# High-Speed Vision Improves Zero-Shot Semantic Understanding of Human Actions

Yongpeng Cao, Yuji Yamakawa • The University of Tokyo

Yamakawa Laboratory
The University of Tokyo

High-speed Flexible Robotics
http://www.hfr.iis.u-tokyo.ac.jp

**Question: Can high-speed video help pretrained models understand fast actions without task-specific training?**

Key result: complete raw Kendo actions are more separable at 120 Hz.

| 120 Hz | 60 Hz | 30 Hz |
|---|---|---|
| **83.3%** | **33.3%** | **41.7%** |
| Full raw | Full raw | Full raw |

## 1 Motivation

- Robots need to infer human intent from rapid, incomplete motion cues.
- Rare / specialized actions make labeled training data expensive.
- Most VLM/LLM action studies use standard frame-rate videos.

**Does higher frame rate improve semantic separation? How can we utilize it?**

## 2 Training-Free Semantic Comparison Framework

Video 120/60/30 Hz → 16-frame clips → InternVideo2.5 descriptions → Qwen3 pairwise comparison → Similarity matrix → NCP classification

**No task-specific training**
Pretrained models only; no Kendo-specific classifier.

**Semantic comparison**
The LLM compares description sequences from two videos.

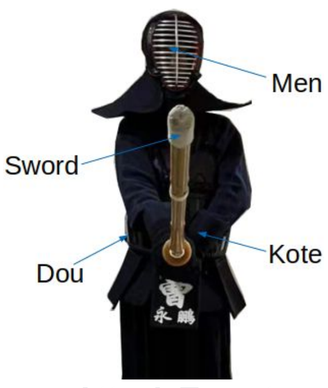
**NCP Classification**
Accuracy summarizes class separation from the similarity matrix.

## 3 Experimental Setup

**Kendo actions used:**
**(Different attack patterns for different targets)**

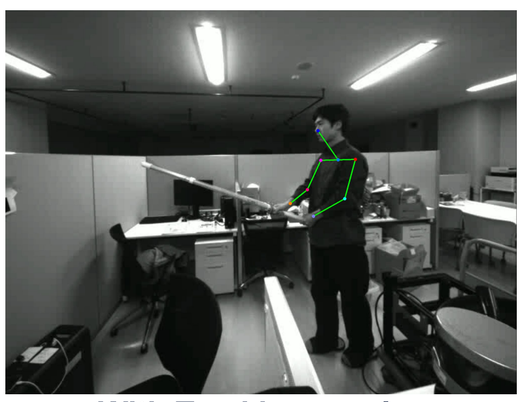
**[Men • Kote • Dou • Regular]**

- Captured at 120 Hz by high-speed camera, then downsampled to 60/30 Hz.
- 3 trials per class; full and partial observations.
- Raw video vs. tracked-joint overlay.




Attack Targets

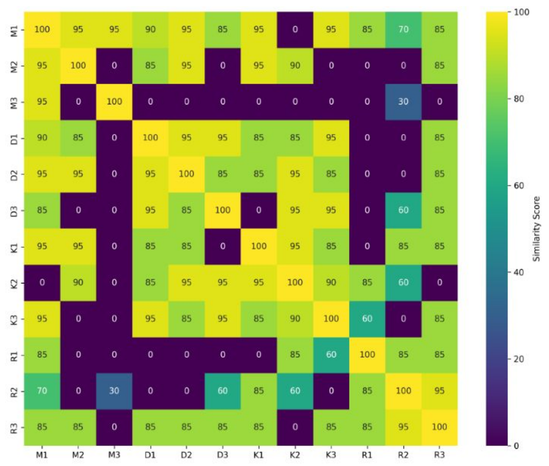
Raw Frame



With Tracking overlay

## 4 Results

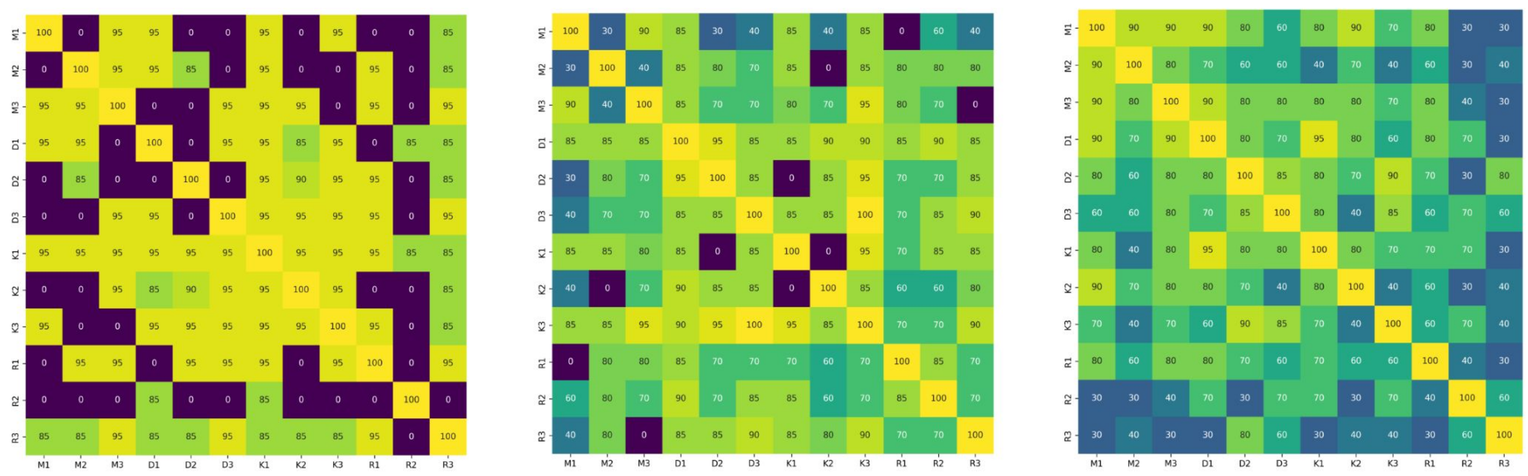
### Higher frame rate gives clearer semantic separation




(a) 120 Hz (b) 60 Hz (c) 30 Hz

Similarity matrices for complete raw Kendo actions at different frame rates.

**Interpretation**: 120 Hz preserves short transitional cues that lower frame rates miss.




(a) 120 Hz with Tracking (b) 60 Hz with Tracking (c) 30 Hz with Tracking

Similarity Matrices for Full Kendo Actions with tracking results overlapped.

### Tracking overlays are context-dependent

**Complete 120 Hz**
Raw 83.3% → +Tracking 16.7%
Overlay can distract.

**Partial 120 Hz**
Raw 33.3% → +Tracking 50.0%
Pose cue can help.

**NCP Accuracy (%)**

| Hz | Full Raw | Full +Track | Partial Raw | Partial +Track |
|---|---|---|---|---|
| 120 | **83.33** | 16.67 | 33.33 | **50.00** |
| 60 | 33.33 | 25.00 | 41.67 | 41.67 |
| 30 | 41.67 | 41.67 | – | – |

*NCP quantifies semantic separability from the similarity matrix.*

## 5 Discussion & Takeaways

### Why 120 Hz helps

- Fast Kendo attacks contain short, discriminative transitions.
- Training-free VLM–LLM reasoning is sensitive to missing motion cues.
- Higher temporal resolution makes generated descriptions more stable.

### Limitations

Small proof-of-concept dataset and predefined partial-observation criterion.

Pose-estimation noise and similar action semantics can reduce reliability.

### Takeaways

1. Temporal resolution matters for foundation-model-based action understanding.
2. Training-free semantic comparison is feasible for rapid, fine-grained actions.
3. Pose cues are useful only when the observation setting needs them.
4. High-speed perception can complement pretrained foundation models for real-time HRI.



Contact: cao@iis.u-tokyo.ac.jp