ByteDance | Seed

Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos

Haobo Yuan[1*], Xiangtai Li[2* †], Tao Zhang[2,3*], Zilong Huang[2], Shilin Xu[4], Shunping Ji[3], Yunhai Tong[4], Lu Qi[2], Jiashi Feng[2], Ming-Hsuan Yang[1]

[1]UC Merced [2]Bytedance Seed [3]Wuhan University [4]Peking University

[†]: Project lead. *: equal technical contribution to the work. E-mail: xiangtai94@gmail.com

Project Page: https://lxtgh.github.io/project/sa2va

Huggingface Model: https://huggingface.co/ByteDance/Sa2VA-4B

汇报人：赖彦涛

任务

视频


Can you describe the video?
The video shows a woman walking down a busy city street at night. She is wearing a red dress and carrying a black handbag. The street is illuminated by bright lights, and there are many people walking around. The woman is walking down the street, passing by many people, and looking around. She is wearing sunglasses and has a confident and stylish demeanor. The video captures the vibrant and lively atmosphere of the city at night, with the bright lights and bustling crowds...
What is the person holding?
The person is holding a black handbag.
Can you segment the person wearing sunglasses on the street?
Sure, it is [SEG].

(a)

图像

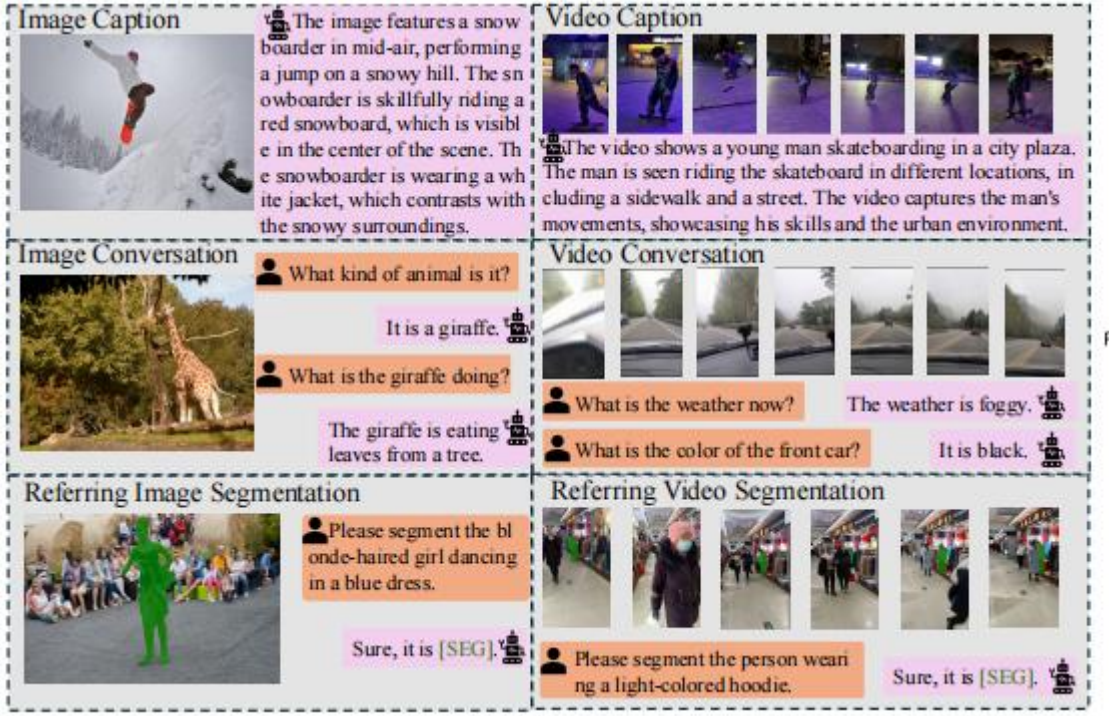


Image Caption
The image features a snowboarder in mid-air, performing a jump on a snowy hill. The snowboarder is skillfully riding a red snowboard, which is visible in the center of the scene. The snowboarder is wearing a white jacket, which contrasts with the snowy surroundings.
Video Caption
The video shows a young man skateboarding in a city plaza. The man is seen riding the skateboard in different locations, including a sidewalk and a street. The video captures the man's movements, showcasing his skills and the urban environment.
Image Conversation
What kind of animal is it?
It is a giraffe.
What is the giraffe doing?
The giraffe is eating leaves from a tree.
Video Conversation
What is the weather now?
The weather is foggy.
What is the color of the front car?
It is black.
Referring Image Segmentation
Please segment the blonde-haired girl dancing in a blue dress.
Sure, it is [SEG].
Referring Video Segmentation
Please segment the person wearing a light-colored hoodie.
Sure, it is [SEG].

研究背景与动机

•**多模态LLM的进展与局限：**多模态大型语言模型（MLLM）在处理图像和视频方面取得显著进展，但现有模型往往局限于特定模态和任务，无法全面支持图像和视频的多种任务。

•**对细粒度视频理解的需求：**在实际应用中，需要对视频内容进行细粒度理解，包括分割、跟踪和基于视觉提示的问答等，但现有模型在这些方面存在不足。

Table 1. **Comparison of capabilities of different representative models.** Our method supports various tasks and modalities. Benefiting from these interactive features on video, Sa2VA can perform multiple promptable tasks in the video, as shown in Fig. 1 (a) and (b).

Method	Support Inputs			Dense Grounding				Conversation				End to End
	Image	Video	Visual Prompts	RES	Ref-VOS	Inter-VOS	GCG	Image-Chat	Video-Chat	Video Caption	Interactive Caption	Training
LLAVA [53]	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓
LLaVA-OneVision [42]	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓
InternVL 2.0 [9]	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓
Osprey [92]	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
LISA [40]	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓
GLAMM [66]	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓
VIP-LLaVA [5]	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
VISA [85]	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
OMG-LLaVA [99]	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✓
PSALM [102]	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓
GSVA [83]	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓
LLaMA-VID [50]	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓
ST-LLM [56]	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓
F-LLM [82]	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
Sa2VA (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

方法概述

•**Sa2VA模型架构：**Sa2VA结合了SAM-2（视频分割基础模型）和LLaVA（先进的视觉语言模型），通过统一文本、图像和视频到共享的LLM token空间，实现对静态和动态视觉内容的密集理解。

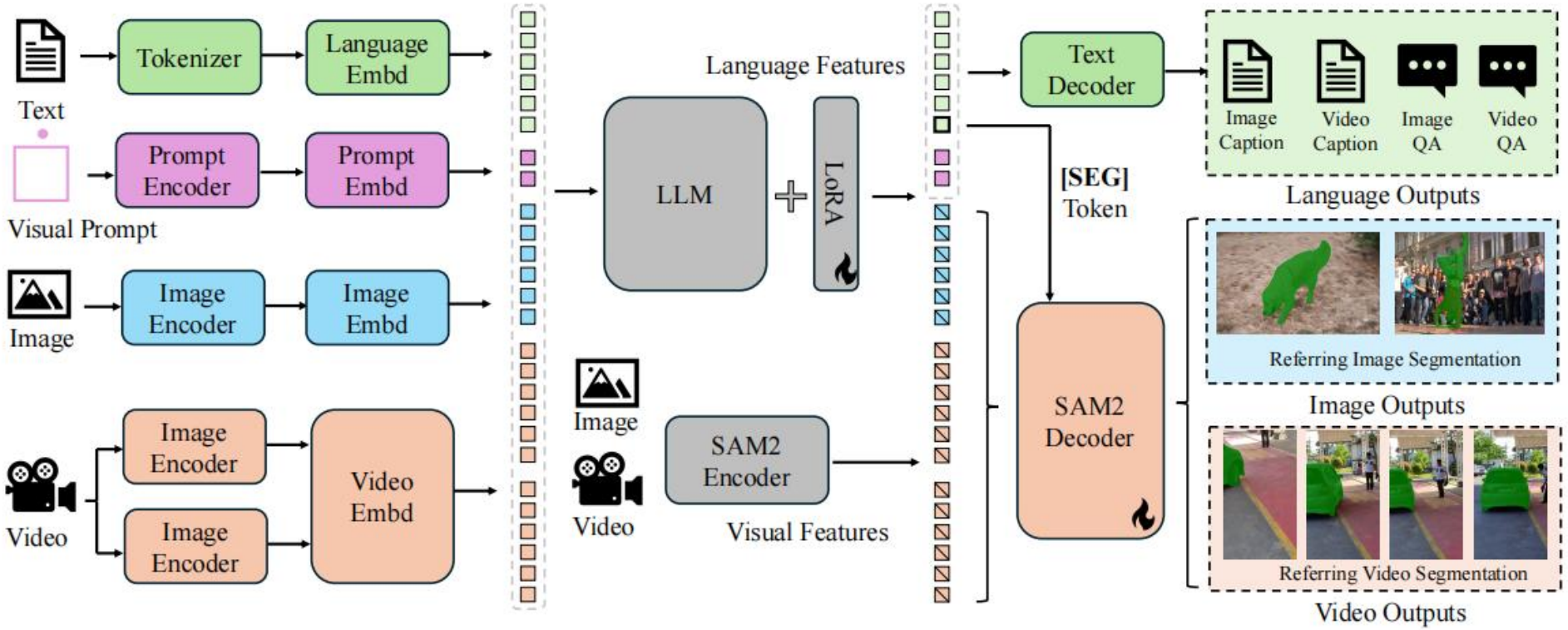


Figure 2. **Our proposed Sa2VA model.** The model first encodes the input texts, visual prompts, images, and videos into token embeddings. These tokens are then processed through a large language model (LLM). The output text tokens are used to generate the "[SEG]" token and associated language outputs. The SAM-2 decoder receives the image and video features from the SAM-2 encoder, along with the "[SEG]" token, to generate corresponding image and video masks.

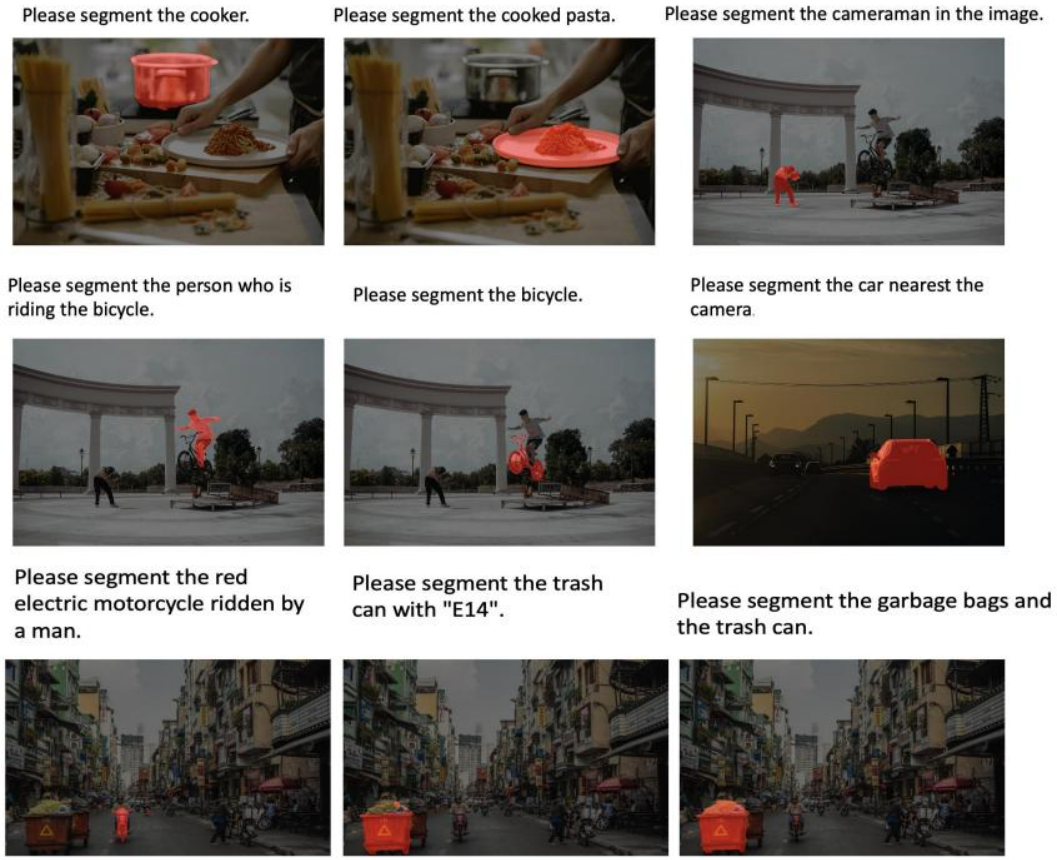
主要实验结果：

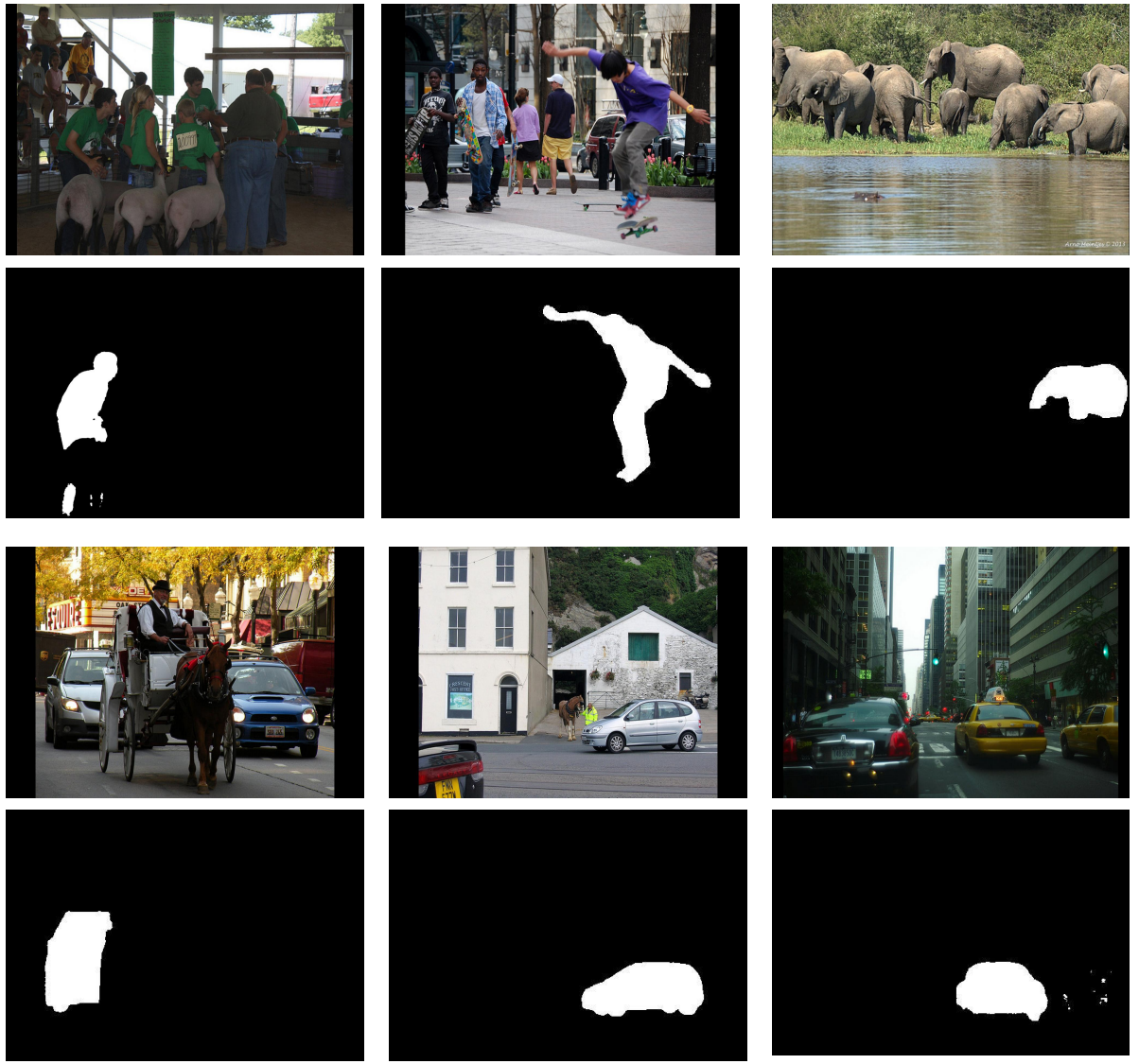
图像分割基准、视频分割基准、聊天任务

Table 2. Experiment results on image/video referring segmentation benchmarks and image/video chat benchmarks. For MME dataset, A / B means the perception (A) and cognition (B) scores, while scores without "/" mean the total score (A + B)

Method	Image Segmentation			Video Segmentation				Image Chat			Video Chat		GCG
	RefCOCO [36]	RefCOCO+ [36]	RefCOCOg [90]	MeViS [13]	Ref-DAVIS17 [38]	Ref-YTVOS [69]	ReVOS [85]	MME [19]	MMBench [59]	SEED-Bench [41]	Video-MME [20]	MMBench-Video [17]	GCG [66]
LLAVA-1.5-13B [54]	-	-	-	-	-	-	-	1531	68.8	70.1	-	-	-
Video-LLaVA-7B [51]	-	-	-	-	-	-	-	-	60.9	-	39.9	1.03	-
LLaMA-VID-7B [50]	-	-	-	-	-	-	-	1521	65.1	59.9	-	1.08	-
mPLUG-Owl3-8B [89]	-	-	-	-	-	-	-	-	77.6	-	53.5	1.35	-
InternVL2-8B [9]	-	-	-	-	-	-	-	-	81.7	76.2	54.0	1.28	-
PixelLM-7B [68]	73.0	66.3	69.3	-	-	-	-	309/135	17.4	-	-	-	-
LaSagnA [77]	76.8	66.4	70.6	-	-	-	-	0/0	0.0	-	-	-	-
LISA-7B [40]	74.1	62.4	66.4	-	-	-	-	1/1	0.4	-	-	-	-
GLaMM-7B [66]	79.5	72.6	74.2	-	-	-	-	14/9	36.8	-	-	-	28.9
LLaVA-G-7B [96]	77.1	68.8	71.5	-	-	-	-	-	-	-	-	-	-
GSVA-13B [83]	79.2	70.3	75.7	-	-	-	-	-	-	-	-	-	-
OMG-LLaVA-7B [99]	78.0	69.1	72.9	-	-	-	-	1177/235	47.9	56.5	-	-	29.9
VISA-13B [85]	72.4	59.8	65.5	44.5	70.4	63.0	50.9	-	-	-	-	-	-
Sa2VA-1B (Ours)	77.4	69.9	72.3	41.7	72.3	65.3	47.6	1381/405	68.3	64.8	39.9	1.07	23.8
Sa2VA-4B (Ours)	78.9	71.7	74.1	46.2	73.8	70.0	53.2	1536/530	77.3	73.3	50.4	1.23	28.2
Sa2VA-8B (Ours)	81.6	76.2	78.7	46.9	75.2	70.7	57.6	1617/511	81.6	75.1	52.1	1.34	31.0
Sa2VA-26B (Ours)	82.5	78.8	79.7	46.2	77.0	70.1	58.4	1691/538	83.7	76.8	52.6	1.45	33.5

可视化结果：


Please segment the cooker.
Please segment the cooked pasta.
Please segment the cameraman in the image.
Please segment the person who is riding the bicycle.
Please segment the bicycle.
Please segment the car nearest the camera.
Please segment the red electric motorcycle ridden by a man.
Please segment the trash can with "E14".
Please segment the garbage bags and the trash can.



消融实验

Effectiveness of Joint Co-training.

Sa2VA's superior performance on image and video QA and segmentation bench marks is attributed to unified instruction tuning and joint co-training

Table 7. Ablation study on co-training effect on multiple datasets.

Data	Image Segmentation			Video Segmentation		Image Chat			Video Chat	
	RefCOCO	RefCOCO+	RefCOCOg	MeViS	Ref-DAVIS17	MME	MMBench	SEED-Bench	Video-MME	MMBench-Video
All Data	77.4	69.9	72.3	50.8	72.3	1381/405	68.3	64.8	39.9	1.07
w/o Image QA	78.0	70.1	72.2	48.3	73.0	1298/359	63.4	63.8	39.7	0.39
w/o Image Segmentation	20.2	20.6	23.2	38.0	48.8	1393/408	70.1	65.7	41.2	1.08
w/o Video QA	78.0	70.4	72.6	50.7	74.3	1370/402	69.1	65.0	41.3	0.71
w/o Video Segmentation	77.4	69.1	72.4	44.4	69.0	1403/398	67.8	64.9	40.4	1.04

Ablation study on the Segmentation Token Design

We have explored several segmentation token designs. When using the repeat strategy, which makes the LLM output N "[SEG]" tokens for N frames, this approach slightly decreases video segmentation performance due to being more prone to missing or repeating "[SEG]" tokens. We also explored using different segmentation tokens for different frames, such as setting "[SEG1]" to "[SEG n]" to segment the object across frames 1 to n

Table 9. Ablation study on "[SEG]" token design.

Type	RefCOCO	RefCOCO+	RefCOCOg	DAVIS	MeVIS
Single	77.4	69.9	72.3	72.3	50.8
Repeat	77.3	70.2	72.5	71.1	49.6
Multiple	77.6	70.3	72.4	68.6	46.3

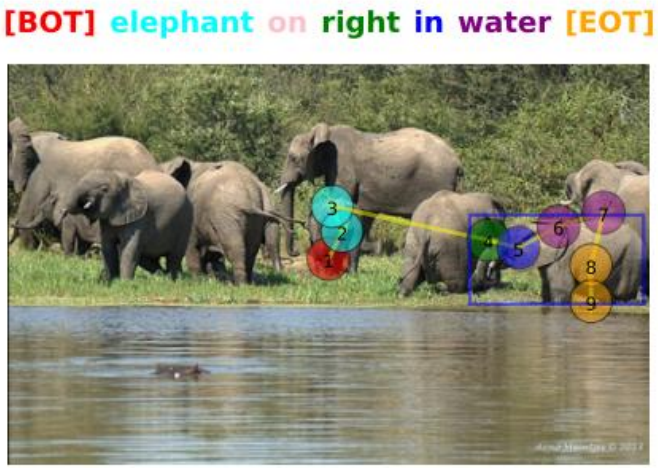
Scale ability of Sa2VA's Data

When we add 3M image QA data from Infinity-MM(only stage-4), Sa2VA-1B shows a 2.1 score improvement on MMBench, with minimal impact on the image and video segmentation benchmark performance. Using our annotated Ref-SAV data, Sa2VA-1B exhibits a 1.7 J&F performance boost on the MeVIS benchmark. These results indicate that Sa2VA's performance can continue to improve as the scale of SFT data increases.

Table 10. Ablation study on using more datasets.

Dataset	Size	RefCOCO	RefCOCOg	MMBench	MME	MeVIS
baseline	1.2M	77.4	72.3	68.3	1381/405	50.8
Inifinity-MM [22]	1.2M+3M	77.1(-0.3)	72.6(+0.3)	70.4(+2.1)	1396/346(-44)	51.2(+0.4)
Ref-SAV	1.2M+37K	77.2(-0.2)	72.6(+0.3)	68.2(-0.1)	1384/418(+16)	52.5(+1.7)

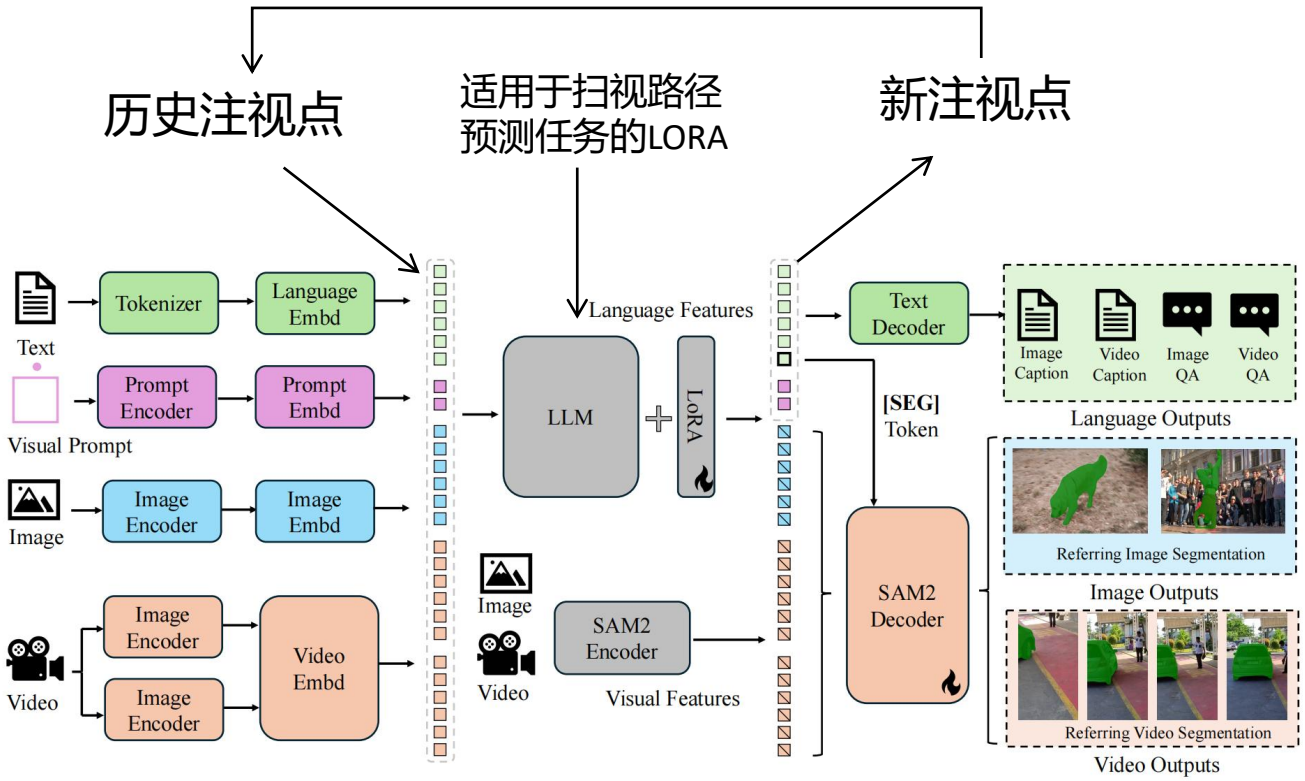
任务：指导表达引导的人类扫视轨迹预测

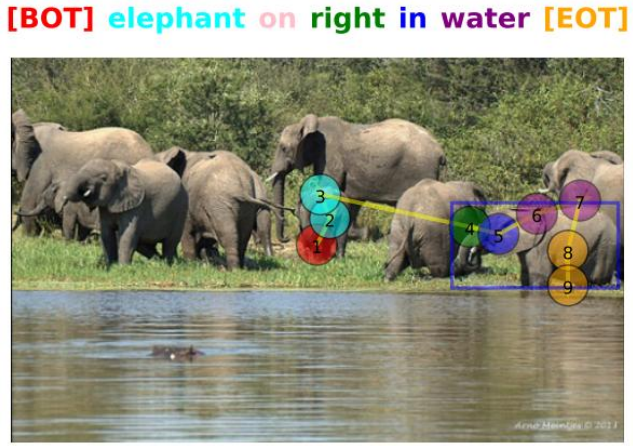


特点：

1. 每一个单词都要预测所对应的注视点Pack(Pack指变化数量的注视点集合)
2. 最终单词所预测的注视点要落在指定表达的指定目标上
3. 在预测当前单词所对应的注视点Pack时不能看到后续的文本信息

方案：


历史注视点
适用于扫视路径
预测任务的LORA
新注视点
Text
Tokenizer
Language Embd
Visual Prompt
Prompt Encoder
Prompt Embd
Image
Image Encoder
Image Embd
Video
Image Encoder
Image Encoder
Video Embd
Language Features
LLM
LoRA
Image
Video
SAM2 Encoder
Visual Features
Text Decoder
[SEG] Token
SAM2 Decoder
Image Caption
Video Caption
Image QA
Video QA
Language Outputs
Referring Image Segmentation
Image Outputs
Referring Video Segmentation
Video Outputs


[BOT] elephant on right in water [EOT]