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

Yongpeng Cao[1,*] and Yuji Yamakawa[2]
{cao, y-ymkw}@iis.u-tokyo.ac.jp

***Abstract*— Understanding human actions is important for human–robot interaction, but collecting data for uncommon and fine-grained actions is difficult. Although pretrained video–language models enable zero-shot action understanding, the effect of temporal resolution on their semantic representations remains unclear. We investigate this issue using high-speed Kendo attacks and propose a training-free pipeline in which a pretrained video–language model describes short clips and a large language model evaluates pairwise action similarity. Experiments at 120, 60, and 30 Hz evaluate complete and partially observed actions, with and without tracked-joint overlays. Nearest-class prototype evaluation shows that 120 Hz video produces more separable semantic representations for complete actions. Joint overlays reduce performance on complete videos but can improve recognition under partial observation. These results demonstrate that temporal resolution affects the zero-shot interpretation of rapid actions and suggest that high-speed perception can improve training-free action understanding.**

I. Introduction

Understanding human actions from visual observations is fundamental to human–robot interaction (HRI), particularly when robots must respond to incomplete and rapidly evolving motion cues. Such interpretation is difficult for rare or specialized actions because collecting sufficient labeled data for supervised learning is often impractical.

Recent zero-shot and training-free approaches based on vision–language models (VLMs) and large language models (LLMs) offer an alternative to task-specific action-recognition models. However, existing work primarily considers actions at standard frame rates and leaves the effect of temporal resolution on rapid, fine-grained motions underexplored.

We study this question using high-speed Kendo attacks and a training-free pipeline that generates descriptions of short video clips with a pretrained video–language model and uses an LLM to compare the resulting description sequences. Controlled experiments at 120, 60, and 30 Hz evaluate complete and partially observed actions, both with and without tracked-joint overlays.

The main contributions of this work are summarized as follows:

- We study the effect of framerate on zero-shot semantic understanding of rapid human actions.
- We introduce a training-free VLM–LLM pipeline for pairwise action comparison.
- We evaluate complete and partial observations and analyze when tracked-joint overlays are beneficial.

*Corresponding author
[1]Institute of Industrial Science, The University of Tokyo, Japan
[2]Interfaculty Initiative in Information Studies, Graduate School of Interdisciplinary Information Studies, The University of Tokyo, Japan

II. Related Work

A. Human Action Recognition

Human action recognition commonly uses skeleton-based graph networks or video models trained on annotated datasets [1], [2], [3], [4], [5], [6], [7], [8], [9]. Although effective on standard benchmarks, these methods are difficult to apply to rare or specialized actions for which labeled data are limited.

B. Vision–Language Models for Action Understanding

Pretrained vision–language and language models have enabled semantic action representations without conventional class-specific classifiers. Existing approaches align motion with text [10], [11], convert motion into linguistic descriptions for LLM reasoning [12], [13], or perform VLM-based recognition [14], [15]. However, these studies largely consider everyday actions recorded at standard frame rates, leaving the effect of temporal resolution on rapid, fine-grained actions underexplored.

C. Early Action Recognition under Partial Observation

Early action recognition predicts an action from incomplete observations and is important for low-latency robotic response [16], [17]. Existing methods generally require supervised training, whereas we study whether high-speed partial observations remain semantically informative in a zero-shot setting.

D. Kendo Motion Analysis and High-Speed Perception

Previous kendo systems have used trajectory estimation, wearable sensing, and markerless tracking for recognition or robotic response [18], [19], [20]. In contrast, we investigate training-free semantic comparison directly from high-speed video.

III. Methodology

The proposed training-free pipeline, illustrated in Fig. 1, converts video clips into textual descriptions, compares description sequences using an LLM, and derives zero-shot classifications from the result similarity matrix.

Each sample video is first processed using a fixed-length clip segmentation strategy. Specifically, a video V_i is divided

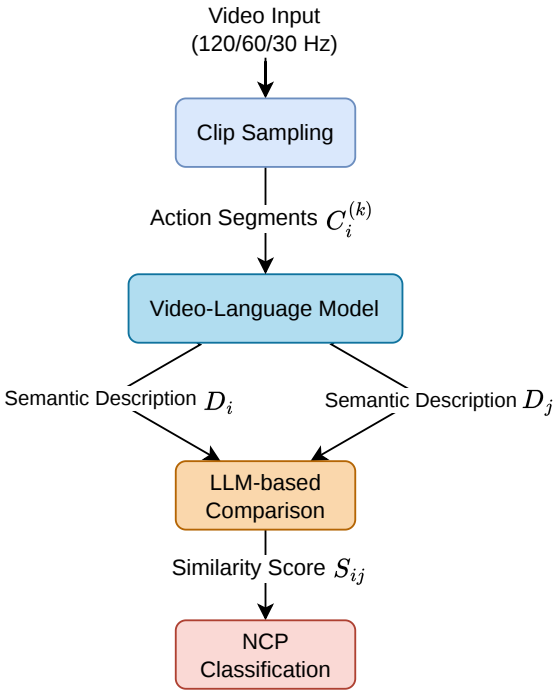


Fig. 1: System Framework.

into non-overlapping clips with a temporal window of L frames. The k-th clip is defined as:

$$C_i^{(k)} = \{f_{i,t} \mid t = (k-1)L + 1, \ldots, kL\}, \tag{1}$$

where $f_{i,t}$ denotes the t-th frame of video V_i. This results in a set of clips $\{C_i^{(k)}\}_{k=1}^{K_i}$ that capture local temporal dynamics of the action. K_i here denotes the total number of clips for video V_i, which depends on the video length and the chosen clip length L. Each clip is processed by a pre-trained video-language model to generate a textual description:

$$d_i^{(k)} = \mathrm{VLM}(C_i^{(k)}), \tag{2}$$

where $d_i^{(k)}$ denotes the description corresponding to the k-th clip. Each video V_i is therefore represented as a sequence of textual descriptions:

$$D_i = \{d_i^{(k)}\}_{k=1}^{K_i}. \tag{3}$$

To measure semantic similarity between actions, we adopt a pairwise comparison strategy using an LLM. For each pair of videos (i, j), the LLM takes the corresponding sequences of descriptions (D_i, D_j) as input and assigns a similarity score to each action pair following an LLM-as-a-judge paradigm, reflecting their semantic proximity:

$$S_{ij} = \mathrm{LLM}(D_i, D_j), \tag{4}$$

where $\mathrm{LLM}(\cdot, \cdot)$ denotes the similarity score generated based on the comparison of the two description sequences. The resulting similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ captures pairwise semantic relationships across all samples.

Finally, to obtain classification results without training, we adopt the Nearest-Class Prototype (NCP) strategy [21]. It assigns a query sample to the class whose prototype is closest. This approach does not require training a discriminative classifier and is therefore suitable for training-free and zero-shot recognition settings. It computes the average action similarity between each sample and all samples belonging to a class, and selects the class with the highest average similarity score as the prediction result. In our case, if two or more classes have the same highest score for a given sample, the prediction is regarded as a failed classification. For each sample i and class c, we compute the average similarity between i and all samples belonging to class c:

$$\mu_i^{(c)} = \frac{1}{|\mathcal{I}_c|} \sum_{j \in \mathcal{I}_c} S_{ij}, \tag{5}$$

where $\mathcal{I}_c$ denotes the index set of samples in class c. The predicted label for sample i is then given by:

$$\hat{y}_i = \arg\max_c \ \mu_i^{(c)}. \tag{6}$$

Additionally, to evaluate the proposed framework under real-time human–robot interaction conditions, we add early action recognition setting using partial observations. Instead of using the full action sequence, only the initial stage of each action is retained according to the segmentation criterion introduced in our previous work [20]. This setting simulates practical scenarios where robotic systems must infer human intent before the action is fully completed in order to enable timely responses.

IV. Experiments

A. High-Speed Video Collection and Preprocessing

To investigate the effect of temporal resolution on zero-shot semantic understanding, we collected high-speed video data of kendo attack actions. Videos were captured at 120 Hz using a monochrome high-speed camera, with action durations ranging from 2 to 3 seconds. We use the Ximea MQ013MG-ON monochrome camera, which provides high frame rates for capturing fine-grained motion details in sufficient resolution.

Each video was down-sampled to 60 Hz and 30 Hz to create multiple temporal resolution conditions for comparison.

In addition, we applied a pose estimation model to extract upper body joint positions, which were overlaid on the original video frames to examine the impact of human joint information on semantic understanding. We adopt the joint tracking approach proposed in our previous work [20] for smooth joint trajectory. The tracking results were visualized as colored skeletons, with different colors representing different joints. Fig. 2 shows an example of the tracking results overlaid on the original video frames.

B. Data Preparation

We prepare three trials for each attack pattern: Men, Kote, and Dou, together with a blank category without an attack action to serve as a baseline for comparison. For clarity, we denote the action labels as M (Men), K (Kote), D (Dou), and R (Regular), with trial indices from 1 to 3. Original Kendo action videos were captured at 120 Hz.

(a) Without Joints Tracking

(b) With Joints Tracking

Fig. 2: Example of High-Speed Kendo Action Video Frame with and without Joints Tracking Results Overlaid.

First, video-level semantic representations are generated from action frames. InternVideo2.5 [22] is adopted to generate semantic summaries from short video clips. In the experiments, we group 16 consecutive frames into one chunk to produce a semantic description.

In these experiments, *Qwen3-4B-Instruct-2507* is selected as the pre-trained LLM for both reasoning and comparison. The model is prompted to focus on temporal ordering, attack patterns, target regions, and key motion transitions, while no task-specific fine-tuning is applied.

C. Results

Fig. 3a visualizes the resulting similarity matrix at 120 Hz. A diagonal-dominant structure can be observed, indicating higher similarity scores for action pairs within the same category and lower scores across different categories.

We then evaluate the effect of temporal resolution on zero-shot semantic comparison. The original 120 Hz videos are down-sampled to 60 Hz and 30 Hz before being processed by the same pipeline. The resulting similarity matrices for the full actions without tracking are shown in Fig. 3. Compared to the 120 Hz results, lower frame-rate inputs exhibit less structured similarity patterns. At 60 Hz (Fig. 3b), although some intra-class similarity is preserved, high similarity scores are frequently assigned to semantically unrelated action pairs. At 30 Hz (Fig. 3c), the similarity scores become more uniform across most of action pairs, resulting in reduced interpretability and weaker semantic separation. All similarity matrix visualizations share the same color bar scale; therefore, the color bar is only shown in Fig. 3a to avoid redundancy.

To explore how joints tracking overlay could improve the classification, we overlap the tracking results of the upper body joints set on the video frames to examine how this additional human joint information would affect action understanding. Results are visualized in Fig. 4. Compared to results without tracking, these similarity maps are less regular and exhibit a larger number of failure cases, but still present the same trend of the high similarity score on high frequency. Interestingly, the classification accuracy is higher for low-frequency samples than for high-frequency samples, which indicates that the simple overlapping strategy for tracked joints can be conflicting for zero-shot Kendo video understanding, especially in high-speed motion scenarios.

Both partially observed action videos with and without tracking-result overlays are evaluated under the same conditions. Similarity matrices are visualized in Fig. 5. To clarify, the 30 Hz data do not satisfy the input requirements of the captioning model and are therefore excluded from evaluation. Compared with complete videos, early-stopped videos exhibit degraded video understanding performance. However, applying tracking overlays helps suppress outlier similarities, compared to the corresponding samples without tracking information.

The primary objective of this experiment is semantic comparison rather than strict action classification, which better reflects human perception of action similarity. Nevertheless, to provide a quantitative evaluation of semantic separability, we additionally adopt the NCP method to estimate classification accuracy from the similarity matrix. The results summarized in Table I demonstrate that higher temporal resolution contributes to more stable zero-shot semantic discriminability for fast kendo actions. It highlights the complementary role of high-speed vision in training-free semantic analysis. This experiment further supports that high-speed perception can enhance interpretation ability of rapid actions.

TABLE I: NCP Accuracy (%) on Full Videos

	Full		Partial	
Hz	Raw	Track.	Raw	Track.
120	83.33	16.67	33.33	50.00
60	33.33	25.00	41.67	41.67
30	41.67	41.67	–	–

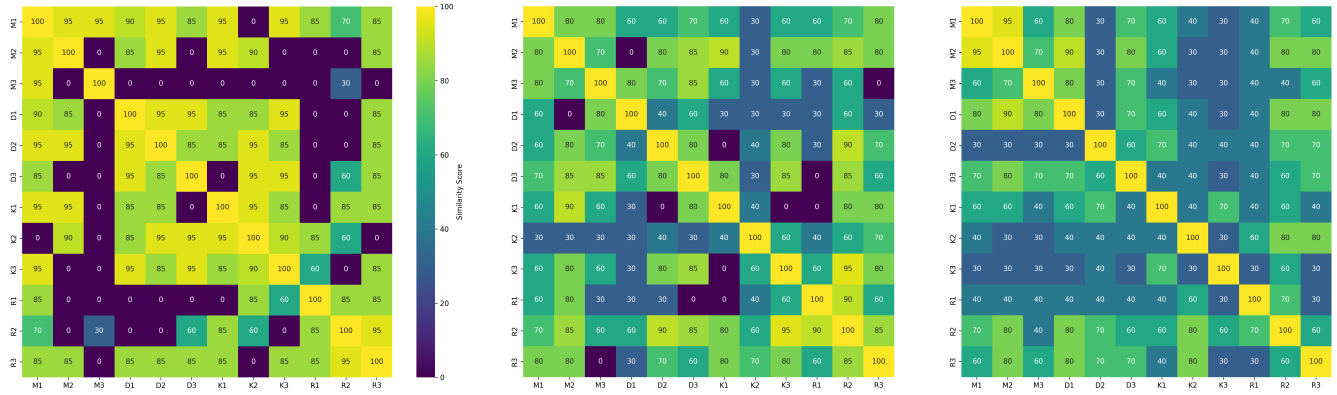
D. Discussion

From the experimental results, we observe that 120 Hz videos provide more reliable semantic separation between different kendo actions.

This can be attributed to several factors: First, higher frame rates preserve critical transitional moments that are essential for distinguishing fast and similar actions. Second, unlike training-based models that learn to detect key discriminative features, the proposed training-free pipeline relies on LLM-based semantic captioning and reasoning, which is more sensitive to missing or sparsely sampled motion cues. Finally, higher temporal resolution introduces greater fault tolerance, making the overall semantic comparison more robust for swift and fine-grained movements.

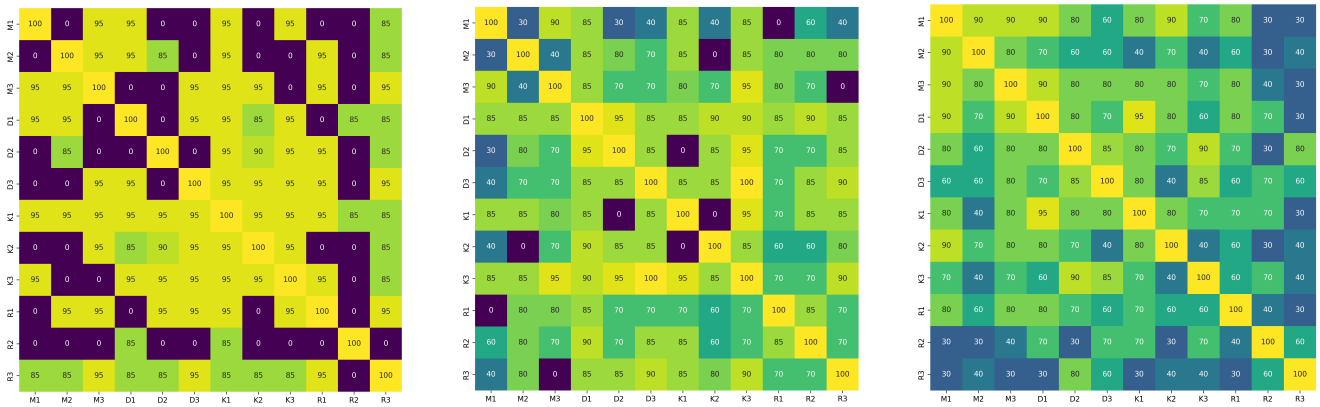
Moreover, tracking overlays are not clearly beneficial for complete action recognition, but become useful when the action is temporally incomplete. This observation indicates that tracking information is most effective when applied to key action periods for understanding the action, while it may be redundant or even distracting when overlaid across the entire action sequence.

This study is limited by its small proof-of-concept dataset and predefined early-stopping criterion. Pose errors and occlusion can also degrade the tracking condition, while



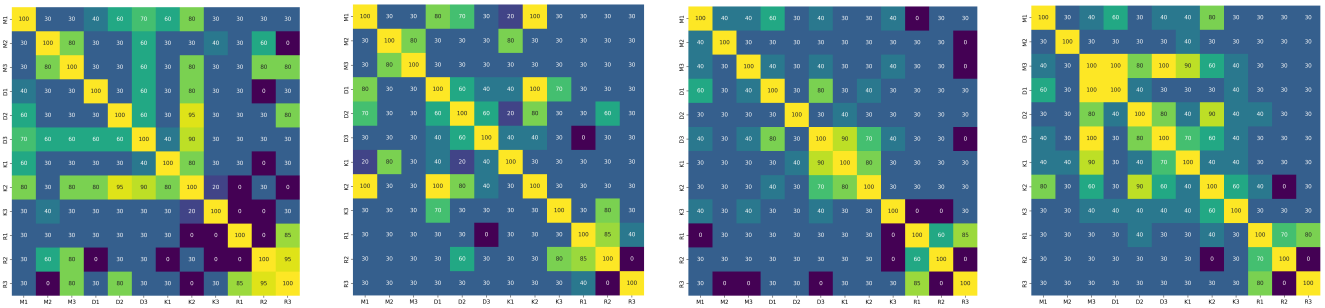
(a) 120 Hz (b) 60 Hz (c) 30 Hz

Fig. 3: Similarity Matrices for Full Kendo Actions at Different Frame Rates.



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

Fig. 4: Similarity Matrices for Full Kendo Actions with tracking results overlapped.



(a) 120 Hz Segmented (b) 60 Hz Segmented (c) 120 Hz Segmented+Tracking (d) 60 Hz Segmented+Tracking

Fig. 5: Similarity Matrices for Partially Observed Actions.

generally high cross-action scores indicate that LLM focus on broad semantic resemblance over fine-grained execution differences.

V. Conclusion

To conclude, this work investigates the impact of temporal resolution on zero-shot semantic understanding of high-speed human actions, using kendo as a representative case. We propose a training-free pipeline that combines video-language models and LLM to compare action pairs without task-specific training. Our experiments demonstrate that higher frame rates significantly improve semantic discriminability, while tracking information is most beneficial for early action recognition. These findings highlight the importance of temporal resolution in training-free action understanding and suggest that high-speed perception can enhance semantic reasoning capabilities for real-time robotic systems. Future work will investigate more robust multimodal integration strategies for zero-shot action understanding, as well as the adoption of hierarchical, coarse-to-fine reasoning strategies to better capture key features from different modalities.

REFERENCES

[1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[2] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 359–13 368.

[3] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 183–192.

[4] W. Myung, N. Su, J.-H. Xue, and G. Wang, "Degcn: Deformable graph convolutional networks for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 33, pp. 2477–2490, 2024.

[5] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, and X.-S. Hua, "Blockgcn: Redefine topology awareness for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2049–2058.

[6] H. Liu, Y. Liu, M. Ren, H. Wang, Y. Wang, and Z. Sun, "Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition," *arXiv preprint arXiv:2411.18941*, 2024.

[7] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "Infogcn: Representation learning for human skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 186–20 196.

[8] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[9] J. Liu, X. Wang, C. Wang, Y. Gao, and M. Liu, "Temporal decoupling graph convolutional network for skeleton-based gesture recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 811–823, 2023.

[10] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 358–374.

[11] H. Zhang, M. C. Leong, L. Li, and W. Lin, "Pevl: Pose-enhanced vision-language model for fine-grained human action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 857–18 867.

[12] H. Qu, Y. Cai, and J. Liu, "Llms are good action recognizers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 395–18 406.

[13] W. Xiang, C. Li, Y. Zhou, B. Wang, and L. Zhang, "Generative action description prompts for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 276–10 285.

[14] C. Han, H. Wang, J. Kuang, L. Zhang, and J. Gui, "Training-free zero-shot temporal action detection with vision-language models," *arXiv preprint arXiv:2501.13795*, 2025.

[15] M. Bosetti, S. Zhang, B. Liberatori, G. Zara, E. Ricci, and P. Rota, "Text-enhanced zero-shot action recognition: A training-free approach," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 327–342.

[16] A. Stergiou and D. Damen, "The wisdom of crowds: Temporal progressive attention for early action prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 709–14 719.

[17] X. Liu, J. Yin, D. Guo, and H. Liu, "Rich action-semantic consistent knowledge for early action prediction," *IEEE Transactions on Image Processing*, vol. 33, pp. 479–492, 2023.

[18] A. Namiki and F. Takahashi, "Motion generation for a sword-fighting robot based on quick detection of opposite player's initial motions," *Journal of Robotics and Mechatronics*, vol. 27, no. 5, pp. 543–551, 2015.

[19] M. Takata, Y. Nakamura, Y. Torigoe, M. Fujimoto, Y. Arakawa, and K. Yasumoto, "Strikes-thrusts activity recognition using wrist sensor towards pervasive kendo support system," in *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2019, pp. 243–248.

[20] Y. Cao and Y. Yamakawa, "Marker-less kendo motion prediction using high-speed dual-camera system and lstm method," in *2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, 2022, pp. 159–164.

[21] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Distance-based image classification: Generalizing to new classes at near-zero cost," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2624–2637, 2013.

[22] Y. Wang, X. Li, Z. Yan, Y. He, J. Yu, X. Zeng, C. Wang, C. Ma, H. Huang, J. Gao, *et al.*, "Internvideo2. 5: Empowering video mllms with long and rich context modeling," *arXiv preprint arXiv:2501.12386*, 2025.