VISA: Reasoning Video Object Segmentation via Large Language Models

Cilin Yan[1], Haochen Wang[2], Shilin Yan[3], Xiaolong Jiang[3], Yao Hu[3],
Guoliang Kang[1], Weidi Xie[4,5], Efstratios Gavves[2]

[1]Beihang University [2]University of Amsterdam [3]Xiaohongshu Inc.
[4]Shanghai Jiao Tong University [5]Shanghai AI Laboratory

https://github.com/cilinyan/VISA

Contribution

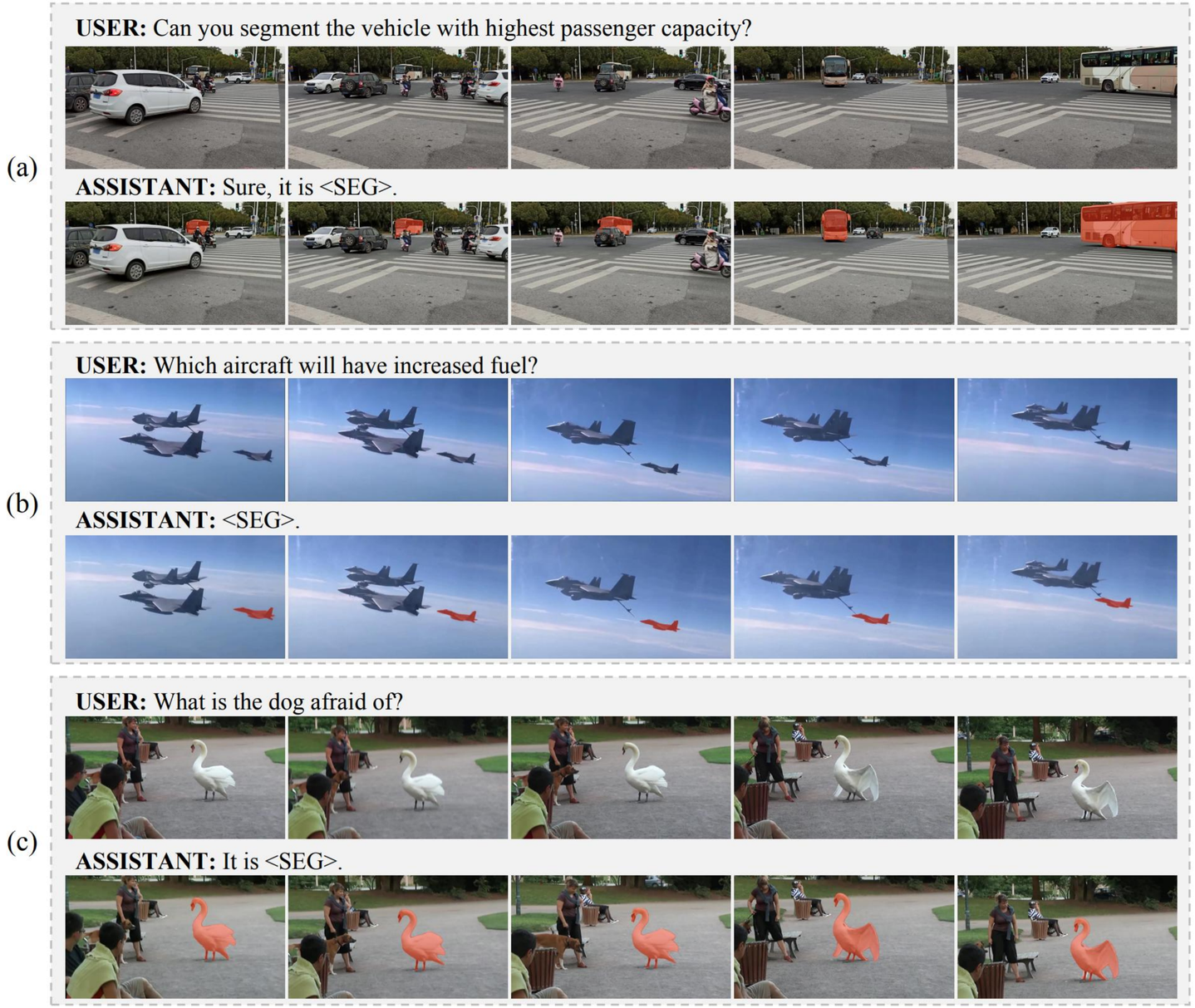
- Generalise to Reasoning Video Object Segmentation (ReasonVOS)
- Construct a comprehensive benchmark, termed ReVOS
- Establish a video-based large language Instructed Segmentation Assistant (VISA) for ReasonVOS

Reasoning Video Object Segmentation

a) complex reasoning of world knowledge

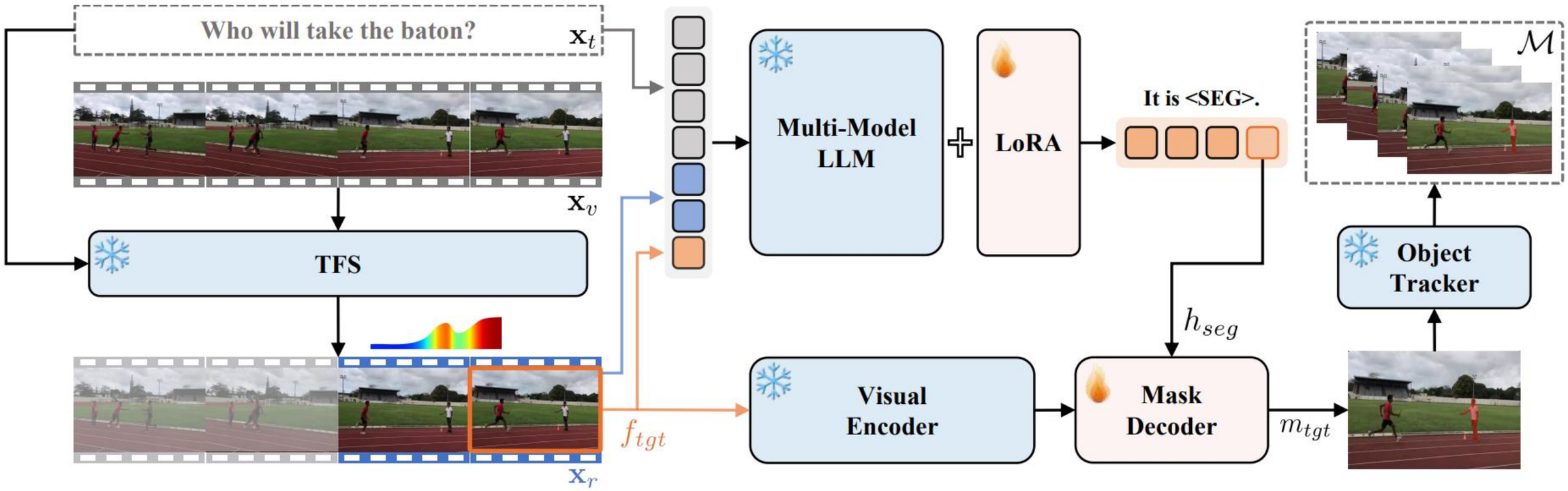
b) inference of upcoming events

c) comprehensive understanding of video content

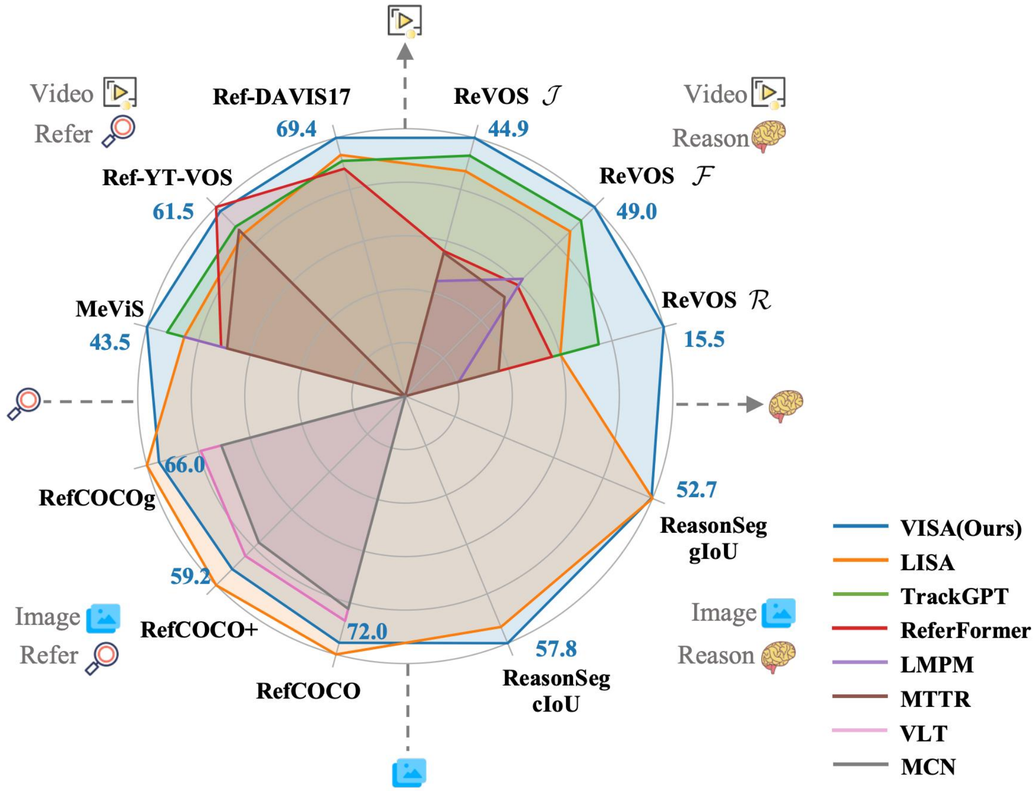


VISA Architecture

- Text-Guided Frame Sampling for Relevant Context Selection
- Multimodal Large Language Model Generates Segmentation Cues
- Bi-Directional Mask Propagation by Object Tracker for Video Segmentation



Experiment Results



Method	Backbone	referring			reasoning			overall			$\mathcal{R}$
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	
ReferFormer [47]	Resnet50	16.6	17.1	16.9	11.9	13.8	12.8	14.3	15.4	14.9	4.9
MTTR [2]	Video-Swin-T	29.8	30.2	30.0	20.4	21.5	21.0	25.1	25.9	25.5	5.6
LMPM [8]	Swin-T	29.0	39.1	34.1	13.3	24.3	18.8	21.2	31.7	26.4	3.2
ReferFormer [47]	Video-Swin-B	31.2	34.3	32.7	21.3	25.6	23.4	26.2	29.9	28.1	8.8
LLaMA-VID [21]+LMPM	Swin-T	29.0	39.1	34.1	12.8	23.7	18.2	20.9	31.4	26.1	3.4
LISA [17]	LLaVA-7B	44.3	47.1	45.7	33.8	38.4	36.1	39.1	42.7	40.9	9.3
LISA* [17]	LLaVA-13B	45.2	47.9	46.6	34.3	39.1	36.7	39.8	43.5	41.6	8.6
TrackGPT(IT)* [38]	LLaVA-7B	46.7	49.7	48.2	36.8	41.2	39.0	41.8	45.5	43.6	11.6
TrackGPT(IT)* [38]	LLaVA-13B	48.3	50.6	49.5	38.1	42.9	40.5	43.2	46.8	45.0	12.8
VISA	Chat-UniVi-7B	51.1	54.7	52.9	36.7	41.7	39.2	43.9	48.2	46.1	7.9
VISA	Chat-UniVi-13B	52.3	55.8	54.1	38.3	43.5	40.9	45.3	49.7	47.5	8.3
VISA(IT)	LLaVA-7B	49.4	52.6	51.0	40.5	45.8	43.2	44.9	49.2	47.1	15.3
VISA(IT)	LLaVA-13B	**55.7**	59.0	**57.4**	41.9	46.5	44.2	**48.8**	52.8	50.8	15.1
VISA(IT)	Chat-UniVi-7B	49.2	52.6	50.9	40.6	45.4	43.0	44.9	49.0	46.9	**15.5**
VISA(IT)	Chat-UniVi-13B	55.6	**59.1**	**57.4**	**42.0**	**46.7**	**44.3**	**48.8**	**52.9**	**50.9**	14.5

Methods	Backbone	MeViS			Ref-YT-VOS			Ref-DAVIS17		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
URVOS [37]	ResNet50	25.7	29.9	27.8	45.3	49.2	47.2	47.3	56.0	51.6
LBDT [11]	ResNet50	27.8	30.8	29.3	48.2	50.6	49.4	-	-	54.1
MTTR [2]	Video-Swin-T	28.8	31.2	30.0	54.0	56.6	55.3	-	-	-
ReferFormer [47]	Video-Swin-B	29.8	32.2	31.0	61.3	64.6	62.9	58.1	64.1	61.1
LMPM [8]	Swin-T	34.2	40.2	37.2	-	-	-	-	-	-
OnlineRefer [46]	Swin-L	-	-	-	**61.6**	**65.5**	**63.5**	61.6	67.7	64.8
LISA [17]	LLaVA-7B	35.1	39.4	37.2	53.4	54.3	53.9	62.2	67.3	64.8
LISA [17]	LLaVA-13B	35.8	40.0	37.9	54.0	54.8	54.4	63.2	68.8	66.0
TrackGPT [38]	LLaVA-7B	37.6	42.6	40.1	55.3	57.4	56.4	59.4	67.0	63.2
TrackGPT [38]	LLaVA-13B	39.2	43.1	41.2	58.1	60.8	59.5	62.7	70.4	66.5
VISA (Ours)	Chat-UniVi-7B	40.7	46.3	43.5	59.8	63.2	61.5	66.3	72.5	69.4
VISA (Ours)	Chat-UniVi-13B	**41.8**	**47.1**	**44.5**	61.4	64.7	63.0	**67.0**	**73.8**	**70.4**

Experiment Results

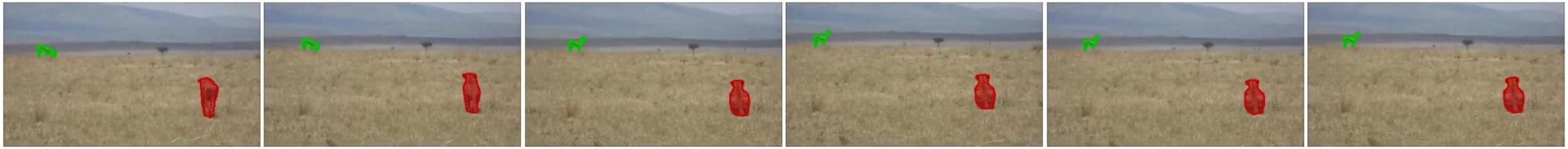
the walrus that loses the most gravitational potential energy.

object that brushes off raindrops and dusty. What object in the video suggests there is a fire nearby?

labrador. Which dog's neck experiences more external force?

Which object has the capacity to fly? What the person might be resting on after the presentation?

Experiment Results



object ready to start hunting. Which animal in the video is most likely to die?

objects that simulate small animals.

the ball that should be hit first according to the rules. Which ball is the target of this shot?

What are the rats being trained to pick up in the video?