Iris: Empowering Video MLLMs with High-Frequency Pose Priors via Spatiotemporal Binding

STRUCT @ PKU

KlingAI Research

Jiahang Zhang[1], Yushuo Guan[2], Yuanxing Zhang[2], Pengfei Wan[2], Jiaying Liu[1]

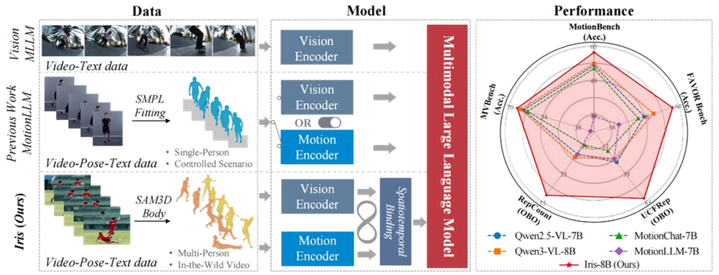
[1]Wangxuan Institute of Computer Technology, Peking University [2]KlingAI Research

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Human motion is a bottleneck for video MLLMs

- 2D patches carry no structural cues, so body dynamics stay implicit.
- Sparse sampling (16 frames or 1 fps) destroys high-frequency kinematics.
- Rendering motion cues onto RGB occludes appearance and inflates tokens.
- Pose is lightweight and structured: **10x denser sampling** at almost no token cost.

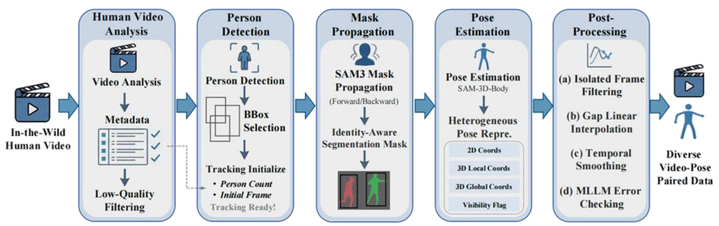


Contributions

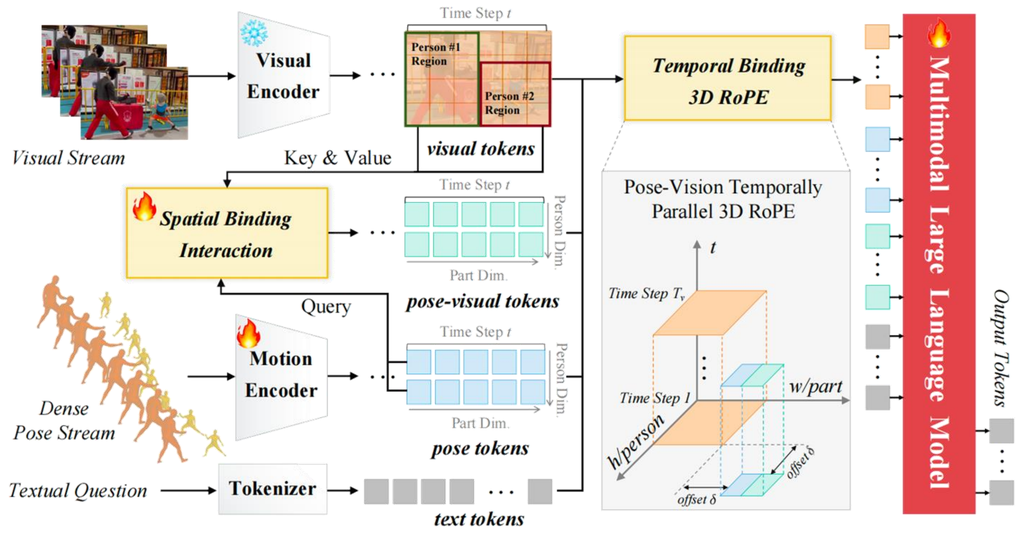
- **Asymmetric dual-stream MLLM.** Sparse vision plus a high-frequency pose stream for dense motion priors.
- **Spatiotemporal binding.** Parallel 3D RoPE for time, person-grounded cross-attention for space.
- **Automatic in-the-wild curation.** Markerless pipeline, 550K instruction samples.

Automatic Pose Data Pipeline

- Filtering, tracking, mask propagation, 3D pose estimation, smoothing, MLLM checking → **550K samples**.



Iris Framework



1. Part-Aware Pose Tokenization with Asymmetric Sampling

- 16 RGB frames pair with **160 pose frames** (20x denser); 24 keypoints form 5 body parts.
- A strided temporal conv and an 8-layer Transformer (74M) give part tokens; each keypoint is a 9-D descriptor (2D, pelvis-relative 3D, absolute pelvis, visibility).

2. Temporal Binding: Pose-Vision Parallel 3D RoPE

- Pose tokens share the temporal index of the visual tokens at the same step.
- Spatial axes become (person, part) with an offset, isolating pose in position space.

3. Spatial Binding: Person-Grounded Cross-Attention

- 2D keypoints give each person a box that masks out visual patches outside it.
- Pose tokens attend only to their own person; a zero-init head activates binding gradually.

4. Three-Stage Training on Curated Data

- Stage 1 pose-caption pretraining (30K); Stage 2 proxy-task alignment (400K); Stage 3 full SFT on 550K samples, 150K vision-only.

Experiments

Fine-grained motion understanding (Accuracy, %)

Model	MotionBench	FAVORBench
MotionLLM-7B	32.4	38.8
MotionChat-7B	54.8	45.7
Qwen2.5-VL-7B	56.2	47.8
Iris-7B (ours)	**60.5**	**54.1**
Qwen3-VL-8B	57.1	51.1
Iris-8B (ours)	**62.3**	**57.8**

Repetitive counting (MAE ↓ / OBO % ↑)

Model	RepCount	UCFRep
Qwen2.5-VL-7B	1.115 / 17.3	1.099 / 28.2
Iris-7B (ours)	**0.420 / 38.0**	**0.466 / 43.7**
Qwen3-VL-8B	0.664 / 20.0	0.696 / 25.2
Iris-8B (ours)	**0.346 / 50.0**	**0.344 / 62.1**

Ablation: how to inject pose (Accuracy, %)

Pose injection	MotionBench	FAVORBench
Pose as text	55.8	47.5
Pose heatmap	57.8	50.9
Motion encoder	58.4	52.0
+ Temporal binding	59.7	53.0
+ Spatial binding	**60.5**	**54.1**

Efficiency. Iris-8B spends 1,419 tokens / 24.67 TFLOPs, the cost of a 16-frame baseline, and beats 160-frame RGB (235.67 TFLOPs) by 2.6 pts.

Project Page

zjh2020@pku.edu.cn

Scan for paper, code and demos