Yangliu HU, Zikai SONG, Junqing YU, et al., 2025. TimeJudge: empowering video-LLMs as zero-shot judges for temporal consistency in video captions. *Front Inform Technol Electron Eng*, 26(11):2204-2214. https://doi.org/10.1631/FITEE.2500412

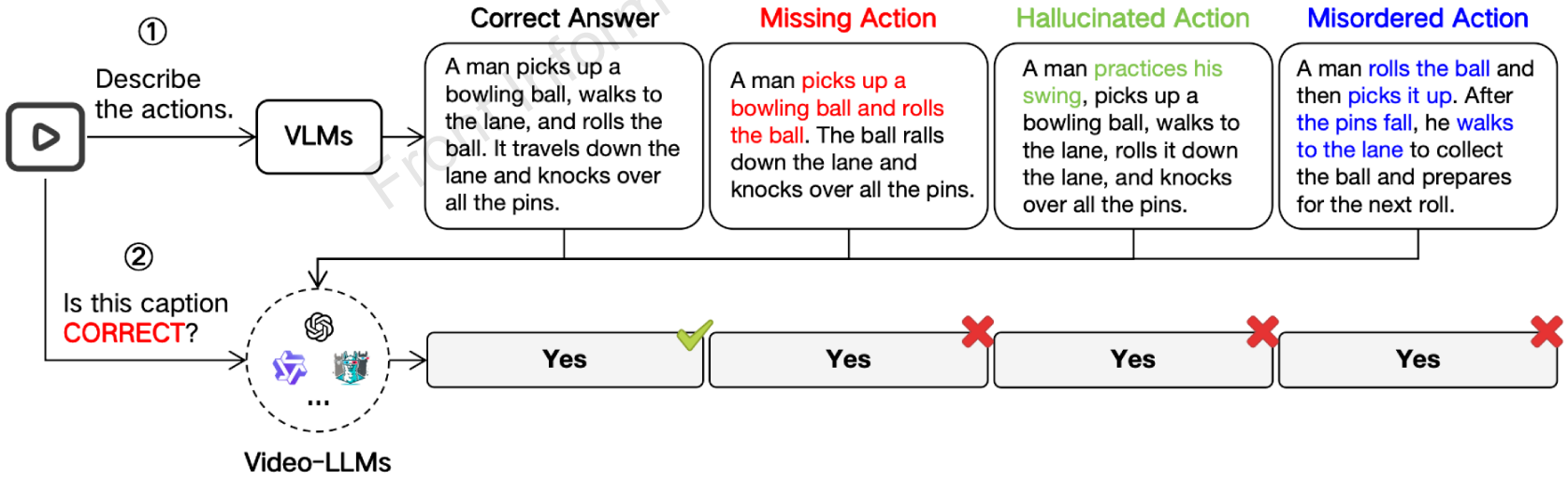
TimeJudge: empowering video-LLMs as zero-shot judges for temporal consistency in video captions

Key words: Video large language model (Video-LLM); Multimodal large language model (MLLM); MLLM-as-a-Judge; Video caption; Benchmark

Corresponding author: Zikai SONG
E-mail: skyesong@hust.edu.cn
ORCID: https://orcid.org/0009-0006-6651-2027

Motivation

Targeting the limitation of current video caption evaluation methods, which are insensitive to fine-grained temporal errors and may be easily misled by fluent but incorrect descriptions, a novel zero-shot temporal consistency judgment framework, termed TimeJudge, is proposed to reliably detect temporal inconsistencies in video captions without requiring any additional training or human annotations.

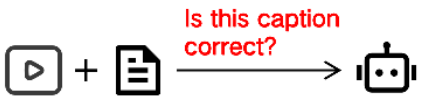


Main idea

- A novel temporal-consistency-oriented evaluation framework is proposed which reformulates video caption evaluation as a set of fine-grained binary question answering tasks, enabling explicit reasoning over temporal presence, order, and completeness of events.
- A cross-modal answer consistency mechanism is introduced which independently answers the same temporal questions based on the video and the caption, and determines temporal correctness by measuring the consistency between visual-based and text-based responses.
- A modality-aware calibration and consistency-aware voting strategy is proposed which mitigates language priors and visual bias, and enables robust zero-shot detection of missing, hallucinated, and misordered events without any additional training.

Method

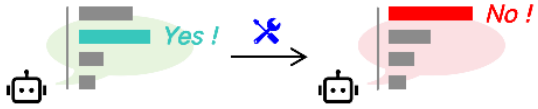
① Quesiton Decomposition

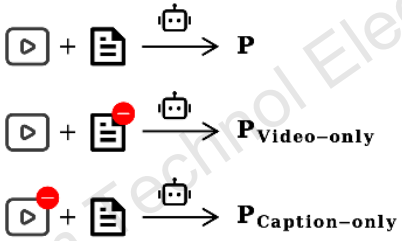


Generate binary questions to assess video-caption alignment in plausible captions, emphasizing temporal over spatial cues. For each question, create an inverse to support cross-validation.

Q_1: Does the caption describe every action that leads to a visible change in the video?
Q_1 inverse: Are there any visible state changes in the video not caused by actions mentioned in the caption?
Q_2: Would the caption lead the reader to imagine something that doesn't happen in the video?
Q_1 inverse: Does the caption suggest events that may mislead the reader about what really happens?
Q_3: ...
Q_3 inverse: ...
...

② Modality-Sensitive Calibration

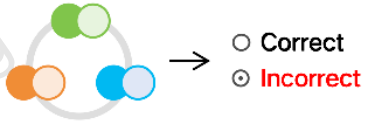


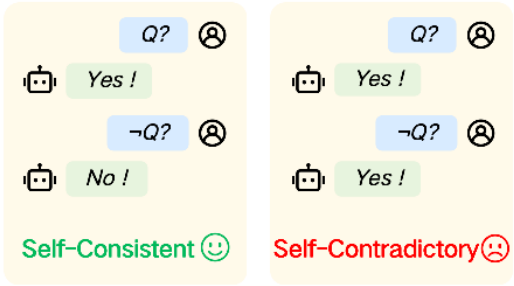


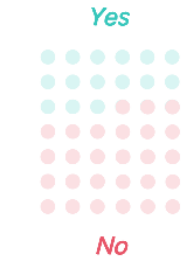
$$\Delta c = P - P_{Video-only}$$
$$\Delta v = P - P_{Caption-only}$$
$$P' = P + \lambda f(\Delta c, \Delta v)$$

$\Delta c, v > 0$ → *Helpful !*
$\Delta c, v = 0$ → *Ignored !*
$\Delta c, v < 0$ → *Misguided!*

③ Judgment Aggregation

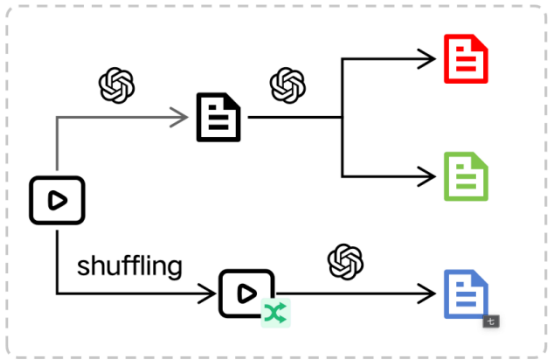


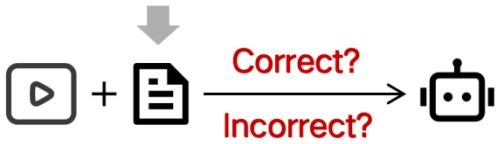


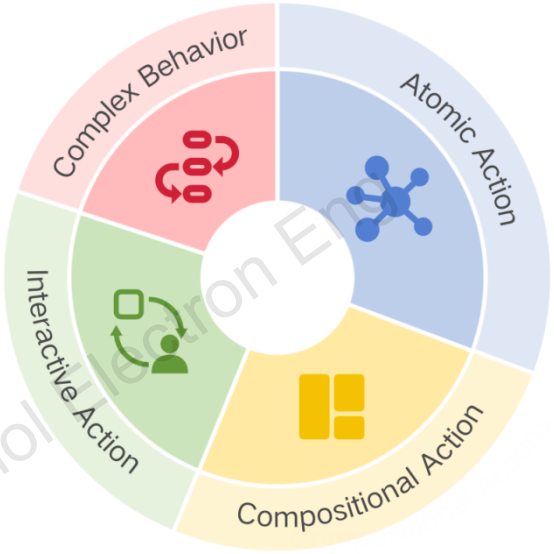


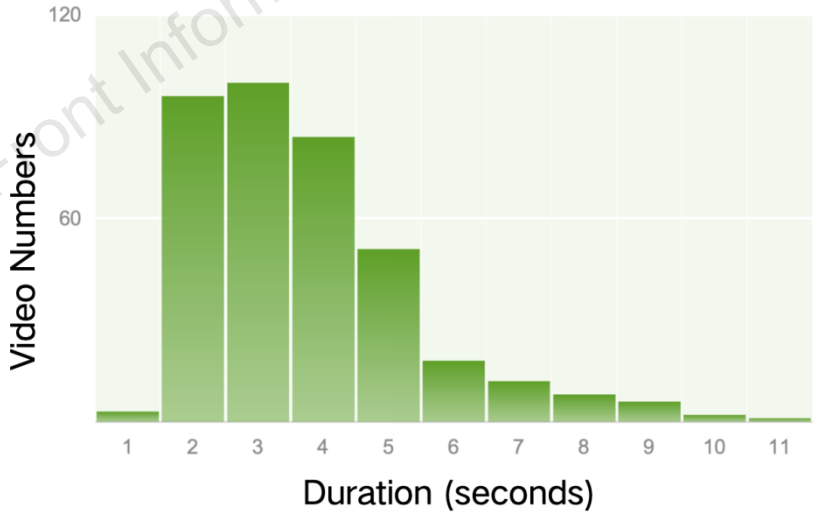
Answer: Incorrect
Type: Missing actions
Reason: The action "walks to the lane" is clearly shown in the video but is absent from the caption.

Benchmark









Results

Table 2 Performance on the TEDBench

Video-LLMs	N_{par}	Method	Missing action			Hallucinated action			Misordered action		
			Accuracy (%)	Recall (%)	F1	Accuracy (%)	Recall (%)	F1	Accuracy (%)	Recall (%)	F1
Random			25.00	25.00	0.2500	25.00	25.00	0.2500	25.00	25.00	0.2500
Qwen2.5-VL	7B	Base	50.26	50.16	0.6295	54.59	52.59	0.6723	50.13	50.08	0.6282
		Ours	**61.29**	**57.03**	**0.7029**	**68.77**	**62.84**	**0.7463**	**72.05**	**72.34**	**0.7186**
VideoLLaMA 3	7B	Base	53.41	53.63	0.5196	59.38	61.76	0.5478	58.40	58.08	0.5920
		Ours	**66.67**	**83.96**	**0.5528**	**66.67**	**65.38**	**0.6801**	**69.16**	**73.55**	**0.6599**
InternVL3	8B	Base	47.11	47.44	0.5031	55.51	54.25	0.6126	56.25	55.71	0.5821
		Ours	**66.01**	**72.93**	**0.5997**	**75.20**	**84.04**	**0.7149**	**71.13**	**85.46**	**0.6382**
MiniCPM-o 2.6	8B	Base	54.16	62.00	0.4059	62.73	69.48	0.5492	56.56	62.38	0.4322
		Ours	**68.50**	**62.39**	**0.7474**	**72.83**	**75.44**	**0.7137**	**61.30**	**75.57**	**0.4776**
GPT-4o mini		Base	61.29	57.62	0.6878	65.88	61.02	0.7204	63.12	60.25	0.6766
		Ours	**68.50**	**62.35**	**0.7479**	**74.67**	**67.81**	**0.7859**	**75.72**	**72.17**	**0.7752**

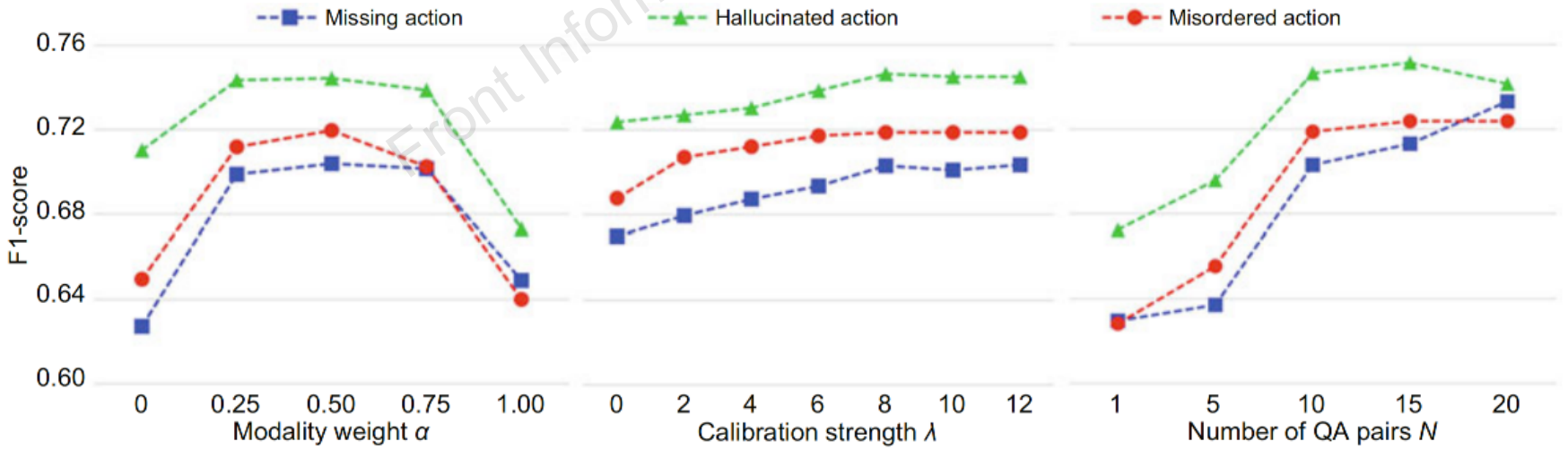
N_{par}: number of parameters. Better results are in bold. Base: results with the original model predictions. Ours: results after applying TimeJudge. All models consistently improve across three temporal error types after applying TimeJudge, demonstrating the latter's general effectiveness. Notably, Qwen2.5-VL and GPT-4o mini achieved comprehensive gains across all metrics, while other models showed significant gains in terms of recall

Results

Table 3 Impact of each component of the TimeJudge

Method	Missing action			Hallucinated action			Misordered action		
	Accuracy (%)	Recall (%)	F1	Accuracy (%)	Recall (%)	F1	Accuracy (%)	Recall (%)	F1
Random	25.00	25.00	0.2500	25.00	25.00	0.2500	25.00	25.00	0.2500
Base	50.26	50.16	0.6295	54.59	52.59	0.6723	50.13	50.08	0.6282
QD	51.32	50.70	0.6599	62.34	57.56	0.7139	66.14	64.82	0.6759
QD+MSC	55.25	52.84	0.6857	67.06	61.32	0.7372	71.26	71.89	0.7084
QD+MSC+JA (TimeJudge)	61.29	57.03	0.7029	68.77	62.84	0.7463	72.05	72.34	0.7186

Progressive addition of QD, MSC, and JA steadily improves all metrics across temporal error types, with the full framework achieving the best overall performance



Conclusions

In this work, we propose TimeJudge, a novel zero-shot temporal consistency judgment framework for video caption evaluation, aiming to reliably detect fine-grained temporal errors without requiring additional training or human annotations. The design of TimeJudge demonstrates the effectiveness of question decomposition and cross-modal consistency reasoning, and provides new insights for building reliable, scalable, and annotation-free evaluation mechanisms for video large language models.

Yangliu Hu is a PhD candidate at the School of Computer Science and Technology, Huazhong University of Science and Technology (HUST), with research interests in multimodal large language models, video understanding, object detection, and virtual reality.

Zikai Song received his PhD degree in computer application technology from HUST in 2023. His research interests include computer vision, multimedia models, video analysis, and social media analysis.

Wei Yang received his PhD degree in computer science from University of Delaware in 2017. He worked at Google ATAP and the startup DGene on vision-based scene understanding and virtual reality before joining HUST in 2021, where he is currently an Associate Researcher and a recipient of the Hubei Provincial “Hundred Talents Program” and the Donghu Scholar Award. His research interests include 3D vision understanding and reconstruction, physics-based vision and graphics, advanced sensing, and novel imaging sensors. He has published over 20 papers in top venues such as TPAMI, TOG, IJCV, CVPR, ICCV, and ECCV. He serves as Area Chair for ICML 2025, ICCV 2025, CVPR 2023/2024/2025, and NeurIPS 2023/2024, and has been a PC member of AAAI 2020/2021, WACV 2021, and BMVC 2018.

Yiping Phoebe Chen is Professor and Chair of the Department of Computer Science and Information Technology at La Trobe University, Australia. She received her First Class Honours BInfTech and PhD degree from the University of Queensland. She has obtained around 30 research grants totaling about 8 million AUD, including 13 ARC grants, and serves as Chief Investigator of the ARC Centre of Excellence in Bioinformatics. Before joining La Trobe in 2010, she held academic positions at Deakin University and Queensland University of Technology. Her multidisciplinary research spans bioinformatics, multimedia, AI, data mining, scientific visualization, and medical image analysis, with recent work focusing on RNA structure knowledge discovery, deep learning applications, disease diagnosis, drug design, and biomolecular network mining. She has published over 300 papers in top venues such as TNNLS, Nature Machine Intelligence, AI, Bioinformatics, T-BME, PR, NAR, IS, CVPR, AAAI, SIGMOD, and ACM MM. She has served as Associate Editor of TNNLS and T-MM, Editor-in-Chief of Current Bioinformatics, Vice Chair of ACM SIGMM, and chair or PC member of more than 100 international conferences.

Junqing Yu received his PhD degree in computer science from Wuhan University, Wuhan, China, in 2002. From 2002 to 2003, he was with University of Manitoba, Winnipeg, Canada, conducting research in digital media processing and networked multimedia systems. Since 2003, he has been with HUST, Wuhan, China, serving as a Deputy Director (Acting) of the Network and Computing Center, Associate Director of the Institute of Digital Media and Intelligent Technology, and Leader of the Digital Media Processing and Retrieval Research Team. Since 2023, he has also served on secondment as a Member of the Leading Party Members' Group and Executive Secretary of the Secretariat of the China Association for Science and Technology, and concurrently as a Standing Committee Member of the Party Committee and Vice President of HUST. His research interests include computer networks, digital media processing and retrieval, multicore computing, and stream compilation. He has led over 20 national and provincial research projects, published more than 200 papers, and held over 60 patents and software copyrights. He has authored one textbook, three monographs, and one translated volume.