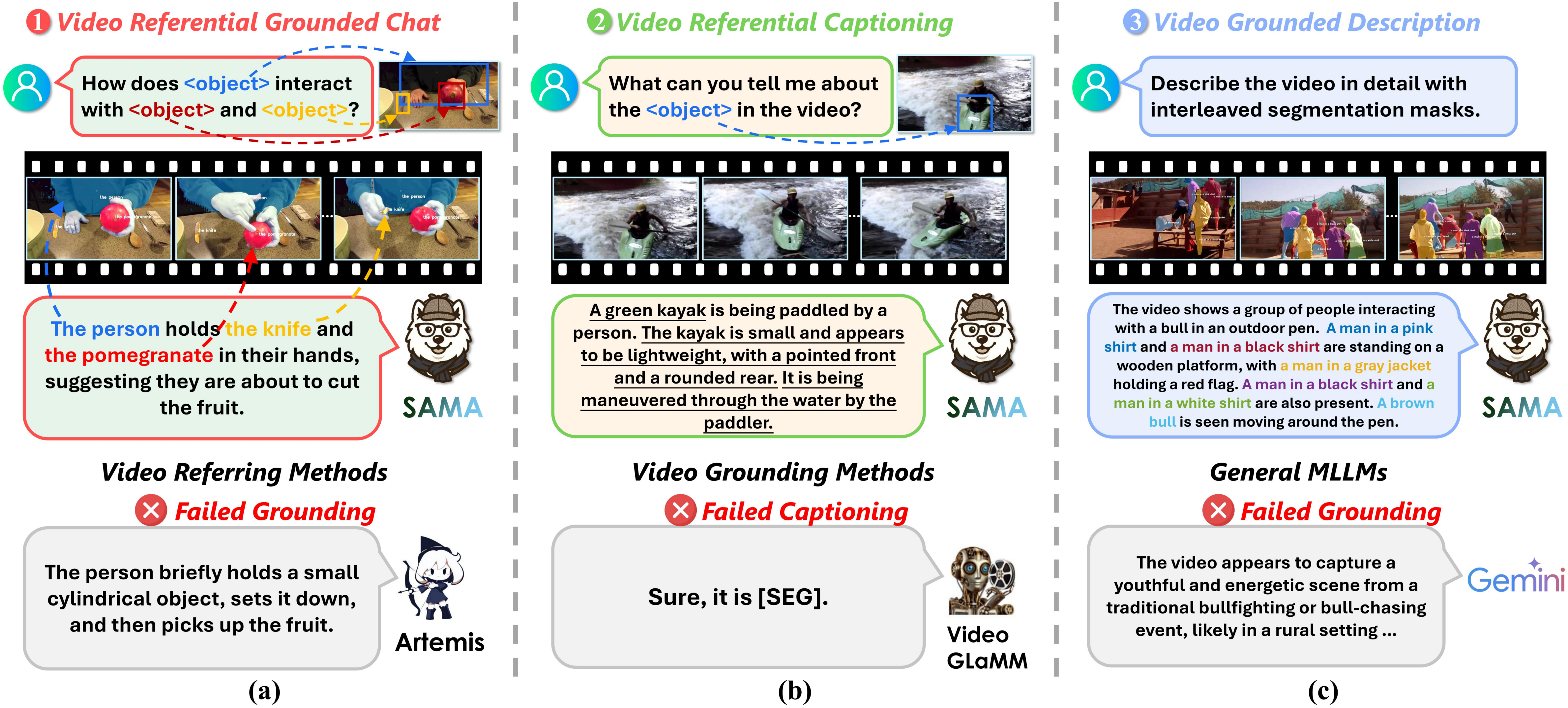
NEURAL INFORMATION
PROCESSING SYSTEMS

SAMA: Towards Multi-Turn Referential Grounded Video Chat with Large Language Models

Ye Sun · Hao Zhang · Henghui Ding · Tiehua Zhang · Xingjun Ma · Yu-Gang Jiang

Fudan University · HKUST · Tongji University

Fine-grained video understanding capabilities of SAMA



How does <object> interact with <object> and <object>?

The person holds the knife and the pomegranate in their hands, suggesting they are about to cut the fruit.

The person briefly holds a small cylindrical object, sets it down, and then picks up the fruit.

(a)

What can you tell me about the <object> in the video?

A green kayak is being paddled by a person. The kayak is small and appears to be lightweight, with a pointed front and a rounded rear. It is being maneuvered through the water by the paddler.

Sure, it is [SEG].

(b)

Describe the video in detail with interleaved segmentation masks.

The video shows a group of people interacting with a bull in an outdoor pen. A man in a pink shirt and a man in a black shirt are standing on a wooden platform, with a man in a gray jacket holding a red flag. A man in a black shirt and a man in a white shirt are also present. A brown bull is seen moving around the pen.

The video appears to capture a youthful and energetic scene from a traditional bullfighting or bull-chasing event, likely in a rural setting ...

(c)

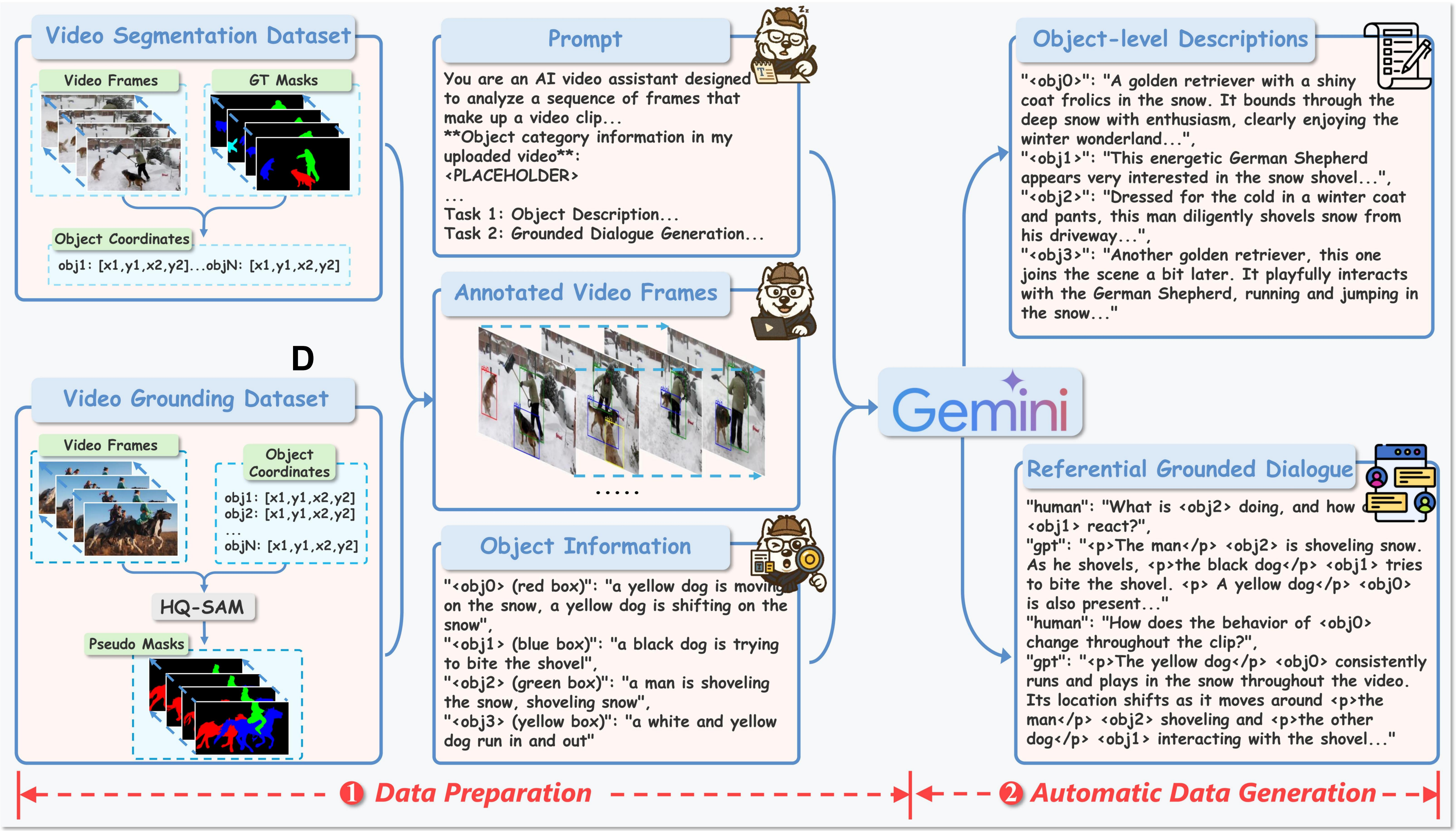
Contributions of SAMA

I. SAMA-239K Dataset: A large-scale dataset comprising 15K videos specifically curated to enable joint learning of video referring understanding, grounding, and multi-turn video chat.

II. SAMA Model: Incorporating a versatile spatio-temporal context aggregator and the SAM to jointly enhance fine-grained video comprehension and precise grounding capabilities.

III. SAMA-Bench: A meticulously designed benchmark consisting of 5,067 questions from 522 videos, to comprehensively evaluate the integrated capabilities of Video LMMs in multi-turn, spatio-temporal referring understanding and grounded dialogue.

Data creation pipeline of SAMA-239K



Prompt

You are an AI video assistant designed to analyze a sequence of frames that make up a video clip...
Object category information in my uploaded video:
<PLACEHOLDER>
...
Task 1: Object Description...
Task 2: Grounded Dialogue Generation...

Object Information

"<obj0> (red box)": "a yellow dog is moving on the snow, a yellow dog is shifting on the snow",
"<obj1> (blue box)": "a black dog is trying to bite the shovel",
"<obj2> (green box)": "a man is shoveling the snow, shoveling snow",
"<obj3> (yellow box)": "a white and yellow dog run in and out"

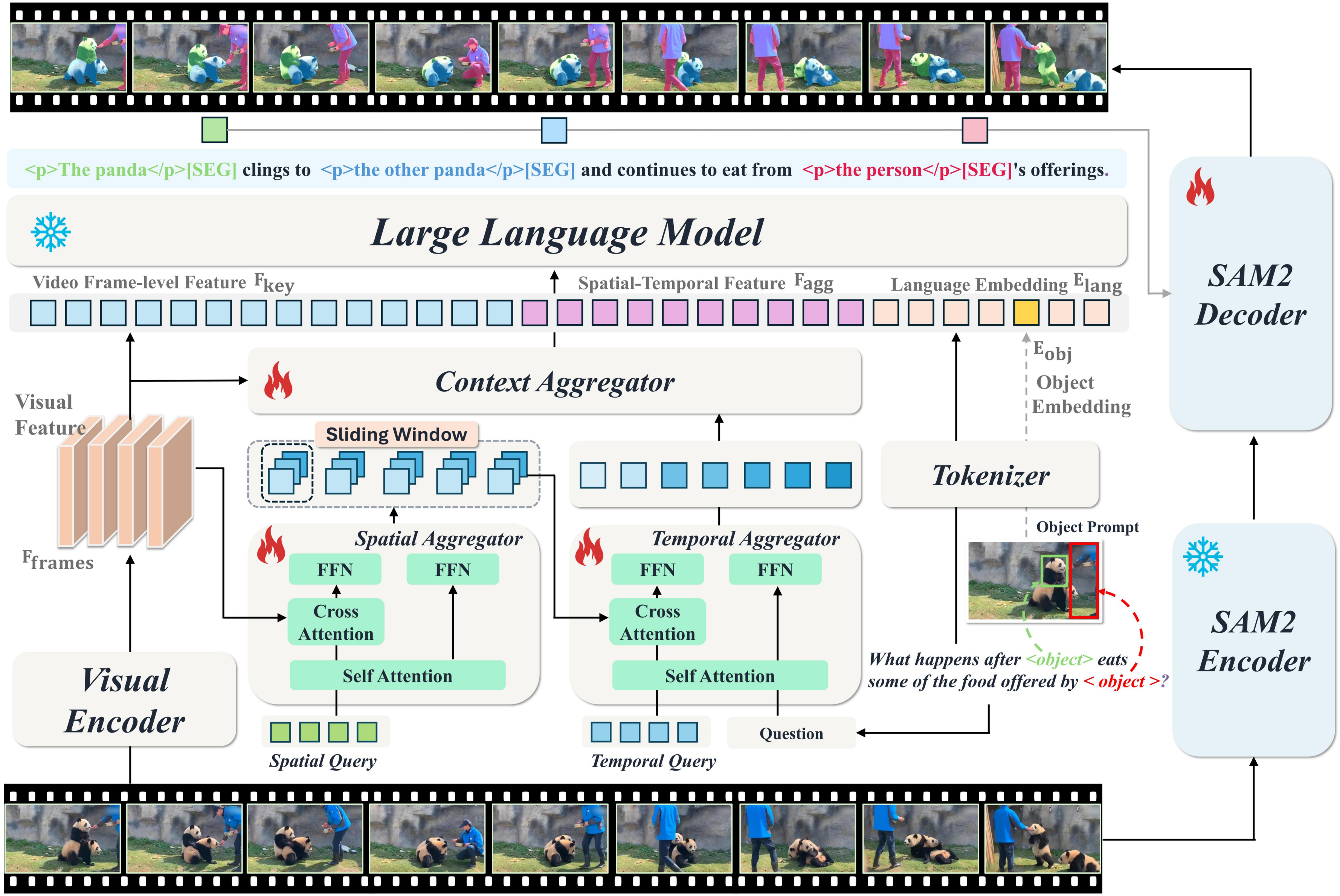
Object-level Descriptions

"<obj0>": "A golden retriever with a shiny coat frolics in the snow. It bounds through the deep snow with enthusiasm, clearly enjoying the winter wonderland...",
"<obj1>": "This energetic German Shepherd appears very interested in the snow shovel...",
"<obj2>": "Dressed for the cold in a winter coat and pants, this man diligently shovels snow from his driveway...",
"<obj3>": "Another golden retriever, this one joins the scene a bit later. It playfully interacts with the German Shepherd, running and jumping in the snow..."

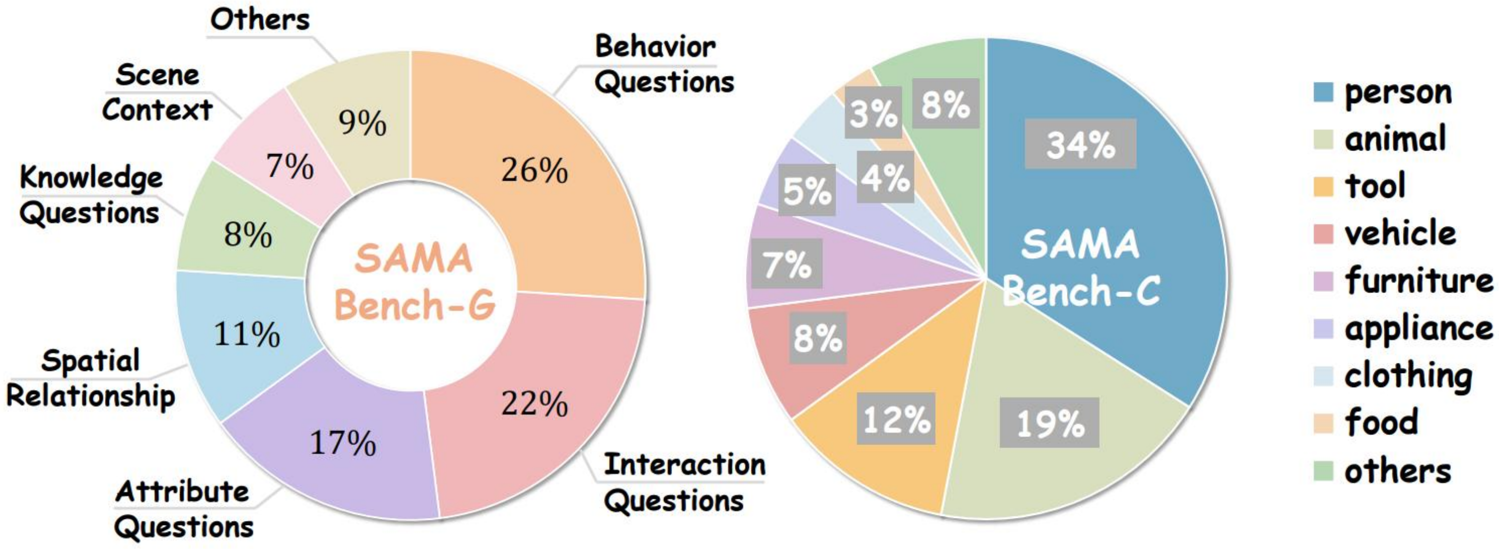
Referential Grounded Dialogue

"human": "What is <obj2> doing, and how d
<obj1> react?",
"gpt": "<p>The man</p> <obj2> is shoveling snow. As he shovels, <p>the black dog</p> <obj1> tries to bite the shovel. <p> A yellow dog</p> <obj0> is also present..."
"human": "How does the behavior of <obj0> change throughout the clip?",
"gpt": "<p>The yellow dog</p> <obj0> consistently runs and plays in the snow throughout the video. Its location shifts as it moves around <p>the man</p> <obj2> shoveling and <p>the other dog</p> <obj1> interacting with the shovel..."

SAMA Framework



Data characteristics of SAMA-Bench



(a) Question types in **Bench-G** (b) Category list in **Bench-C**

Experiments

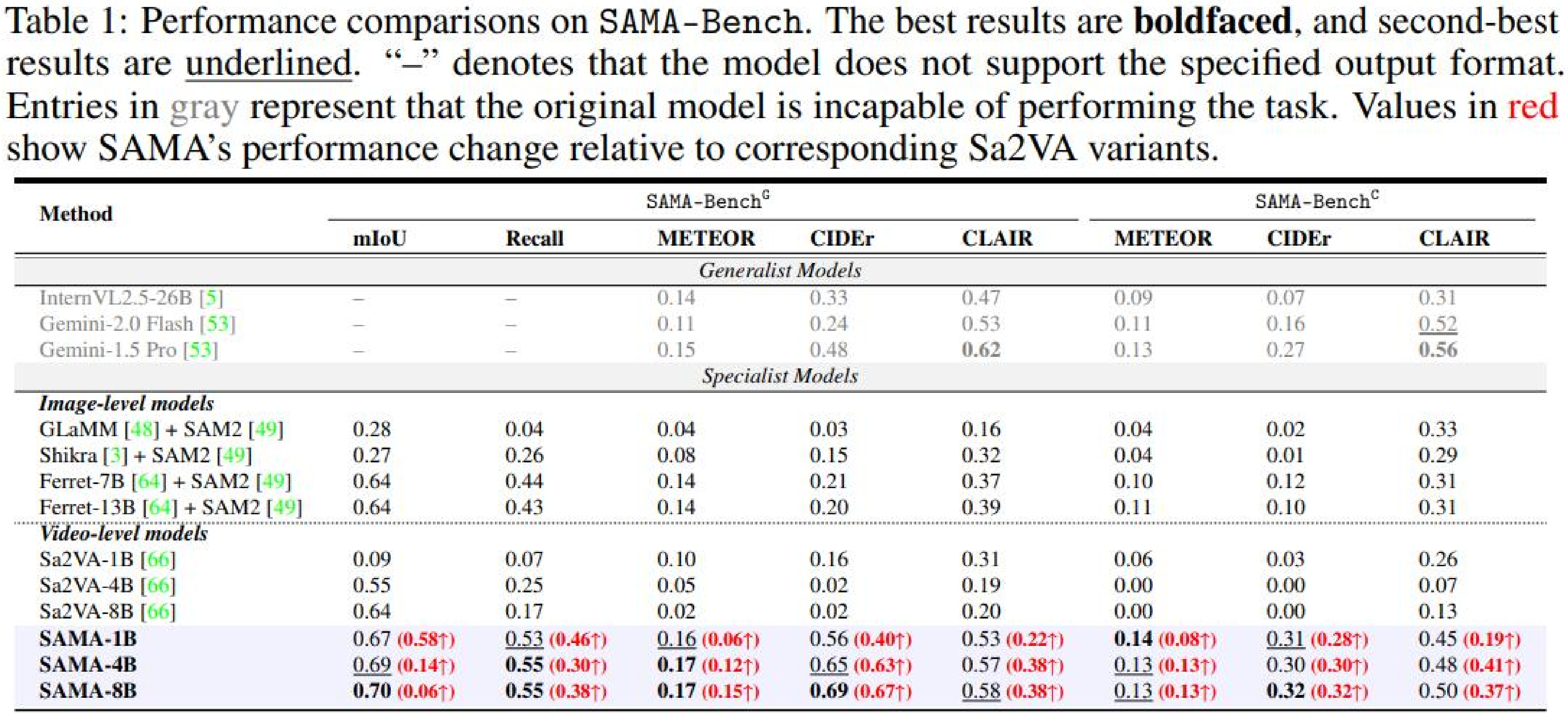


Table 1: Performance comparisons on `SAMA-Bench`. The best results are **boldfaced**, and second-best results are underlined. "–" denotes that the model does not support the specified output format. Entries in gray represent that the original model is incapable of performing the task. Values in red show SAMA's performance change relative to corresponding Sa2VA variants.

Method	`SAMA-Bench`^{G}					`SAMA-Bench`^{C}		
	mIoU	**Recall**	**METEOR**	**CIDEr**	**CLAIR**	**METEOR**	**CIDEr**	**CLAIR**
Generalist Models								
InternVL2.5-26B [5]	–	–	0.14	0.33	0.47	0.09	0.07	0.31
Gemini-2.0 Flash [53]	–	–	0.11	0.24	0.53	0.11	0.16	0.52
Gemini-1.5 Pro [53]	–	–	0.15	0.48	**0.62**	0.13	0.27	**0.56**
Specialist Models								
Image-level models								
GLaMM [48] + SAM2 [49]	0.28	0.04	0.04	0.03	0.16	0.04	0.02	0.33
Shikra [3] + SAM2 [49]	0.27	0.26	0.08	0.15	0.32	0.04	0.01	0.29
Ferret-7B [64] + SAM2 [49]	0.64	0.44	0.14	0.21	0.37	0.10	0.12	0.31
Ferret-13B [64] + SAM2 [49]	0.64	0.43	0.14	0.20	0.39	0.11	0.10	0.31
Video-level models								
Sa2VA-1B [66]	0.09	0.07	0.10	0.16	0.31	0.06	0.03	0.26
Sa2VA-4B [66]	0.55	0.25	0.05	0.02	0.19	0.00	0.00	0.07
Sa2VA-8B [66]	0.64	0.17	0.02	0.02	0.20	0.00	0.00	0.13
SAMA-1B	0.67 **(0.58↑)**	0.53 **(0.46↑)**	0.16 **(0.06↑)**	0.56 **(0.40↑)**	0.53 **(0.22↑)**	**0.14** **(0.08↑)**	0.31 **(0.28↑)**	0.45 **(0.19↑)**
SAMA-4B	0.69 **(0.14↑)**	**0.55** **(0.30↑)**	**0.17** **(0.12↑)**	0.65 **(0.63↑)**	0.57 **(0.38↑)**	0.13 **(0.13↑)**	0.30 **(0.30↑)**	0.48 **(0.41↑)**
SAMA-8B	**0.70** **(0.06↑)**	**0.55** **(0.38↑)**	**0.17** **(0.15↑)**	**0.69** **(0.67↑)**	0.58 **(0.38↑)**	0.13 **(0.13↑)**	**0.32** **(0.32↑)**	0.50 **(0.37↑)**

Experiments

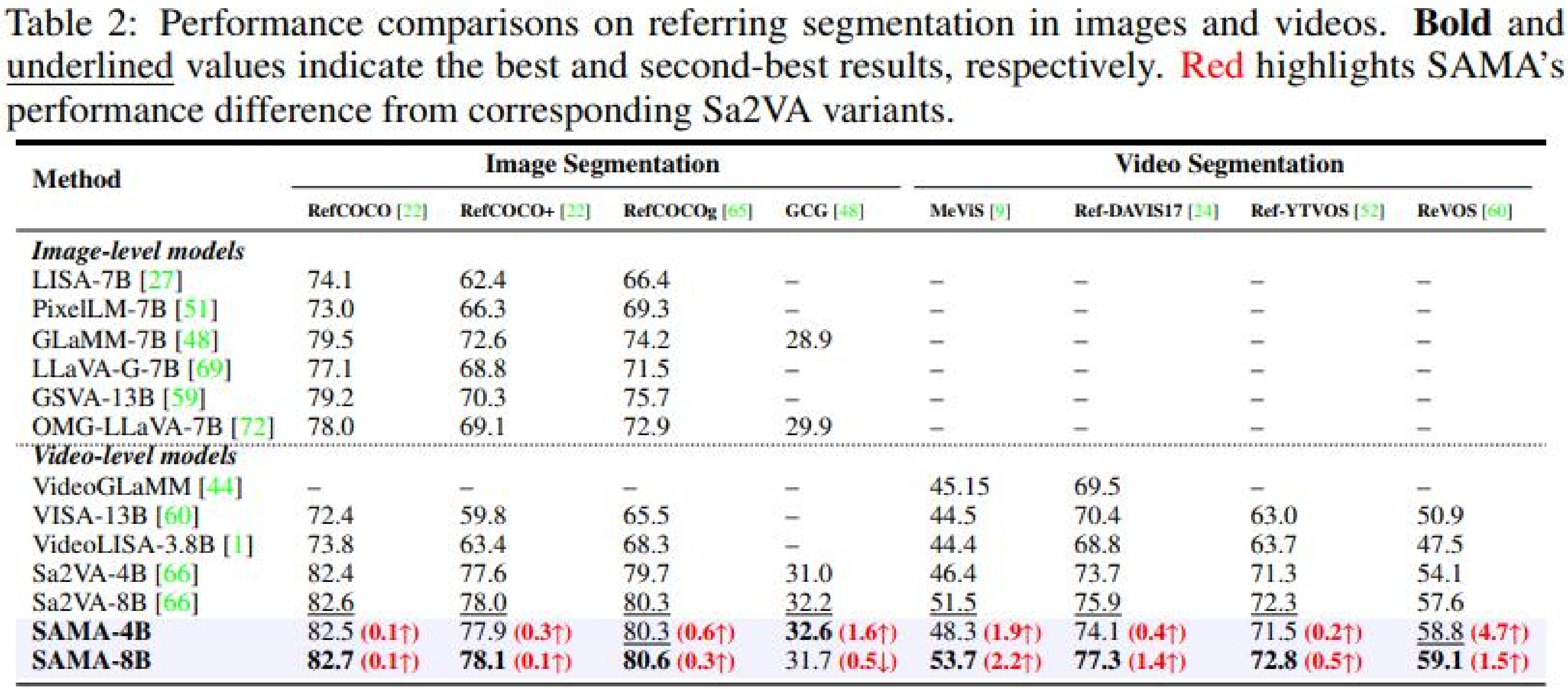
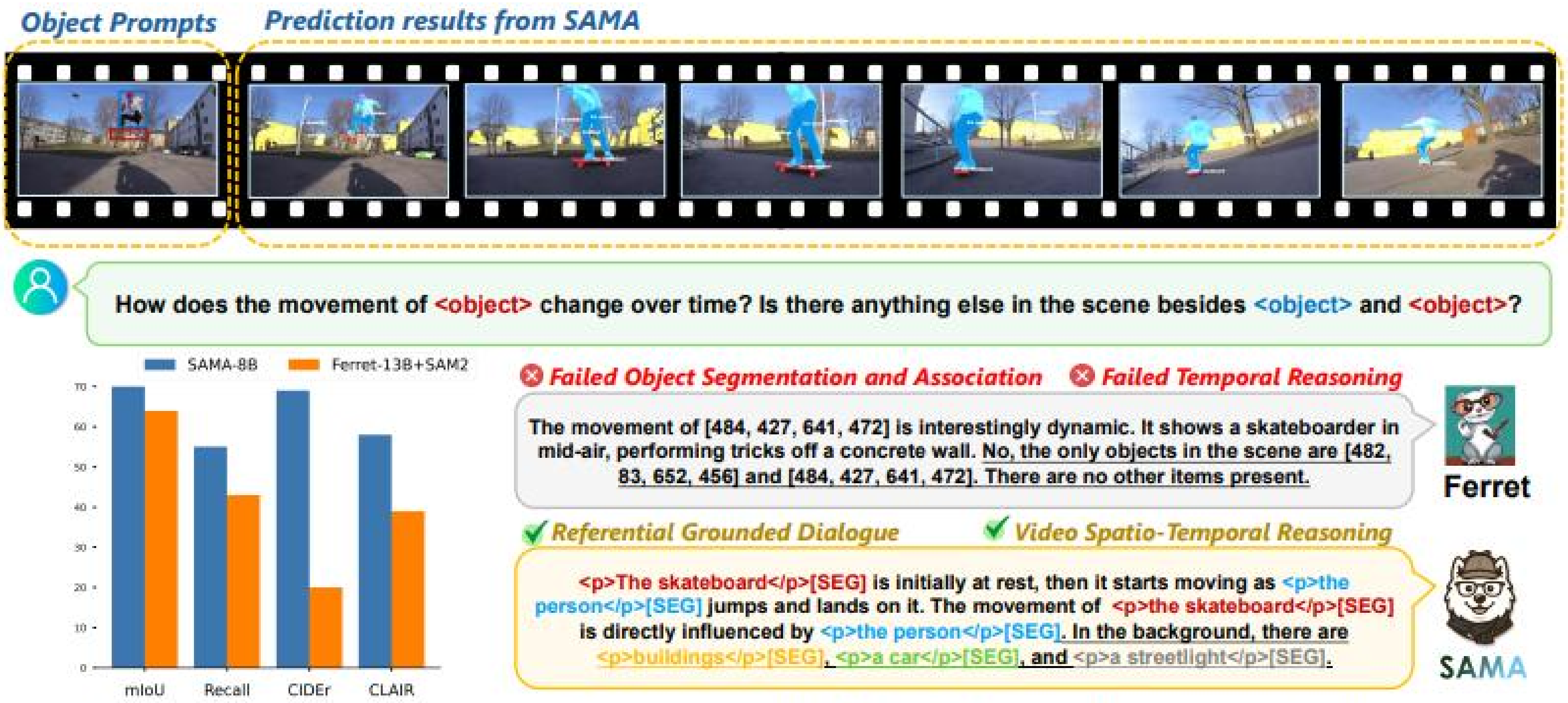


Table 2: Performance comparisons on referring segmentation in images and videos. **Bold** and underlined values indicate the best and second-best results, respectively. Red highlights SAMA's performance difference from corresponding Sa2VA variants.

Method	Image Segmentation				Video Segmentation			
	RefCOCO [22]	**RefCOCO+** [22]	**RefCOCOg** [65]	**GCG** [48]	**MeViS** [9]	**Ref-DAVIS17** [24]	**Ref-YTVOS** [52]	**ReVOS** [60]
Image-level models								
LISA-7B [27]	74.1	62.4	66.4	–	–	–	–	–
PixelLM-7B [51]	73.0	66.3	69.3	–	–	–	–	–
GLaMM-7B [48]	79.5	72.6	74.2	28.9	–	–	–	–
LLaVA-G-7B [69]	77.1	68.8	71.5	–	–	–	–	–
GSVA-13B [59]	79.2	70.3	75.7	–	–	–	–	–
OMG-LLaVA-7B [72]	78.0	69.1	72.9	29.9	–	–	–	–
Video-level models								
VideoGLaMM [44]	–	–	–	–	45.15	69.5	–	–
VISA-13B [60]	72.4	59.8	65.5	–	44.5	70.4	63.0	50.9
VideoLISA-3.8B [1]	73.8	63.4	68.3	–	44.4	68.8	63.7	47.5
Sa2VA-4B [66]	82.4	77.6	79.7	31.0	46.4	73.7	71.3	54.1
Sa2VA-8B [66]	82.6	78.0	80.3	32.2	51.5	75.9	72.3	57.6
SAMA-4B	82.5 **(0.1↑)**	77.9 **(0.3↑)**	80.3 **(0.6↑)**	**32.6** **(1.6↑)**	48.3 **(1.9↑)**	74.1 **(0.4↑)**	71.5 **(0.2↑)**	58.8 **(4.7↑)**
SAMA-8B	**82.7** **(0.1↑)**	**78.1** **(0.1↑)**	**80.6** **(0.3↑)**	31.7 **(0.5↓)**	**53.7** **(2.2↑)**	**77.3** **(1.4↑)**	**72.8** **(0.5↑)**	**59.1** **(1.5↑)**

Visualization


Object Prompts
Prediction results from SAMA
SAMA-8B
Ferret-13B+SAM2
70
60
50
40
30
20
10
0
mIoU
Recall
CIDEr
CLAIR
Failed Object Segmentation and Association
Failed Temporal Reasoning
Ferret
Referential Grounded Dialogue
Video Spatio-Temporal Reasoning
SAMA

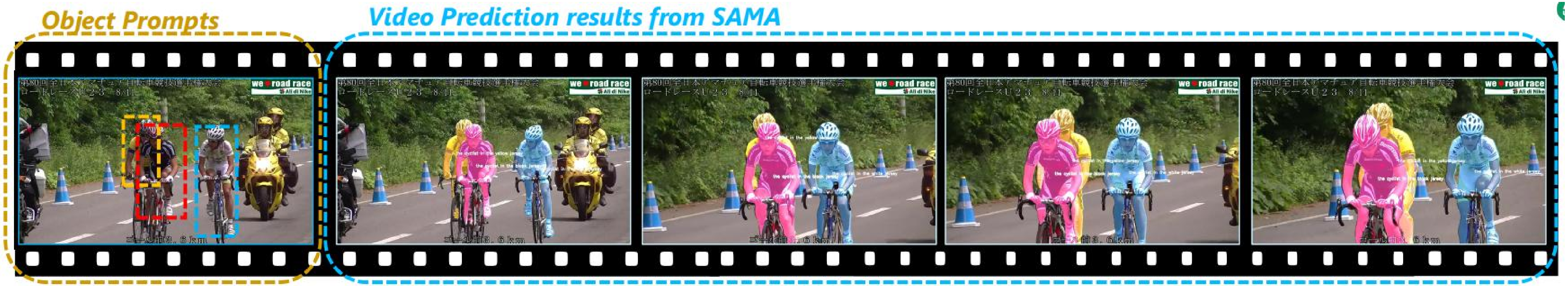
How does the movement of <object> change over time? Is there anything else in the scene besides <object> and <object>?

The movement of [484, 427, 641, 472] is interestingly dynamic. It shows a skateboarder in mid-air, performing tricks off a concrete wall. No, the only objects in the scene are [482, 83, 652, 456] and [484, 427, 641, 472]. There are no other items present.

<p>The skateboard</p>[SEG] is initially at rest, then it starts moving as <p>the person</p>[SEG] jumps and lands on it. The movement of <p>the skateboard</p>[SEG] is directly influenced by <p>the person</p>[SEG]. In the background, there are <p>buildings</p>[SEG], <p>a car</p>[SEG], and <p>a streetlight</p>[SEG].

Figure 5: Visual comparisons between our SAMA and Ferret. Best viewed with zoom.

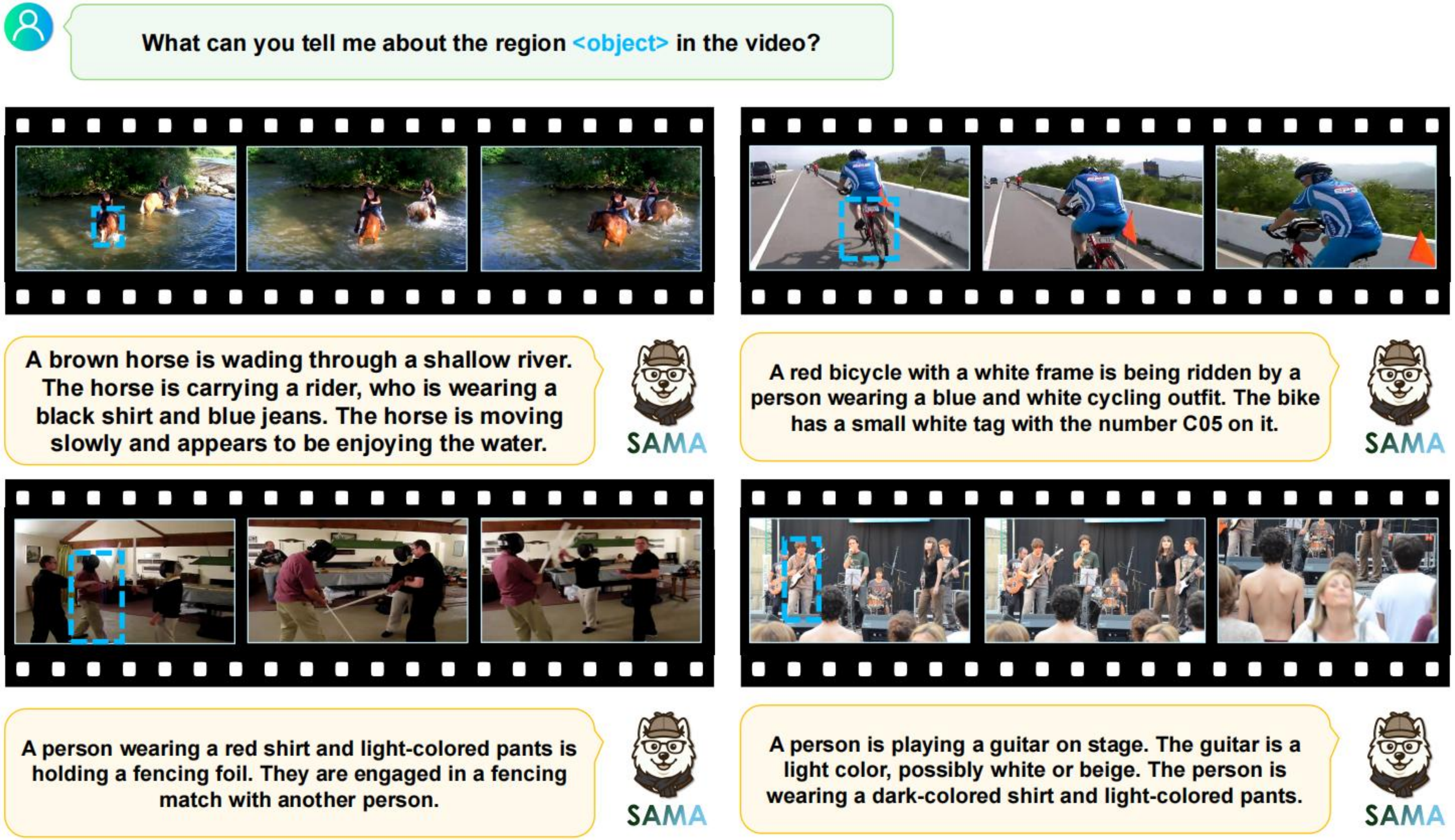
Visualization



What is the spatial relationship between the cyclists <object>, <object>, and <object>?
Response with interleaved segmentation masks.

<p> The cyclist in the white jersey </p> [SEG] is riding behind <p> the cyclist in the black jersey </p> [SEG] . <p> The cyclist in the yellow jersey </p> [SEG] is also riding behind <p> the cyclist in the black jersey </p> [SEG] .

Visualization



What can you tell me about the region <object> in the video?

A brown horse is wading through a shallow river. The horse is carrying a rider, who is wearing a black shirt and blue jeans. The horse is moving slowly and appears to be enjoying the water.

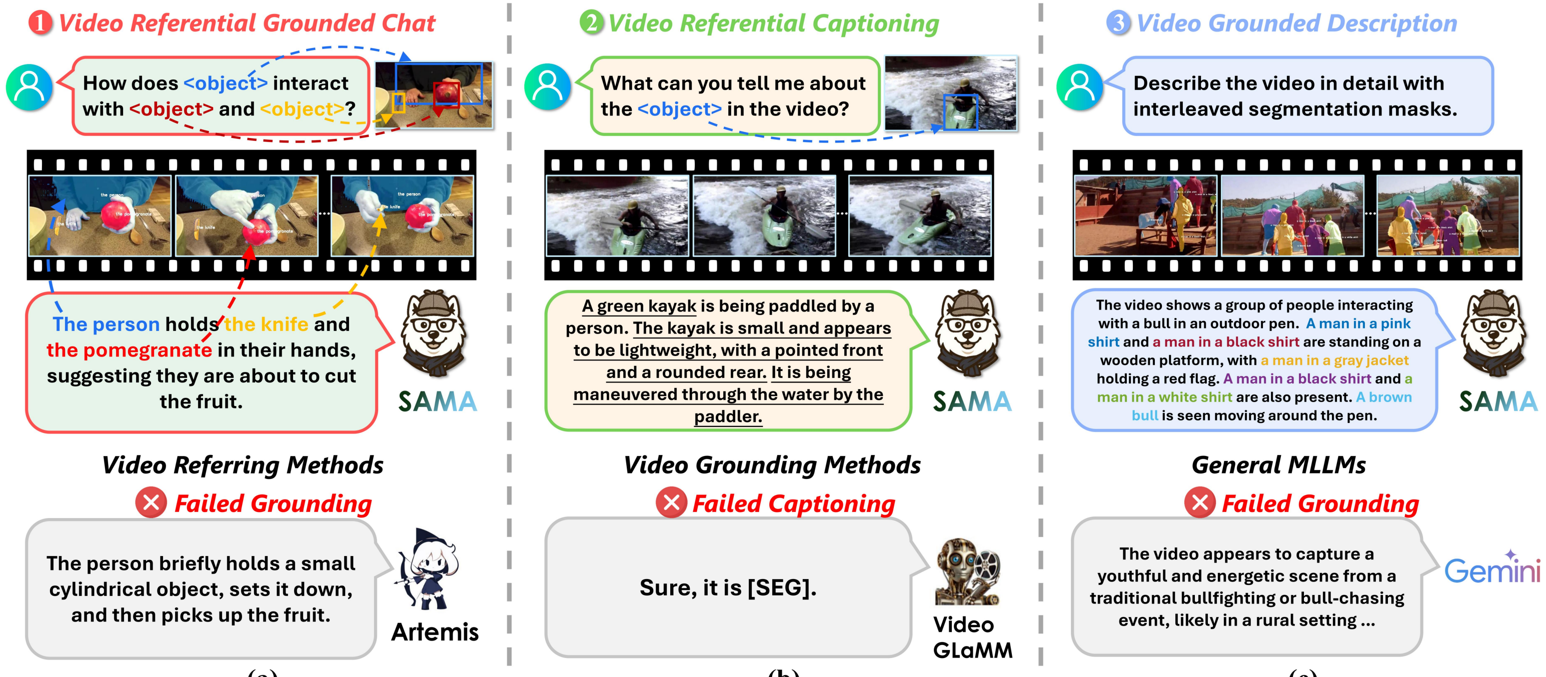
A red bicycle with a white frame is being ridden by a person wearing a blue and white cycling outfit. The bike has a small white tag with the number C05 on it.

A person wearing a red shirt and light-colored pants is holding a fencing foil. They are engaged in a fencing match with another person.

A person is playing a guitar on stage. The guitar is a light color, possibly white or beige. The person is wearing a dark-colored shirt and light-colored pants.

Thank you!

NEURAL INFORMATION
PROCESSING SYSTEMS



How does <object> interact with <object> and <object>?

The person holds the knife and the pomegranate in their hands, suggesting they are about to cut the fruit.

The person briefly holds a small cylindrical object, sets it down, and then picks up the fruit.

What can you tell me about the <object> in the video?

A green kayak is being paddled by a person. The kayak is small and appears to be lightweight, with a pointed front and a rounded rear. It is being maneuvered through the water by the paddler.

Sure, it is [SEG].

Describe the video in detail with interleaved segmentation masks.

The video shows a group of people interacting with a bull in an outdoor pen. A man in a pink shirt and a man in a black shirt are standing on a wooden platform, with a man in a gray jacket holding a red flag. A man in a black shirt and a man in a white shirt are also present. A brown bull is seen moving around the pen.

The video appears to capture a youthful and energetic scene from a traditional bullfighting or bull-chasing event, likely in a rural setting ...

(a) (b) (c)

See the paper for more details!

Fudan University · HKUST · Tongji University