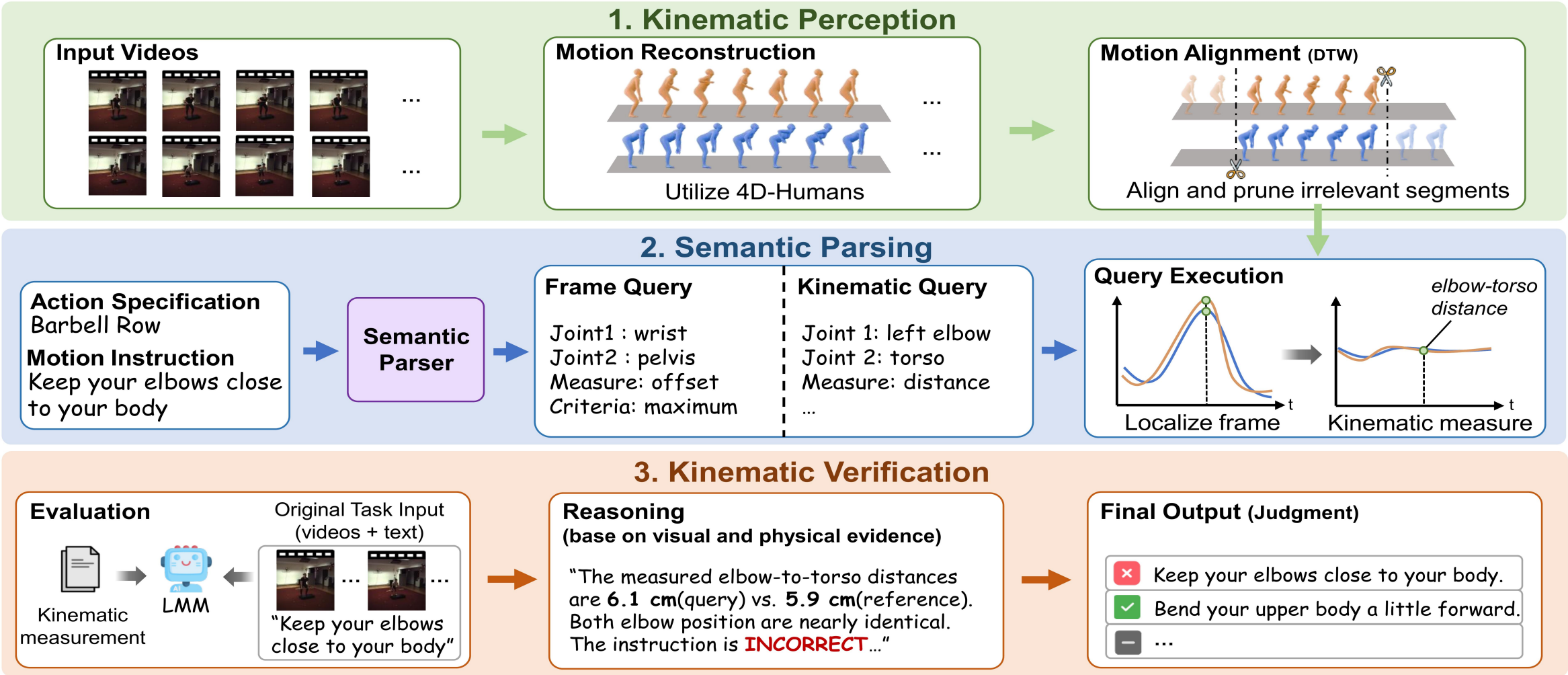
- ✍️ Revise ambiguous annotations
- 📊 Correct 6.5% annotations

MotionHalluc

- 📊 553 pairs
- 📄 2,673 annotations
- 🎥 Multi-view videos
+22-joint skeleton
+Corrective instructions
+Frame-level timestamps

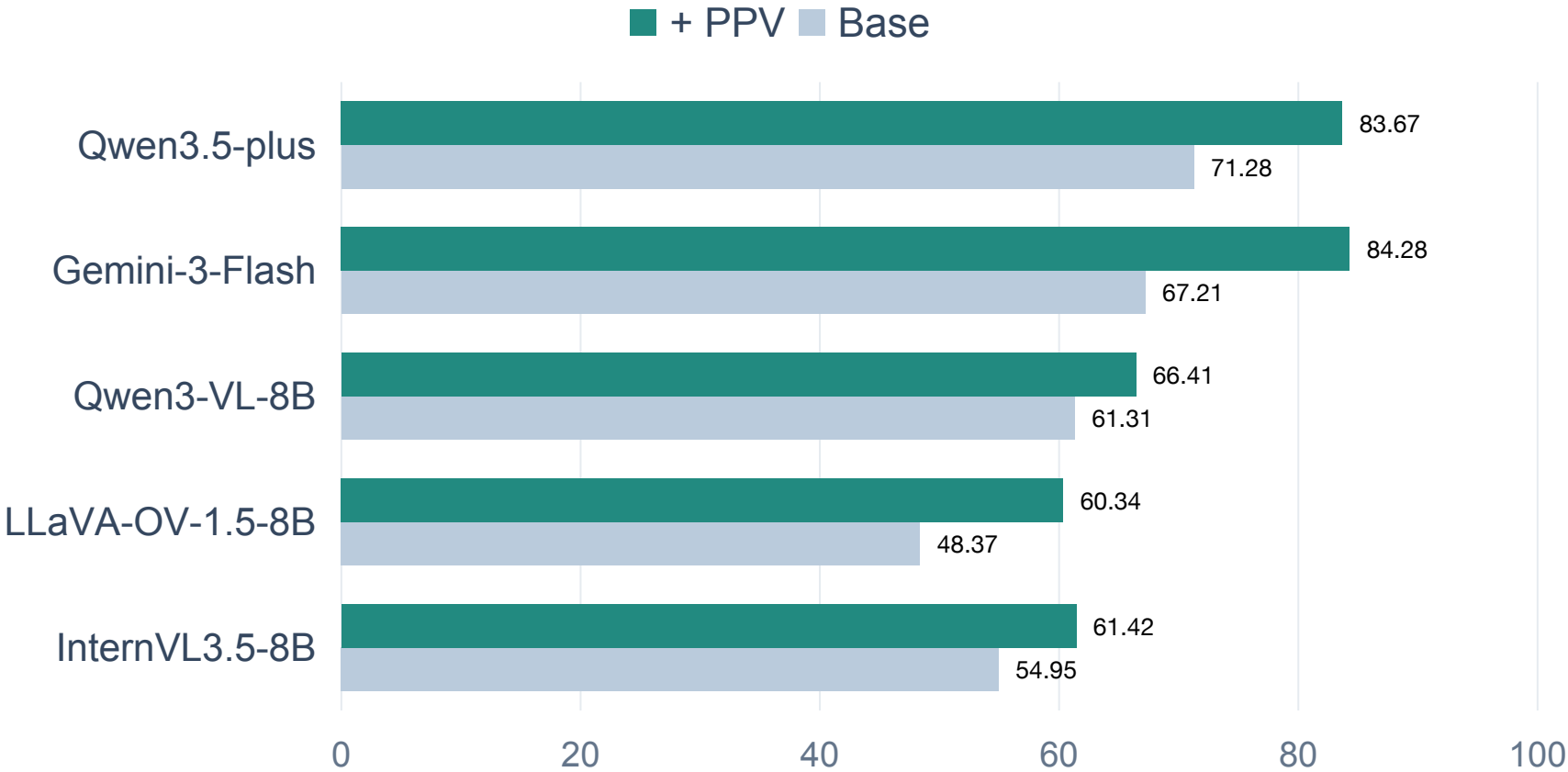
Perceive-Parse-Verify (PPV)

A training-free baseline that uses numerical kinematic evidence at inference time



PPV improves overall accuracy across five models

Overall accuracy (%)

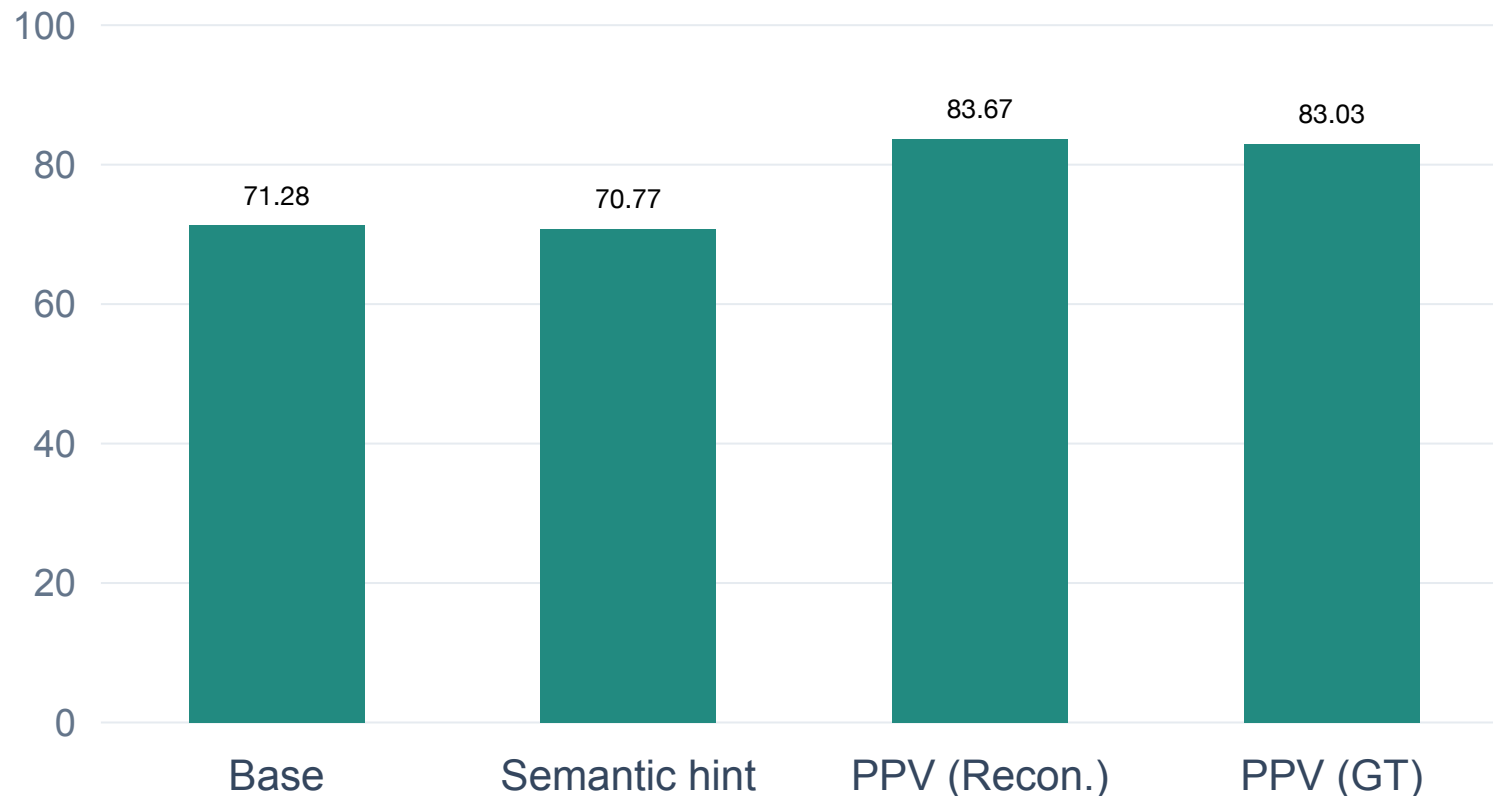


+10.6

percentage points
mean overall gain

Kinematic measurements drive the gain

Qwen3.5-plus



Semantic hints alone show no improvement.

Executing the queries adds physical measurements and raises accuracy by 12.39 pp over the base model.

Reconstructed and GT motion yield similar overall accuracy in this ablation.

Conclusions

Reliable motion feedback requires explicit kinematic comparison.

MotionHalluc diagnoses directional, attributional and temporal errors in paired-video reasoning.

PPV helps models understand motion more accurately, improving mean overall accuracy by 10.6 percentage points across five evaluated models.