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]Carnegie Mellon University
[4]Purdue University

Abstract

In this work, we introduce CAT-V (Caption Anything in Video), a training-free framework for fine-grained object-centric video captioning of user-selected instances. CAT-V combines (i) a SAMURAI-based Segmenter for precise object masks across frames, (ii) a TRACE-Uni Temporal Analyzer for event boundary detection and coarse event descriptions, and (iii) an InternVL-2.5 Captioner that, conditioned on spatiotemporal visual prompts and chain-of-thought (CoT) guidance, produces detailed, temporally coherent captions about object attributes, actions, states, interactions, and context. The system supports point, box, and region prompts and maintains temporal sensitivity by tracking object states across segments. Unlike overly abstract vanilla video captioning or a terse dense captioning, CAT-V achieves object-level specificity with spatial accuracy and temporal coherence through modular prompting. While built from pre-trained components, its training-free design balances interpretability, flexibility, and efficiency. The code is available at: https://github.com/yunlong10/CAT-V.

Introduction

Video captioning seeks coherent language descriptions of dynamic visual content. Prompted multimodal LLMs (MLLMs) (Tang et al. 2023) can perform vanilla video captioning, but typically yield abstract, video-level summaries that ignore object-level specificity and temporal variation (Huang et al. 2023; Zhang et al. 2021; Maaz et al. 2023; Li et al. 2023b). Dense video captioning localizes multi-events with captions, yet task-specific models (Wang et al. 2021) are often overly concise; MLLM-based variants fine-tuned on DVC data (Yang et al. 2023; Tang et al. 2024; Guo et al. 2024; Zhang et al. 2025; Krishna et al. 2017; Zhou, Xu, and Corso 2018) may compromise instruction following and still struggle with fine-grained, object-centric control. While controllable image captioning exists (Wang et al. 2023; Huang et al. 2024), controllable, fine-grained, object-centric captioning in videos remains underexplored. Attempts to integrate SAM with MLLMs (Yuan et al. 2025) typically rely on annotated data to train both components.

Contributions. We propose **CAT-V** (***C**aption **A**ny**T**hing in **V**ideo*), a training-free framework for object-centric video captioning that leverages a pre-trained segmentation model and a temporal-aware MLLM. CAT-V couples SAMURAI (Yang et al. 2024) (**Segmenter**) for robust, promptable video object segmentation; TRACE-Uni (Guo et al. 2024) (**Temporal Analyzer**) for event boundaries and coarse captions; and InternVL-2.5 (Chen et al. 2024) (**Captioner**) for detailed object-centric descriptions with spatiotemporal masks of the highlighted object; CoT prompting structures analysis of attributes, actions, states, interactions, and environments. CAT-V thus offers precise, temporally aware, user-controllable captions without fine-tuning.

Related Work

Dense video captioning localizes and describes events by modeling objects, space, and time. Early pipelines decoupled event localization and captioning (Krishna et al. 2017), while recent unified approaches such as PDVC (Zhu et al. 2022) and TRACE-Uni (Guo et al. 2024) couple timestamp prediction with caption generation. **Interactive video object segmentation**leverages user inputs for efficient mask propagation; SAM 2 introduces memory-based video awareness (Ravi et al. 2024), and SAMURAI extends it with motion-aware filtering for robust tracking (Yang et al. 2024). In **multimodal LLMs**, Flamingo (Alayrac et al. 2022) and BLIP-2 (Li et al. 2023a) established LLM-based vision–language training, and instruction tuning (Liu et al. 2024; Tang et al. 2024; Bi et al. 2024, 2023) improves downstream generalization, yet object-centric, temporally grounded video description remains limited. Controllable captioning has been explored mainly for images (Huang et al. 2024), and efforts to pair SAM with MLLMs typically rely on annotated training (Yuan et al. 2025; Hua et al. 2024). In contrast, **CAT-V** is *training-free* and *object-centric*: it unifies SAMURAI-based segmentation, TRACE-Uni temporal parsing, and InternVL-2.5 captioning via spatiotemporal mask injection and CoT prompting, yielding spatially precise, temporally coherent, user-controllable captions without fine-tuning.

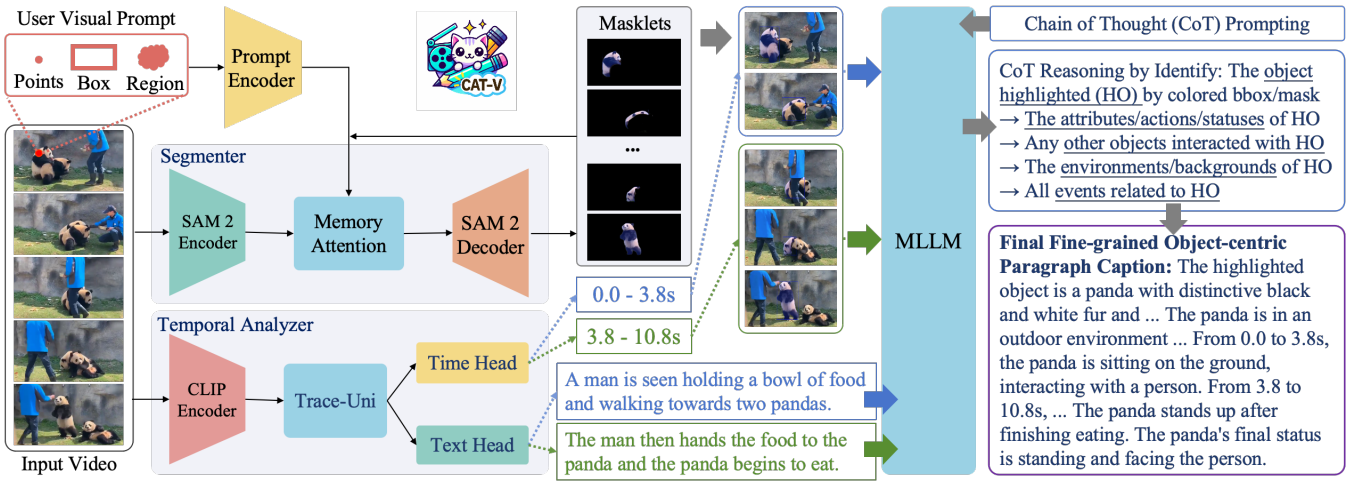

Figure 1: **CAT-V** comprises three modules: *Segmenter*, *Temporal Analyzer*, and *Captioner*.

CAT-V: Caption Anything in Video

CAT-V performs fine-grained, object-centric captioning via spatiotemporal multimodal prompting with three modules: Segmenter $\mathcal{S}$, Temporal Analyzer $\mathcal{T}$, and Captioner $\mathcal{C}$ (Figure 1). Given a video $V = \{I_t\}_{t=1}^{T}$ and user prompt p (points, boxes, or regions), CAT-V proceeds as follows.

Segmenter. $\mathcal{S}$ uses SAMURAI (Yang et al. 2024) to segment the target object from user-provided prompts. For each frame I_t, it outputs a binary mask $M_t = \mathcal{S}(I_t, p)$ with $M_t \in \{0,1\}^{H\times W}$. SAM 2's encoder/decoder (Ravi et al. 2024) and a prompt encoder yield frame embeddings and prompt features; SAMURAI augments these with Kalman filtering and motion-aware memory, improving robustness under occlusion, motion blur, and clutter.

Temporal Analyzer. $\mathcal{T}$, built on TRACE-Uni (Guo et al. 2024), models temporal dynamics hierarchically, producing N events with boundaries $\{(s_i, e_i)\}_{i=1}^{N}$ and coarse captions $c_i = \mathcal{T}(V, s_i, e_i)$. This decomposition exposes interactions and activities at multiple time scales.

Captioner. $\mathcal{C}$ (InternVL-2.5-8B (Chen et al. 2024)) conditions on the input video V, masks $\{M_t\}_{t=1}^{T}$, event tuples $\{(s_i, e_i, c_i)\}_{i=1}^{N}$, and CoT prompts P_{CoT} to produce spatially precise, temporally coherent object-centric captions:

$$C_{\text{final}} = \mathcal{C}\Big(V(\{M_t\}_{t=1}^{T}, f), \{(s_i, e_i, c_i)\}_{i=1}^{N}, P_{\text{CoT}}\Big),$$

where f controls mask injection into V.

Chain-of-Thought Prompting. We design structured prompts $P_{\text{CoT}} = \{A_1, \ldots, A_K\}$ that cover attributes, actions, state changes, interactions, environments, and related events. This modular analysis reduces omissions and yields a coherent, temporally grounded narrative.

CAT-V's Capabilities. CAT-V supports: (i) object-centric captioning for different user-selected targets within the same video with precise temporal segmentation of actions and states; (ii) support for diverse visual prompts (points, bounding boxes, irregular regions) with robust tracking and consistent caption accuracy; (iii) spatiotemporal mask injection that effectively guides the Captioner (boxes/polygons most accurate); (iv) CoT prompting that upgrades generic summaries to detailed, time-stamped narratives; (v) interactive multi-turn, object-centric chatting for follow-up queries on attributes, actions, and temporal behaviors.

Chain-of-Thought Prompt

Above are the event captions given by the user, whose timestamps are very accurate but the subjects of the sentences are not necessarily what we want to highlight. Please pay attention to the **object highlighted (HO)** by colored bounding box and blue mask in the video frames, and generate accurate object-centric caption for the HO. Please make sure in object-centric paragraph caption, the sentences should be detailed and specific, and the subjects of all sentences MUST be HO. Please follow the format:

HO: ..., **HO's attributes**: ..., **All actions done by HO**: ..., **All statuses of HO**: ..., **All other objects interacted with HO**: ..., **All environments/backgrounds of HO**: ..., **All events related to HO**: ..., **Final object-centric paragraph caption**: The HO is [attr.], [env.]. From ... to ...s, the HO [status], [any action], [any status/attr./environment changes]... From ... to ...s, the HO [status], [any action], [any status/attr./env. changes]. The OH's [final status] is ...

Conclusion

We propose CAT-V, a training-free framework for object-centric video captioning. It supports flexible visual prompts, maintains consistent object tracking, and produces detailed, temporally coherent captions without additional training, enabling intuitive, interactive video understanding.

Acknowledgements

This work was supported by Sony Group Corporation. We would like to thank Sayaka Nakamura and Jerry Jun Yokono for their insightful discussion.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bi, J.; Nguyen, N. M.; Vosoughi, A.; and Xu, C. 2023. MISAR: A Multimodal Instructional System with Augmented Reality. arXiv:2310.11699.

Bi, J.; Tang, Y.; Song, L.; Vosoughi, A.; Nguyen, N.; and Xu, C. 2024. EAGLE: Egocentric AGgregated Language-video Engine. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 1682–1691. ACM.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Guo, Y.; Liu, J.; Li, M.; Tang, X.; Liu, Q.; and Chen, X. 2024. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*.

Hua, H.; Liu, Q.; Zhang, L.; Shi, J.; Zhang, Z.; Wang, Y.; Zhang, J.; and Luo, J. 2024. FINECAPTION: Compositional Image Captioning Focusing on Wherever You Want at Any Granularity. *arXiv preprint arXiv:2411.15411*.

Huang, C.; Tian, Y.; Kumar, A.; and Xu, C. 2023. Egocentric Audio-Visual Object Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22910–22921.

Huang, X.; Wang, J.; Tang, Y.; Zhang, Z.; Hu, H.; Lu, J.; Wang, L.; and Liu, Z. 2024. Segment and caption anything. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13405–13417.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Tang, Y.; Shimada, D.; Bi, J.; Feng, M.; Hua, H.; and Xu, C. 2024. Empowering LLMs with Pseudo-Untrimmed Videos for Audio-Visual Temporal Understanding. *arXiv preprint arXiv:2403.16276*.

Wang, T.; Zhang, J.; Fei, J.; Zheng, H.; Tang, Y.; Li, Z.; Gao, M.; and Zhao, S. 2023. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*.

Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; and Luo, P. 2021. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6847–6857.

Yang, A.; Nagrani, A.; Seo, P. H.; Miech, A.; Pont-Tuset, J.; Laptev, I.; Sivic, J.; and Schmid, C. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10714–10726.

Yang, C.-Y.; Huang, H.-W.; Chai, W.; Jiang, Z.; and Hwang, J.-N. 2024. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*.

Yuan, H.; Li, X.; Zhang, T.; Huang, Z.; Xu, S.; Ji, S.; Tong, Y.; Qi, L.; Feng, J.; and Yang, M.-H. 2025. Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos. *arXiv preprint arXiv:2501.04001*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, Y.; Liang, S.; Yang, S.; Liu, X.; Wu, Z.; Shan, S.; and Chen, X. 2021. Unicon: Unified context network for robust active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia*, 3964–3972.

Zhou, L.; Xu, C.; and Corso, J. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhu, W.; Pang, B.; Thapliyal, A. V.; Wang, W. Y.; and Soricut, R. 2022. End-to-end dense video captioning as sequence generation. *arXiv preprint arXiv:2204.08121*.