Object-centric Video Question Answering with Visual Grounding and Referring

Supplementary Material

1. More Implementation Details

Training Data Composition. The complete list of used datasets in training is presented in Tab. 1. For the datasets with overlapping, we select the disjoint samples during training, such as ViP-LLaVA-Instruct and LLaVA-150k.

Task	Datasets	# Samples
Image Segmentation		
Semantic Seg.	ADE20k [28]	20.2k
	COCO-Stuff [3]	118.3k
	Pascal-Part [5]	4.3k
	PACO-LVIS [18]	4.6k
Referring Seg.	RefCOCO+ [8]	17k
	RefCOCO [8]	22k
	RefCOCOg [15]	17k
	RefCLEF [8]	18k
Reasoning Seg.	ReasonSeg [10]	0.2k
Video Segmentation		
VOS	YoutubeVOS [19]	3.5k
Referring VOS	Ref-Youtube-VOS [19]	3.5k
	MeViS [7]	1.6k
	Ref-DAVIS [9]	5.3k
Reasoning VOS	ReVOS [21]	0.6k
Image QA		
VQA	LLaVA-150k [13]	150k
Referring VQA	Osprey-Conv [24]	30k
	Osprey-Desc [24]	60k
	ViP-LLaVA-Instruct [4]	216k
Video QA		
VideoQA	LLaVA-Video-OE [27]	960k
	LLaVA-Video-MC [27]	196k
	NeXT-QA-OE [20]	17k
	NeXT-QA-MC [20]	17k
	ActivityNetQA [23]	24k
	PerceptionTest [16]	2.4k
Referring VideoQA	VideoInfer (Ours)	20k

Table 1. The detailed list of training datasets. For some datasets, we only use a subset of them, such as LLaVA-Video, Osprey, and ViP-LLaVA. The sampling rate is listed in the training script.

Training Details. We utilize the dynamic resolution for Qwen2.5-VL [1] models, the max pixels of videos are set as 320×28×28 for 8 frames, and the max pixels of a single image are set as 1280×28×28. Images exceeding the above max pixels will be resized while maintaining their aspect ratio to fit.

Evaluation Protocols. The results of MeViS *val* and Ref-YouTube-VOS are evaluated through the online server. On VideoRefer-Bench^{Q}, our method processes 16 input frames. For VideoInfer, we utilize GPT-4o-2024-11-20 to evaluate Accuracy (Acc.) and Score, following the same prompt strategy as Video-ChatGPT [14]. To evaluate region-feature based models, we convert the visual prompt (RGBA image) into the mask according to the alpha channel and then input the mask with visual input and question into these models to generate the response.

2. More Experimental Results

2.1. Quantitative Results

Method	Perception Test [16]	MVBench [12]	NExT-QA [20]
Generalist Models			
LLaVA-OV-7B [11]	-	56.7	79.4
VideoLLaMA2.1-7B [6]	54.9	57.3	75.6
LLaVA-Video-7B [27]	67.9	58.6	83.2
Specialist Models			
Artemis [17]	47.1	34.1	-
VideoRefer-7B [25]	56.3	59.6	-
RGA3-7B (Ours)	68.7	63.8	75.3

Table 2. Comparison on general video question-answering tasks.

Results on General VideoQA Benchmarks. In addition to the referring video question-answering benchmarks, we also evaluate our architecture on general VideoQA QA tasks without visual prompts as inputs, through the LMMs-Eval toolkit [26]. As shown in Tab. 2, our model is comparable to popular general VideoQA models while possessing the ability to perform interactive referring and grounding in object-centric scenarios.

Method	val		test					
	overall		short query		long query		overall	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
LISA-7B [10]	52.9	54.0	40.6	40.6	49.4	51.0	47.3	48.4
LISA-13B [10]	56.2	62.9	44.3	42.0	54.0	54.3	51.7	51.1
VISA-7B [21]	52.7	57.8	-	-	-	-	-	-
VideoLISA-3.8B [2]	61.4	67.1	43.8	42.7	56.9	57.7	53.8	54.4
LISA++-7B [22]	64.2	68.1	49.6	51.1	59.3	61.7	57.0	59.5
RAG3-3B (Ours)	65.4	68.5	58.5	54.2	62.3	65.8	61.4	63.3
RAG3-7B (Ours)	**68.7**	**70.2**	**58.7**	**54.1**	**68.5**	**72.1**	**66.1**	**68.3**

Table 3. Comparision on validation and test set of ReasonSeg [10] for image-level reasoning object segmentation.

Results on Image Segmentation Benchmarks For image segmentation evaluation, we utilize gIoU (average per-image IoUs) and cIoU (cumulative intersection over union) on reasoning-based benchmark ReasonSeg [10] and cIoU for referring-based benchmark refCOCO(+/g) [8, 15]. As shown in Tab. 3, RGA3-7B outperforms the state-of-the-art

Method	refCOCO [8]			refCOCO+ [8]			refCOCOg [15]	
	val	testA	testB	val	testA	testB	val(U)	test(U)
LISA-7B [10]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
VISA-7B [21]	72.4	75.5	68.1	59.8	64.8	53.1	65.5	66.4
VideoLISA-3.8B [2]	73.8	76.6	68.8	63.4	68.8	56.2	68.3	68.8
RGA3-3B (Ours)	78.9	81.1	75.0	71.3	77.1	63.5	74.7	75.0
RGA3-7B (Ours)	**79.7**	**82.6**	**76.0**	**73.5**	**78.6**	**67.0**	**76.2**	**75.9**

Table 4. Comparison on image-level referring object segmentation datasets.

method LISA++-7B [22] on the ReasonSeg benchmark by a large margin. Moreover, on general image semantic segmentation benchmarks, such as refCOCO, RGA3-7B still outperforms recent MLLM-based methods, which indicates the strong general grounding ability of RGA3.

Robustness in Extremely Long Videos. Our work primarily addresses object-centric video tasks, which typically involve short video durations (*e.g.,* VideoRefer-Bench^{Q} comprises a few-second clips sourced from DAVIS or MeVIS). Although our VideoInfer incorporates longer clips (sub-minute duration) from LVOS and TAO, we acknowledge the necessity for robustness in ultra-long video. Due to the lack of appropriate benchmarks, we evaluated RGA3 on the validation set of the LongVideoBench with different duration groups:

Duration	$(8s, 15s]$	$(15s, 60s]$	$(180s, 600s]$	$(900s, 3600s]$
Accuracy	72.5	70.9	57.3	46.3

Table 5. Performance on the validation set of LongVideoBench.

More Ablations. The improvement over the previous state-of-the-art methods on video referring segmentation and question-answering is mainly from the base MLLM, the proposed STOM module, and the dataset composition in training. Additionally, we find that under the current training strategy, the performance of the individual task will decrease compared to training separately in most cases. We think this should be further addressed through multi-stage training or more diverse prompting.

Model	Training Data	MLLM	Size	Modules	VideoRefer	ReasonVOS
					Acc.	$\mathcal{J}\&\mathcal{F}$
VideoRefer	QA	SigLIP-Qwen2	7B	-	71.9	-
VideoLISA	Seg	Phi-3-V	3.8B	SAM	-	45.1
Ours	Seg	QwenVL-2.5	3B	SAM	-	47.9
	Seg		3B	SAM2	-	48.7
	Seg+QA		3B	SAM2	62.3	51.7
	Seg+QA		3B	SAM2 +STOM	66.6	51.7
	Seg+QA		7B	SAM2 +STOM	74.0	53.6

Table 6. Additional ablations on the design choices.

2.2. Failure Case and Future Work

In practice, due to computational limitations, we restrict RGA3's input to 16 frames per video (Other existing object-

Figure 1. Failure case on VideoInfer dataset. The frames with grey masks are not selected as input to RGA3. The green box frames the video content which the man on the left reports something to the man on the right. 'GT' is the ground truth, and 'Pred' is the prediction of RGA3.

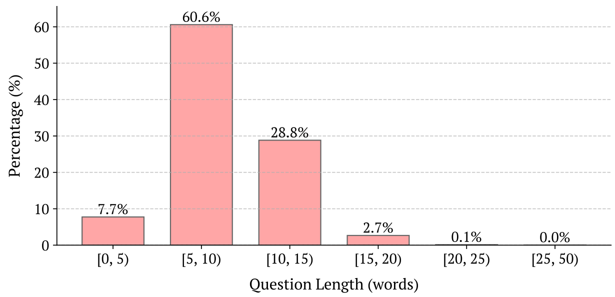
centric VideoLLMs also suffer from this limitation). However, in very long videos, this frame selection introduces large temporal gaps, potentially omitting critical contextual information. Our VideoInfer dataset introduces videos that contain over 1,000 frames, yet the existing models can not process the whole sequence due to computational limitations. For instance, as shown in Fig. 1, the raw video contains over 1,000 frames, yet only 16 frames are used as input. With such sparse frame sampling, the model struggles to capture a coherent sequence of events.

In this specific case, the moment when the left person reports to the right person is skipped, leading to an incorrect prediction. This issue cannot be naively addressed by extracting just one or a few visual tokens per frame, as object-level information must be preserved across frames to enable accurate object-centric reasoning. Therefore, handling long-form object-centric video reasoning remains a challenging open problem, particularly in transforming spatial and temporal detailed object-centric information into a reasonable number of tokens. We plan to explore solutions further to enhance object-centric reasoning in long videos in our future work.

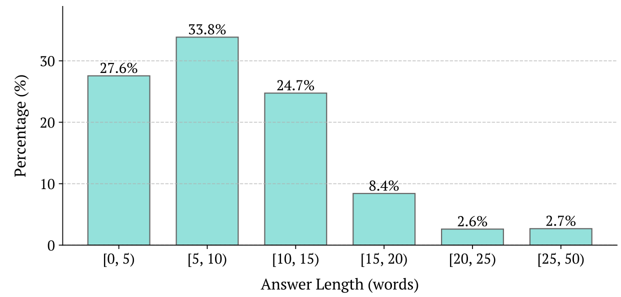
3. Discussion and Visualizations

3.1. Potential Information Loss

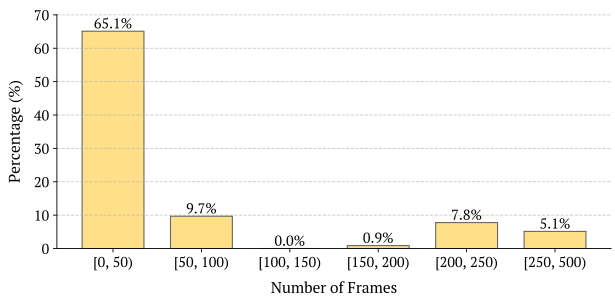
The STOM module blends prompts onto original frames with **transparency** through alpha blending, so that the objects will not be completely occluded, and the features can



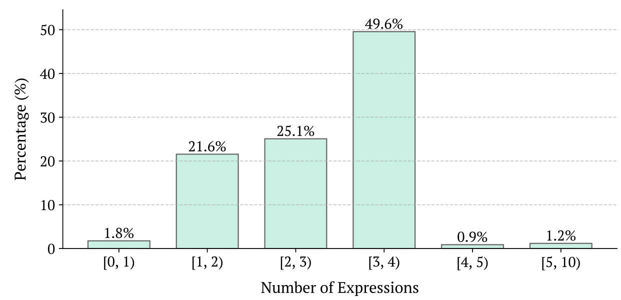
(a) Histogram of question word counts.



(b) Histogram of answer word counts.



(c) Histogram of frames per video counts.



(d) Histogram of objects per video counts.

Figure 2. Visualization of statistics of the test split of VideoInfer.

be reserved.

3.2. Visualization of VideoInfer Benchmark

As shown in Fig. 2, we visualize the statistics of the test split in VideoInfer. The questions range from 3 to 50 words in length, with an average of 8.4 words. Answers vary between 1 and 75 words, averaging 8.7 words. The number of frames per sample spans from 7 to over 2000, with a mean of 189.5 frames. For objects, the count ranges from 1 to 8, averaging 2.3 objects of interest per video.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024. 1, 2

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1209–1218, 2018. 1

[4] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12914–12923, 2024. 1

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1971–1978, 2014. 1

[6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1

[7] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2694–2703, 2023. 1

[8] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Con-*

ference on Empirical Methods in Natural Language Processing (EMNLP), pages 787–798, 2014. 1, 2

[9] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 1

[10] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, 2024. 1, 2

[11] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[12] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, 2024. 1

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[14] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1

[15] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2016. 1, 2

[16] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023. 1

[17] Jihao Qiu, Yuan Zhang, Xi Tang, Lingxi Xie, Tianren Ma, Pengyu Yan, David Doermann, Qixiang Ye, and Yunjie Tian. Artemis: Towards referential understanding in complex videos. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[18] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7141–7151, 2023. 1

[19] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision (ECCV)*, pages 208–223. Springer, 2020. 1

[20] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021. 1

[21] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision (ECCV)*, pages 98–115. Springer, 2024. 1, 2

[22] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 1, 2

[23] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 1

[24] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28202–28211, 2024. 1

[25] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1

[26] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024. 1

[27] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1

[28] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017. 1