Caption Anything in Video: Fine-grained Object-centric Captioning via Spatiotemporal Multimodal Prompting

Yolo Yunlong Tang[1], Jing Bi[1], Chao Huang[1], Susan Liang[1], Daiki Shimada[2], Hang Hua[1], Yunzhong Xiao[3], Yizhi Song[4], Pinxin Liu[1], Mingqian Feng[1], Junjia Guo[1], Zhuo Liu[1], Luchuan Song[1], Ali Vosoughi[1], Jinxi He[1], Liu He[4], Zeliang Zhang[1], Jiebo Luo[1], Chenliang Xu[1]

[1] University of Rochester, [2] Sony Group Corporation, [3] CMU, [4] Purdue University

AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20–22, 2026. 1. SINGAPORS

Why Object-Centric Captions?

We reveal the gap: VidLLMs output generic video-level text, missing **object-level precision** and temporal dynamics. Dense captioning is overly concise. CAT-V restores **user control** with visual prompts, delivering fine-grained, temporally coherent descriptions across evolving **object states** and **interactions**.

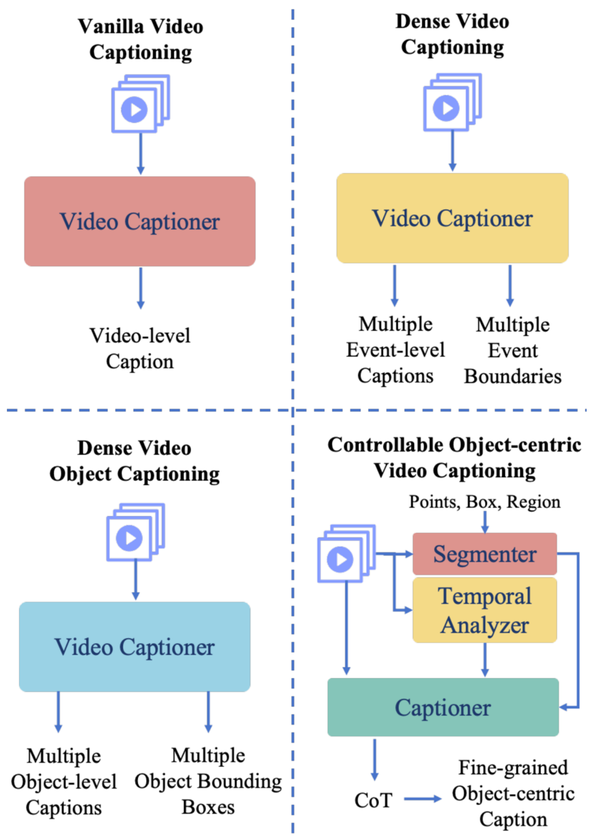


Figure 1. Comparison of video captioning approaches.

Our Contribution

- We introduce **CAT-V**, a training-free stack uniting **SAMURAI**, **TRACE-Uni**, and **InternVL-2.5**.
- Supports points, boxes, regions; yields **spatially precise**, temporally aware object captions.
- Needs no extra data; preserves **instruction-following** and rich **user interaction**.

CAT-V Framework

Given video V with T frames and prompt p, the **Segmenter** outputs masks $\{M_t\}$. We inject **masklets** as **spatiotemporal prompts**; the **Temporal Analyzer** returns $\{(s_i,e_i), c_i\}$. The **Captioner** fuses all—with CoT—to produce coherent, **object-centric** narratives.

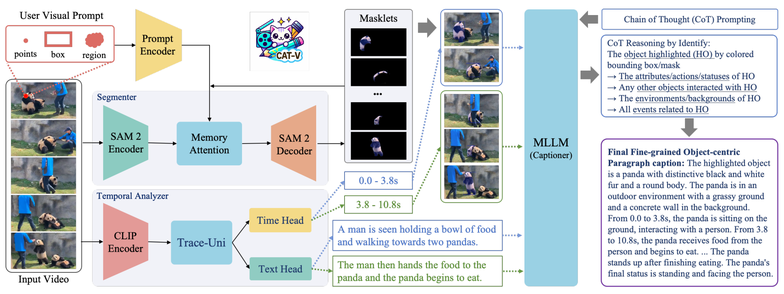


Figure 2. CAT-V Framework.

Precision Segmenter

Built on **SAMURAI** [3] with Kalman filtering and motion-aware memory, the Segmenter delivers pixel-accurate **masklets** under occlusion, blur, and clutter. It encodes frames and prompts, handles points, boxes, irregular regions, and locks onto the **highlighted object** across time.

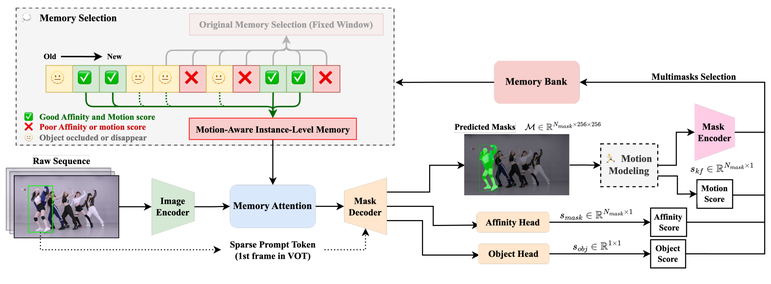


Figure 3. Our Segmenter is built on SAMURAI model.

Temporal Analyzer

Built on **TRACE-Uni** [2], the Temporal Analyzer hierarchically parses V, discovering N events with boundaries $\{(s_i,e_i)\}$ and coarse captions $\{c_i\}$. It powers **status-change tracking**, aligns actions to timestamps, and strengthens **temporal grounding** for object-centric reasoning.

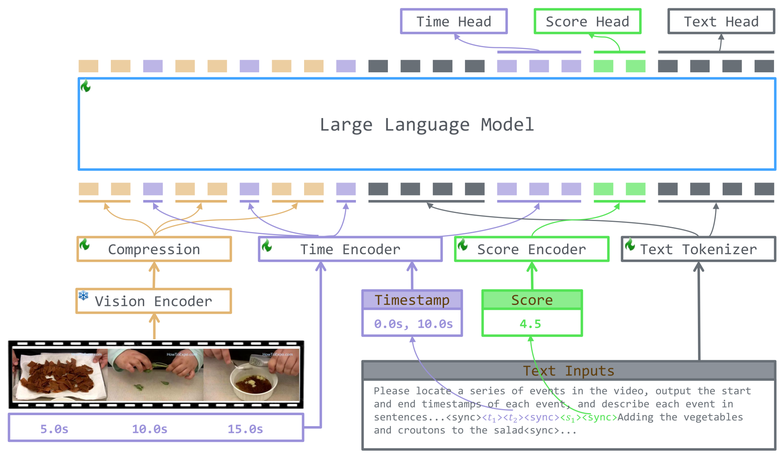
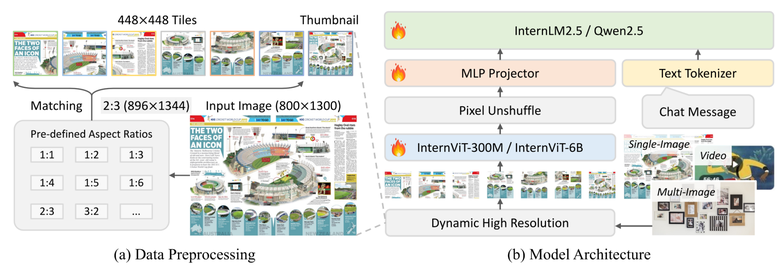


Figure 4. Our Temporal Analyzer is built on TRACE-Uni.

Captioner

Based on **InternVL-2.5-8B** [1], the Captioner ingests V, $\{M_t\}$, $\{(s_i,e_i)\}$, $\{c_i\}$, and P_{CoT}. Structured prompts enumerate **attributes, actions, statuses, interactions, environments**, then synthesize a temporally aware paragraph. CoT boosts **detail** and precision versus generic outputs.



(a) Data Preprocessing (b) Model Architecture

Figure 5. The Caption is built on InternVL-2.5.

Experiments

We use extensive qualitative experiments to demonstrate the versatility and effectiveness of CAT-V through various visual prompting, highlight styles, CoT prompting, and interactive chatting capabilities.

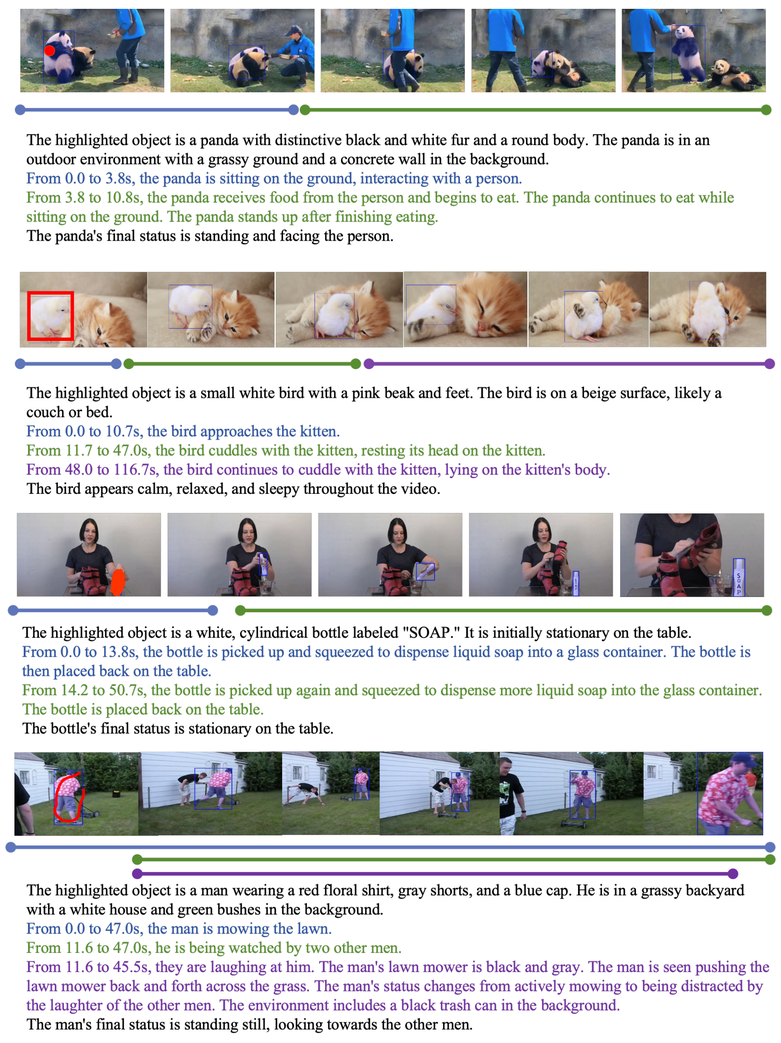


Figure 6. CAT-V's support for various visual prompting formats. The system effectively handles points, bounding boxes, and irregular regions to identify and track diverse objects, including pandas, birds, bottles, and people, demonstrating its flexibility and accuracy in accommodating different user input preferences.

Experiments (continue)

Figure 7. CAT-V can focus on different objects within the same video.

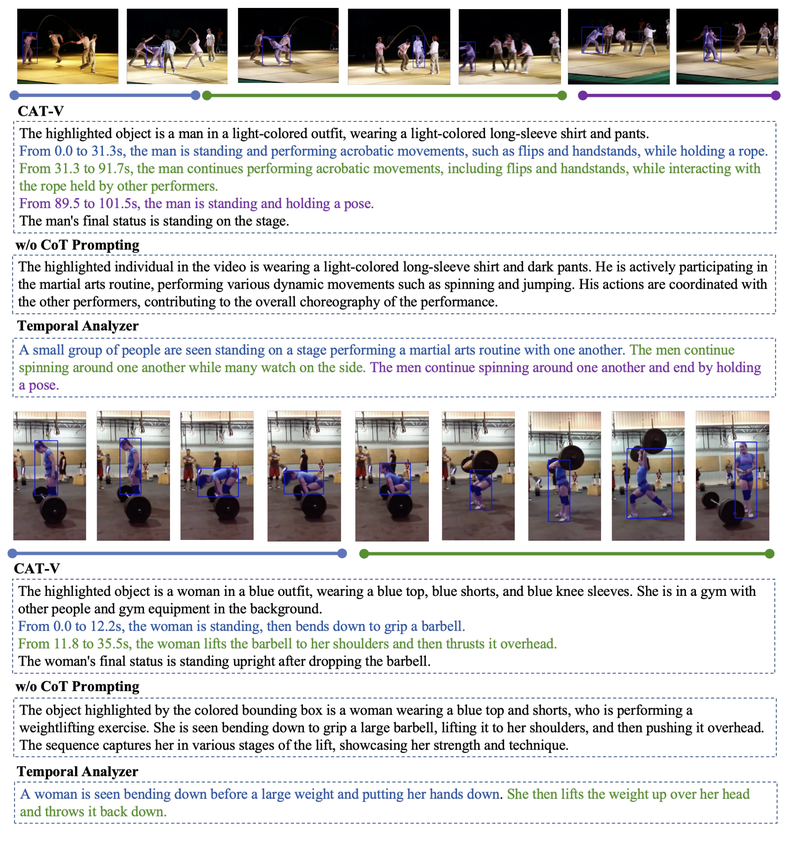


Figure 8. Comparison of object-centric video captioning using CAT-V with CoT prompting, without CoT prompting, and using only the Temporal Analyzer.

Figure 9. bounding boxes and polygons yield the most **accurate** captions, while color blocks/masks can distort appearance (e.g., color) and harm **fidelity**. In contrast, blur/circles better preserve color but provide weaker **spatial cues**.

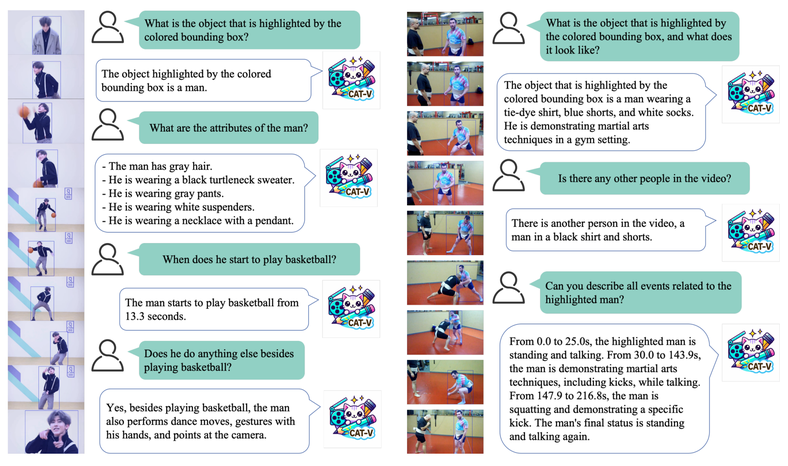


Figure 10. **Multi-round** object-centric dialogue. The system maintains HO reference, answers follow-ups about **attributes**, **actions**, and time spans, and supports switching targets via fresh **visual prompts**.

Reference

[1] Zhe Chen et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv*, 2024.
[2] Yongxin Guo et al. Trace: Temporal grounding video llm via causal event modeling. *ICLR*, 2025.
[3] Cheng-Yen Yang et al. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv*, 2024.

Demo Video: https://youtu.be/2eiPVKXEoxw

HF Daily Paper: https://huggingface.co/papers/2504.05541

GitHub Repo: https://github.com/yunlong10/CAT-V