

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

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

<sup>1</sup> University of Rochester, <sup>2</sup> Sony Group Corporation, <sup>3</sup> CMU, <sup>4</sup> Purdue University



AAAI-26 / IAAI-26 / EAAI-26  
 JANUARY 20-27, 2026, SINGAPORE

## 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**.

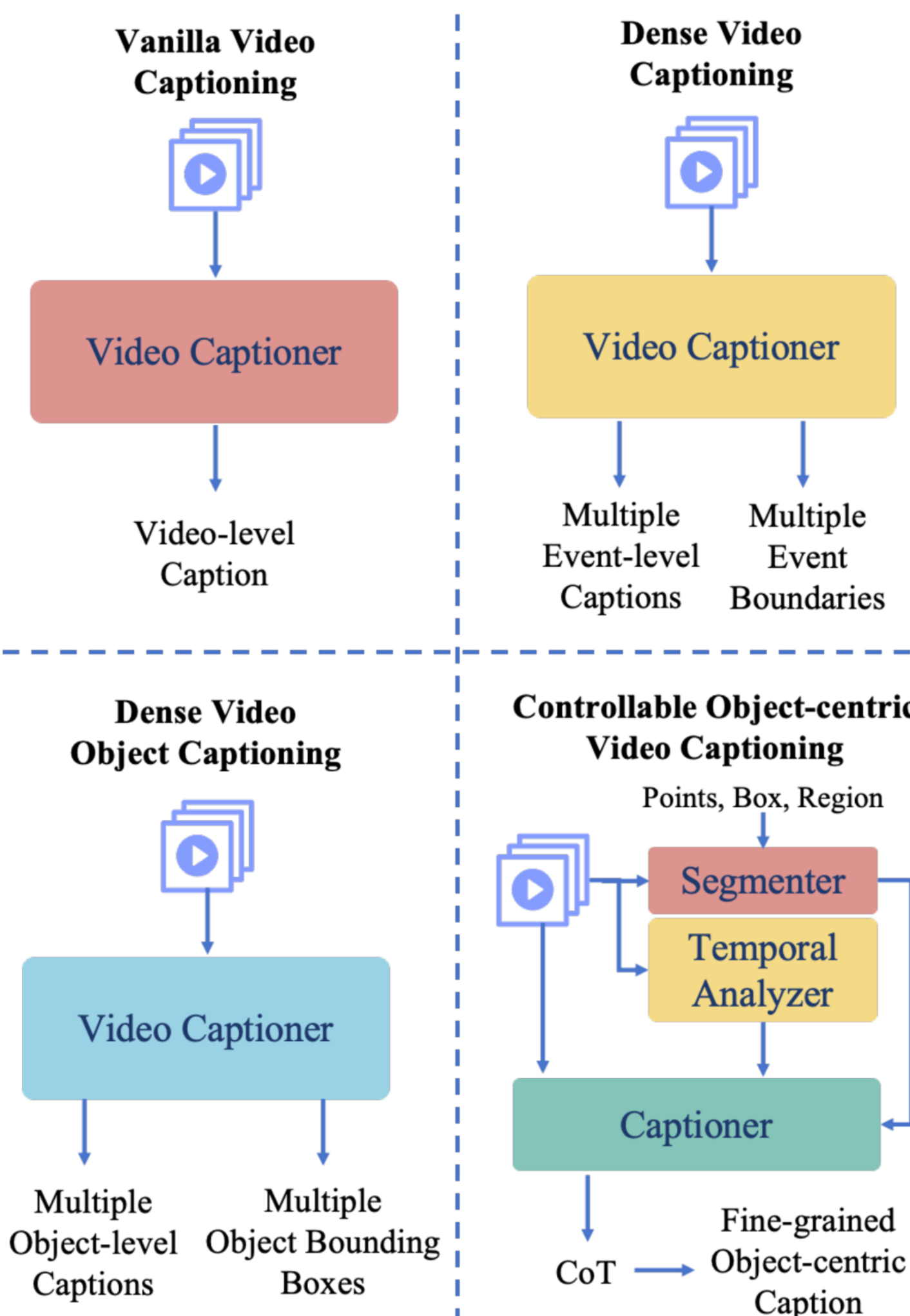


Figure 1. Comparison of video captioning approaches.

## 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.

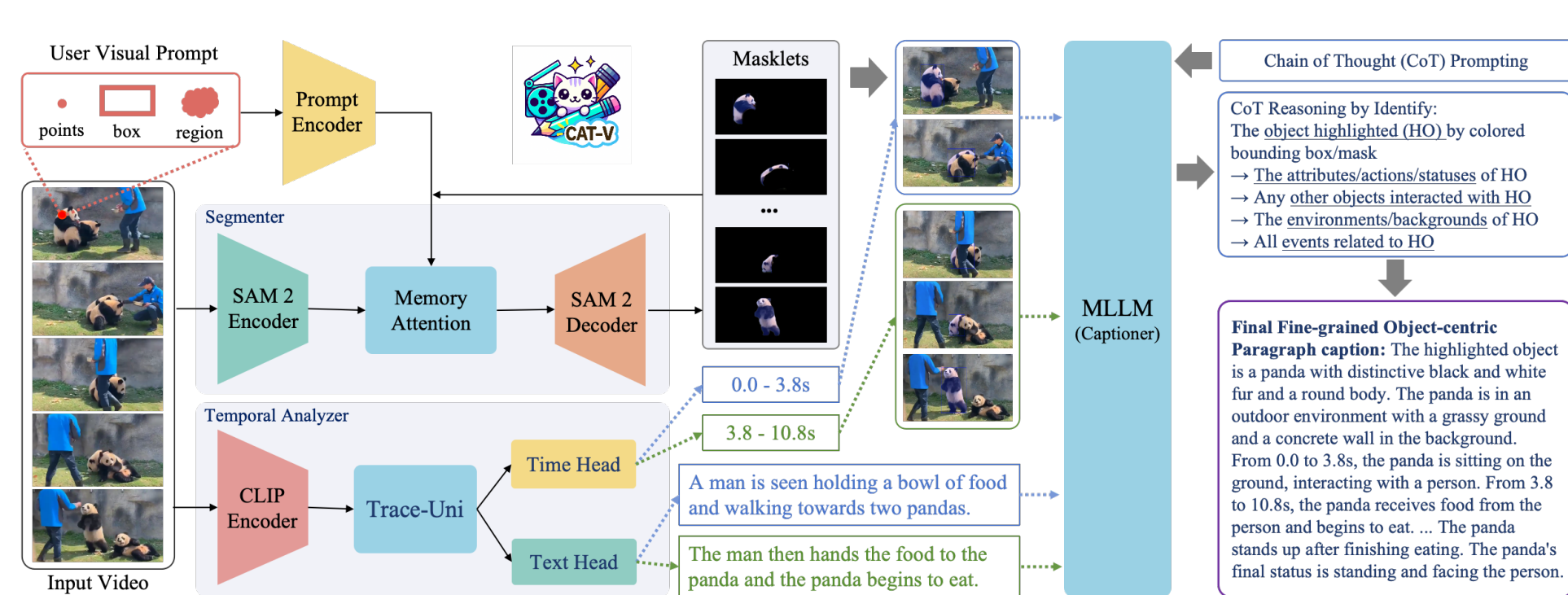


Figure 2. CAT-V Framework.

## 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.

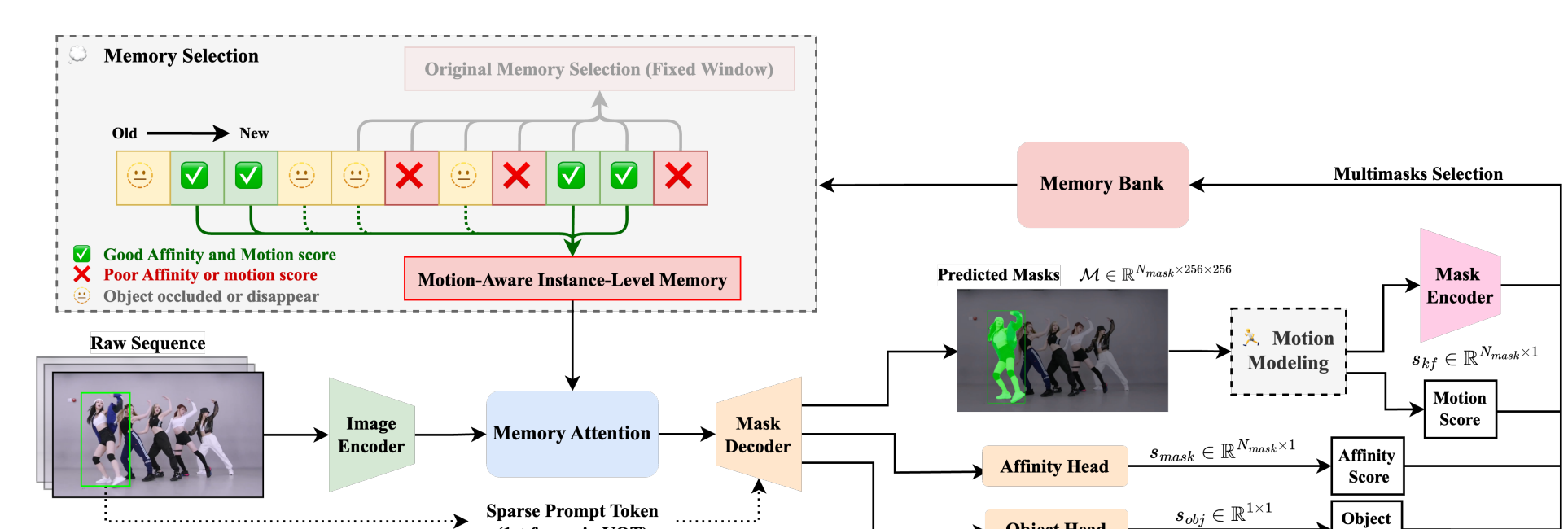


Figure 3. Our Segmenter is built on SAMURAI model.

## 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.

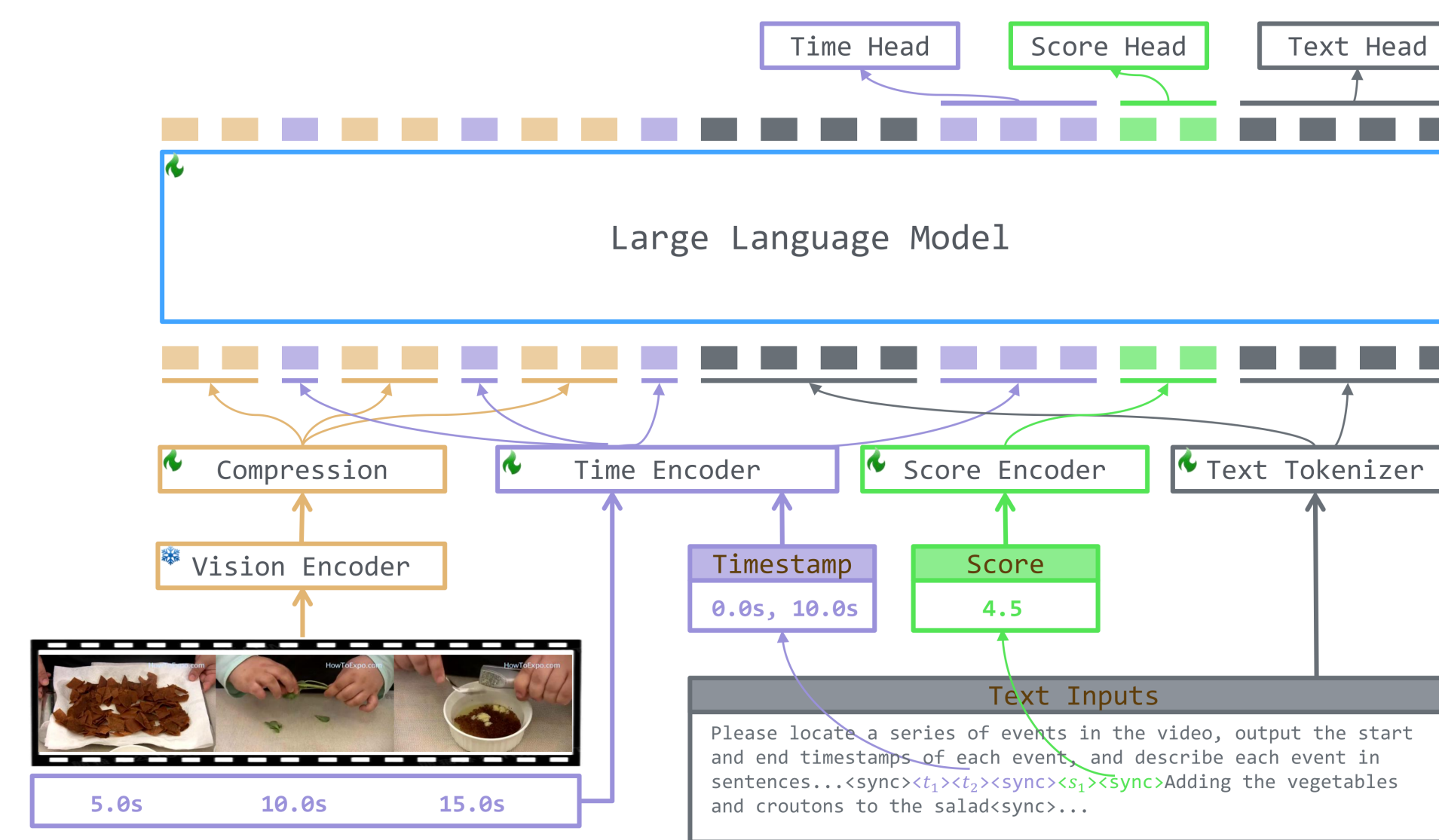


Figure 4. Our Temporal Analyzer is built on TRACE-Uni.

## 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.

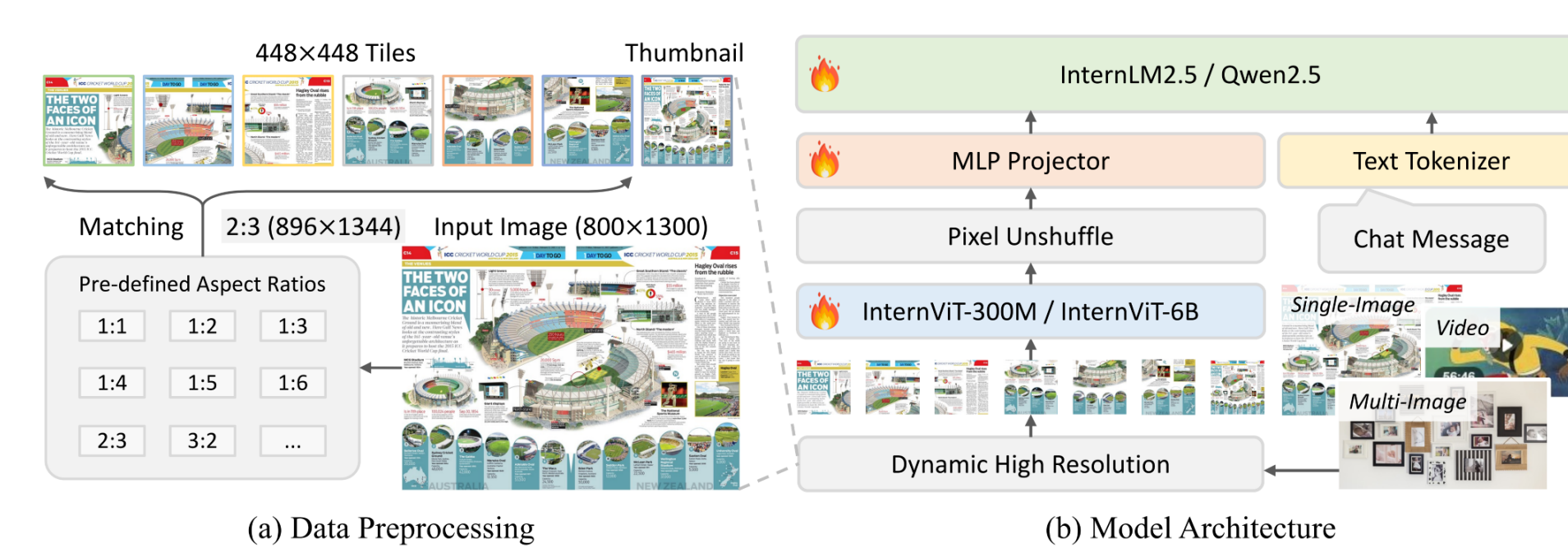
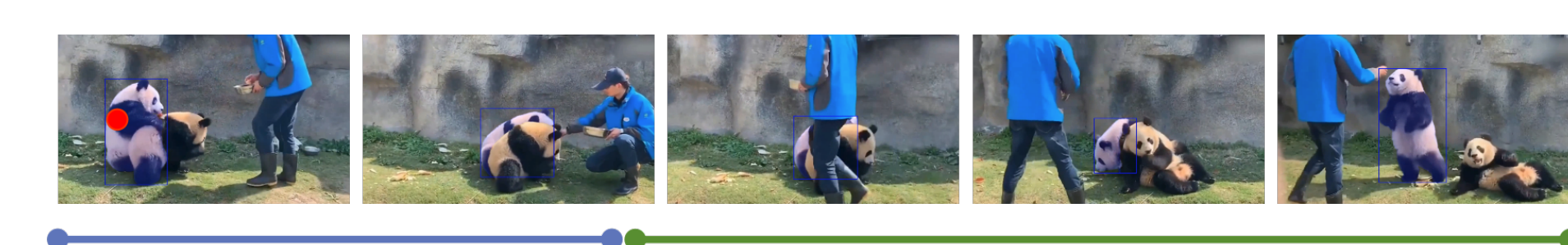


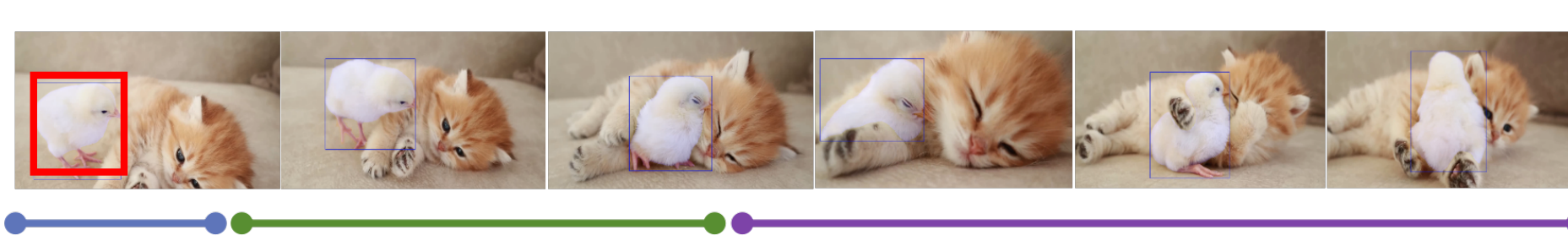
Figure 5. The Captioner is built on InternVL-2.5.

## Experiments

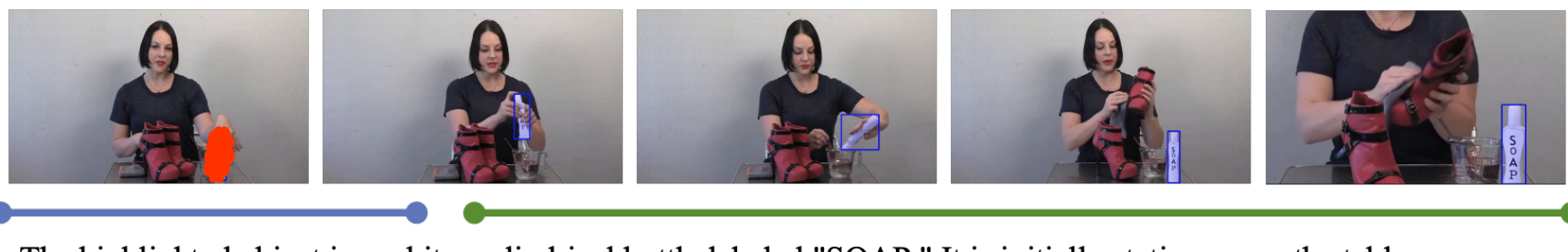
We use extensive qualitative experiments to demonstrate the versatility and effectiveness of CAT-V through various visual prompting, highlight styles, CoT prompting, and interactive chatting capabilities.



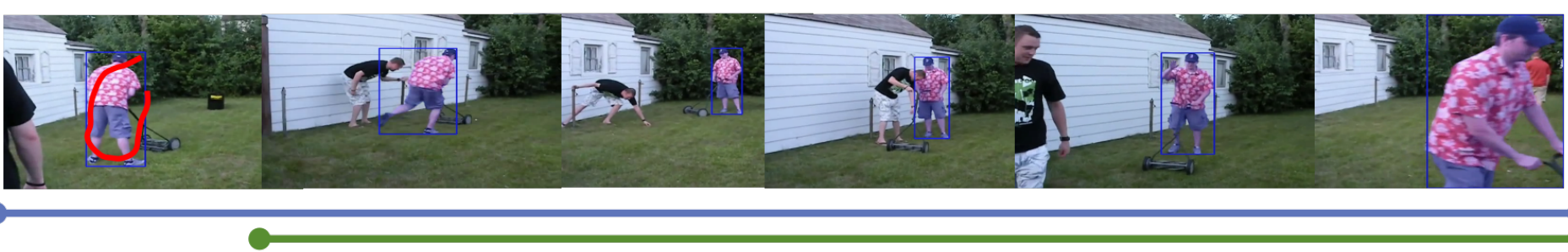
The highlighted object is a panda with distinctive black and white fur and a round body. The panda is in an outdoor environment with a grassy ground and a concrete wall in the background.  
 From 0.0 to 3.8s, the panda is sitting on the ground, interacting with a person.  
 From 3.8 to 10.8s, the panda receives food from the person and begins to eat. The panda continues to eat while sitting on the ground. The panda stands up after finishing eating.  
 The panda's final status is standing and facing the person.



The highlighted object is a small white bird with a pink beak and feet. The bird is on a beige surface, likely a couch or bed.  
 From 0.0 to 10.7s, the bird approaches the kitten.  
 From 11.7 to 47.0s, the bird cuddles with the kitten, resting its head on the kitten.  
 From 48.0 to 116.7s, the bird continues to cuddle with the kitten, lying on the kitten's body.  
 The bird appears calm, relaxed, and sleepy throughout the video.



The highlighted object is a white, cylindrical bottle labeled "SOAP." It is initially stationary on the table.  
 From 0.0 to 13.8s, the bottle is picked up and squeezed to dispense liquid soap into a glass container. The bottle is then placed back on the table.  
 From 14.2 to 50.7s, the bottle is picked up again and squeezed to dispense more liquid soap into the glass container. The bottle is placed back on the table.  
 The bottle's final status is stationary on the table.



The highlighted object is a man wearing a red floral shirt, gray shorts, and a blue cap. He is in a grassy backyard with a white house and green bushes in the background.  
 From 0.0 to 47.0s, the man is mowing the lawn.  
 From 11.6 to 47.0s, he is being watched by two other men.  
 From 11.6 to 45.5s, they are laughing at him. The man's lawn mower is black and gray. The man is seen pushing the lawn mower back and forth across the grass. The man's status changes from actively mowing to being distracted by the laughter of the other men. The environment includes a black trash can in the background.  
 The man's final status is standing still, looking towards the other men.

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)

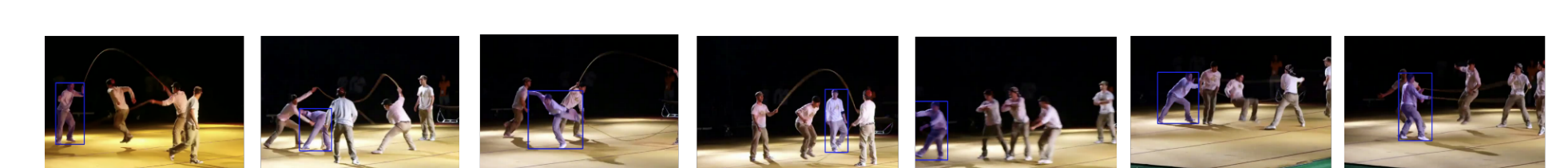


The highlighted object is a horse, wearing a saddle and bridle, in a fenced dirt area with trees and other horses in the background.  
 From 0.0 to 21.1s, the horse moves through the fenced area while a man rides it.  
 From 21.1 to 55.6s, the horse stands still while the man dismounts, ties a calf, and walks back to the horse.  
 The horse's final status is standing still.



The highlighted object is a man wearing a shirt, blue jeans, and a black hat, in a dusty outdoor ranch with fences, trees, and other cattle.  
 From 0.0 to 21.1s, the man is riding a brown horse and swinging a rope around.  
 From 18.1 to 37.6s, the man throws the rope onto a calf and jumps off the horse to tie up the calf.  
 From 37.6 to 55.6s, the man walks back to the horse, mounts it, and prepares to ride again.

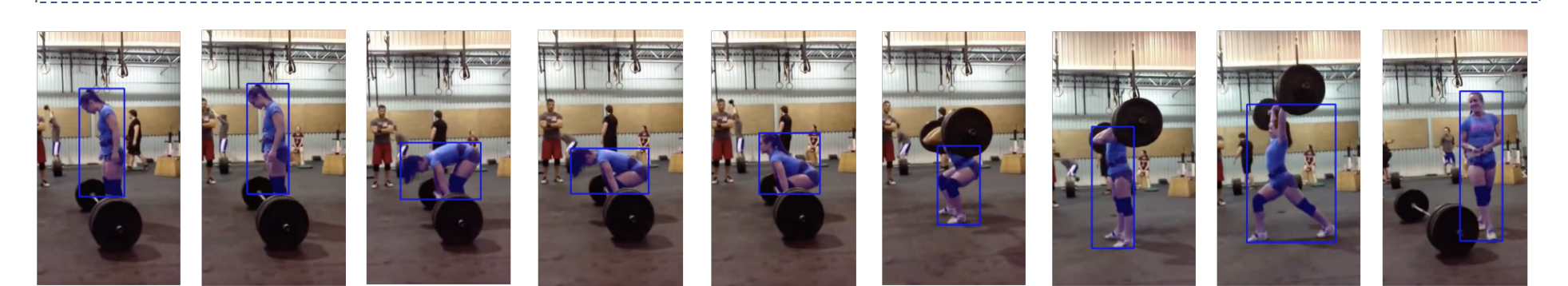
Figure 7. CAT-V can focus on different objects within the same video.



The highlighted object is a man in a light-colored outfit, wearing a light-colored long-sleeve shirt and pants.  
 From 0.0 to 31.3s, the man is standing and performing acrobatic movements, such as flips and handstands, while holding a rope.  
 From 31.3 to 91.7s, the man continues performing acrobatic movements, including flips and handstands, while interacting with the rope held by other performers.  
 From 89.5 to 101.5s, the man is standing and holding a pose.  
 The man's final status is standing on the stage.

**w/o CoT Prompting**  
 The highlighted individual in the video is wearing a light-colored long-sleeve shirt and dark pants. He is actively participating in the martial arts routine, performing various dynamic movements such as spinning and jumping. His actions are coordinated with the other performers, contributing to the overall choreography of the performance.

**Temporal Analyzer**  
 A small group of people are seen standing on a stage performing a martial arts routine with one another. The men continue spinning around one another while many watch on the side. The men continue spinning around one another and end by holding a pose.



The highlighted object is a woman in a blue outfit, wearing a blue top, blue shorts, and blue knee sleeves. She is in a gym with other people and gym equipment in the background.  
 From 0.0 to 12.2s, the woman is standing, then bends down to grip a barbell.  
 From 11.8 to 35.5s, the woman lifts the barbell to her shoulders and then thrusts it overhead.  
 The woman's final status is standing upright after dropping the barbell.

**w/o CoT Prompting**  
 The object highlighted by the colored bounding box is a woman wearing a blue top and shorts, who is performing a weightlifting exercise. She is seen bending down to grip a large barbell, lifting it to her shoulders, and then pushing it overhead. The sequence captures her in various stages of the lift, showcasing her strength and technique.

**Temporal Analyzer**  
 A woman is seen bending down before a large weight and putting her hands down. She then lifts the weight up over her head and throws it back down.

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

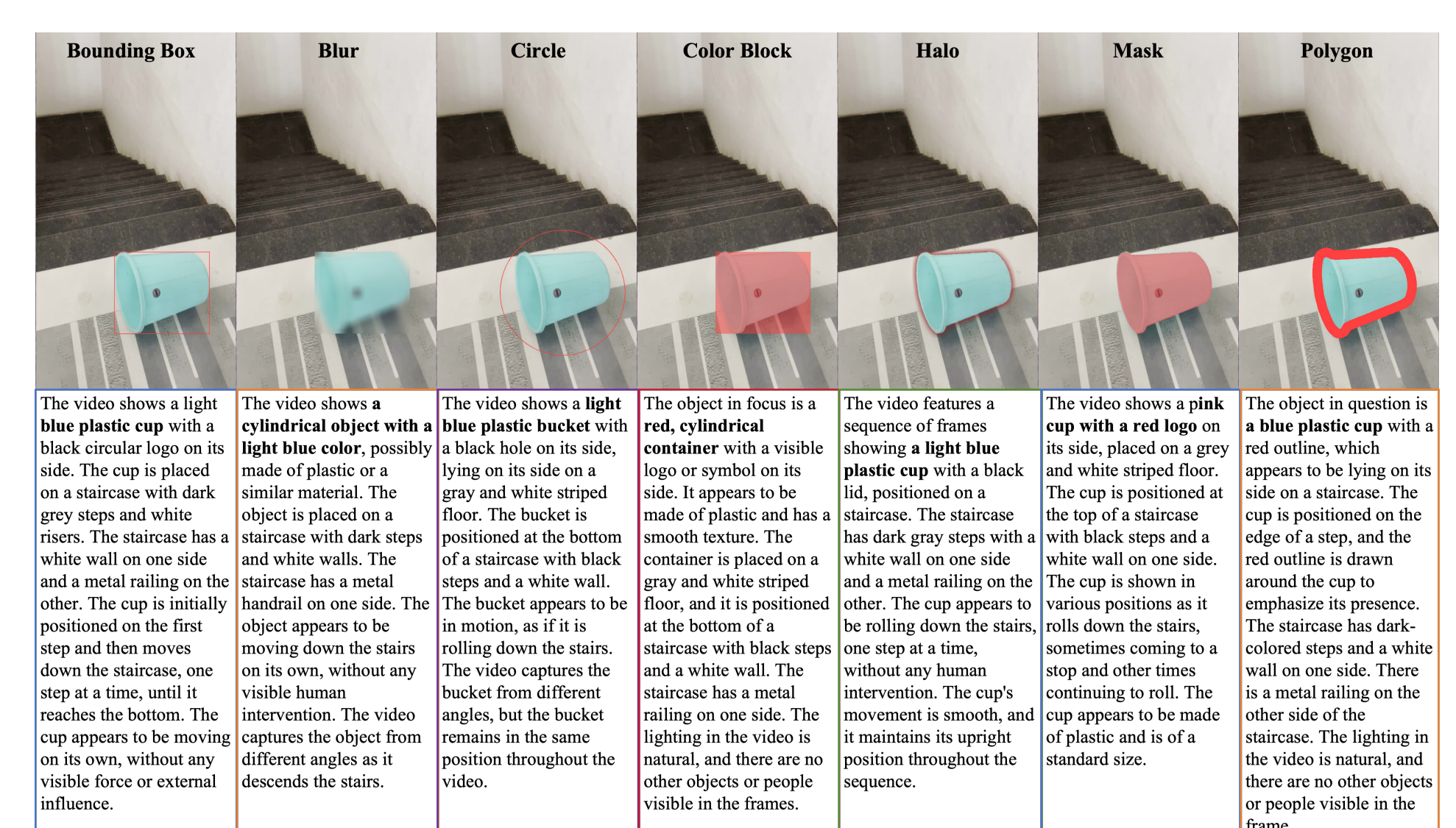


Figure 9. bounding boxes and polygons yield the most accurate captions, while color blocks/masks can distort appearance (e.g., color) and harm fidelity. In contrast, blur/circles better preserve color but provide weaker spatial cues.

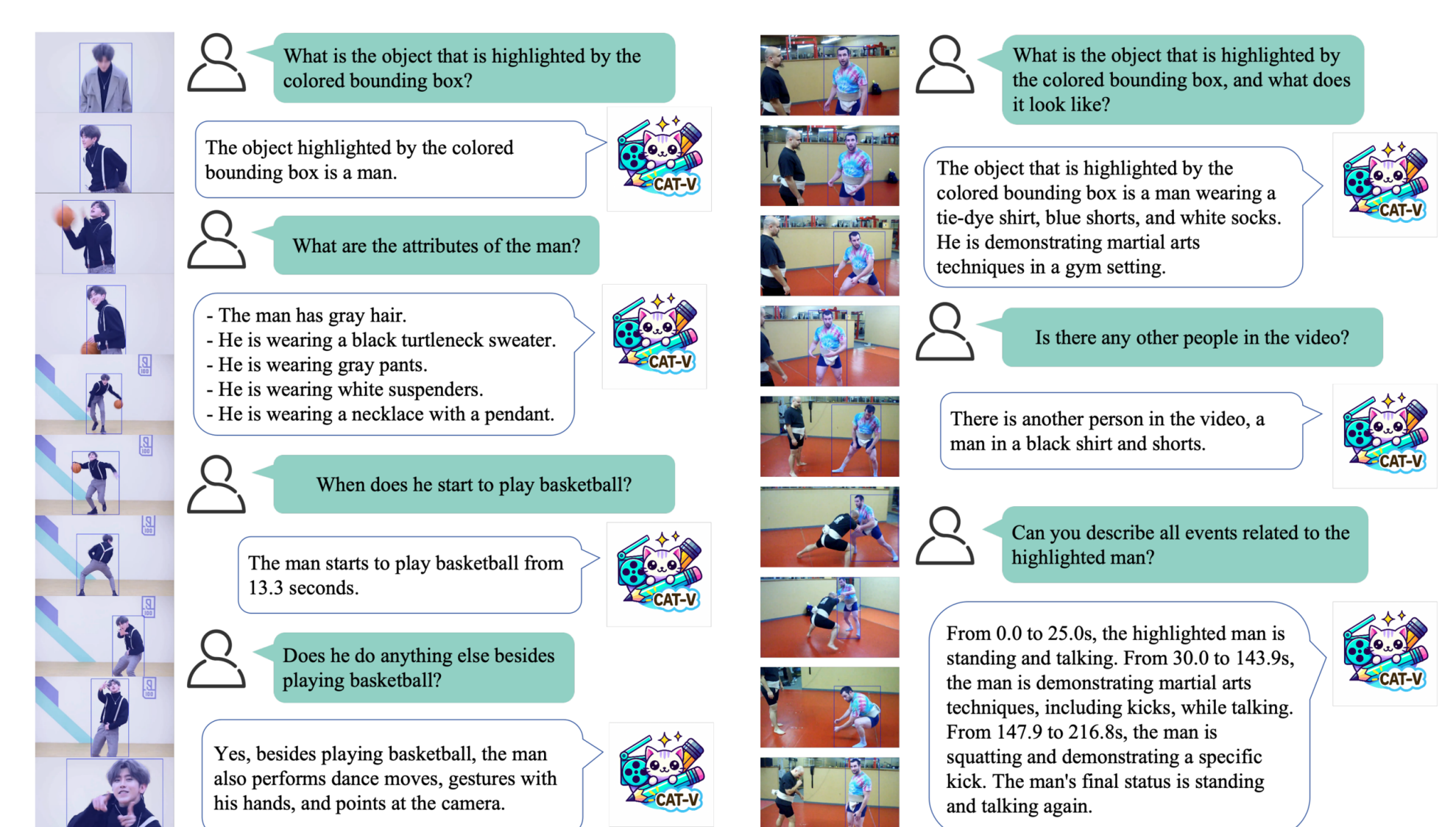


Figure 10. Multi-round object-centric dialogue. The system maintains HO reference, answers follow-ups about attributes, actions, and time spans, and supports switching targets via fresh visual prompts.

## Reference

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