The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

Supplementary Material

This supplementary material provides additional details and analysis of VRS-HQ, expanding on the content presented in the main paper. We begin by evaluating the impact of various training datasets on segmentation performance (§A). Next, we present more detailed implementation information to facilitate reproducibility (§B). We then elaborate on the specific method of utilizing SAM2 [8] for mask decoding and propagation (§C). Subsequently, we show some failure cases with analysis to offer a more comprehensive understanding of VRS-HQ's limitations (§D). Then we compare VRS-HQ with other methods on the multimodal question-answering tasks. (§E) Additionally, we present more qualitative comparisons against VISA, highlighting the strengths of our proposed method (§F). Finally, we visualize the reasoning segmentation results of VRS-HQ on in-the-wild video datasets, demonstrating its strong generalization capabilities (§G).

A. Datasets Ablation

As illustrated in Tab. 1, fine-tuning with the full datasets yields the best performance while excluding the image segmentation dataset, VideoQA dataset [6], or ReVOS dataset [9] individually results in varying degrees of metric degradation. Notably, removing the VideoQA dataset minimally impacts the model's performance, with a decline of **0.9**% in $\mathcal{J}\&\mathcal{F}$ on both the referring and reasoning subsets, as its primary role is to support the MLLM's video comprehension rather than directly contributing to the segmentation process. In contrast, excluding the ReVOS dataset leads to a noticeable drop of **4.4**% and **7.6**% in $\mathcal{J}\&\mathcal{F}$, highlighting its pivotal role in enhancing the model's reasoning segmentation performance in challenging scenarios.

Table 1. Ablation study on the impact of training datasets.

Datasets	referring			reasoning		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Joint	**59.8**	**64.5**	**62.1**	**53.5**	**58.7**	**56.1**
w/o ImageSeg	58.5	63.2	60.8	51.0	56.3	53.6
w/o VideoQA	58.7	63.7	61.2	52.4	58.0	55.2
w/o ReVOS	55.3	60.1	57.7	45.3	51.6	48.5

B. Additional Implementation Details

Due to space constraints of the main document, additional implementation details are provided here. During training, we use varying sampling ratios for different datasets (*cf.* Tab. 2). For video segmentation datasets, 8-12 frames are uniformly sampled at fixed intervals per video, and up to three object categories are selected per image or video. During inference, we utilize CLIP-336 [7] for global sampling, selecting up to 12 frames per video. Input images are resized to 224 × 224 before being input to Chat-UniVi [3]. Data passed to SAM2 is augmented as described in [4] and resized to 1024 × 1024. Moreover, LoRA [2] is applied with a scaling factor of 16 and a dropout rate of 0.05 across all query and value projection layers within the MLLM, enabling efficient fine-tuning.

Table 2. Datasets sampling ratio during training.

Dataset	SemSeg	RIS	ImageQA	ReaSeg	VideoQA	VideoSeg
Ratios	9/32	3/32	3/32	1/32	1/8	3/8

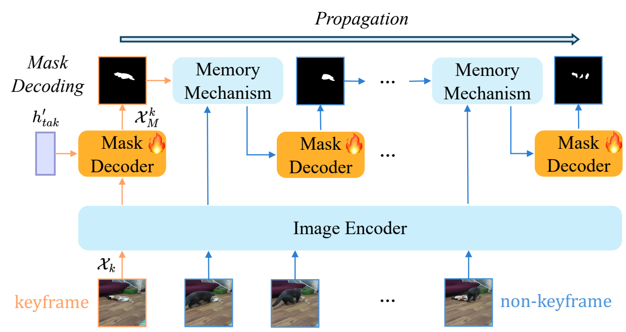

Figure 1. Details of SAM2 for mask decoding and propagation. All the video frames are input into the image encoder for feature extraction. The feature embeddings of the keyframe interact with h'_{tak} through the mask decoder for mask generation and then propagate it to the remaining video frames via the memory mechanism.

C. More Details of SAM2

As depicted in Fig. 1, we provide detailed insights into the process of mask decoding and propagation using SAM2 [8]. Specifically, all input video frames are processed through the image encoder to extract multi-scale visual features. Subsequently, the fused temporal embedding h'_{tak} interacts with the keyframe features in the mask decoder to generate the segmentation mask and perform video-level propagation. The prediction is then encoded by the memory encoder and stored in the memory bank, which maintains a FIFO queue of memories from recent frames. Feature embeddings from subsequent non-keyframes attend to these stored mask features through memory attention and utilize the mask decoder to generate corresponding masks, enabling inter-frame propagation.

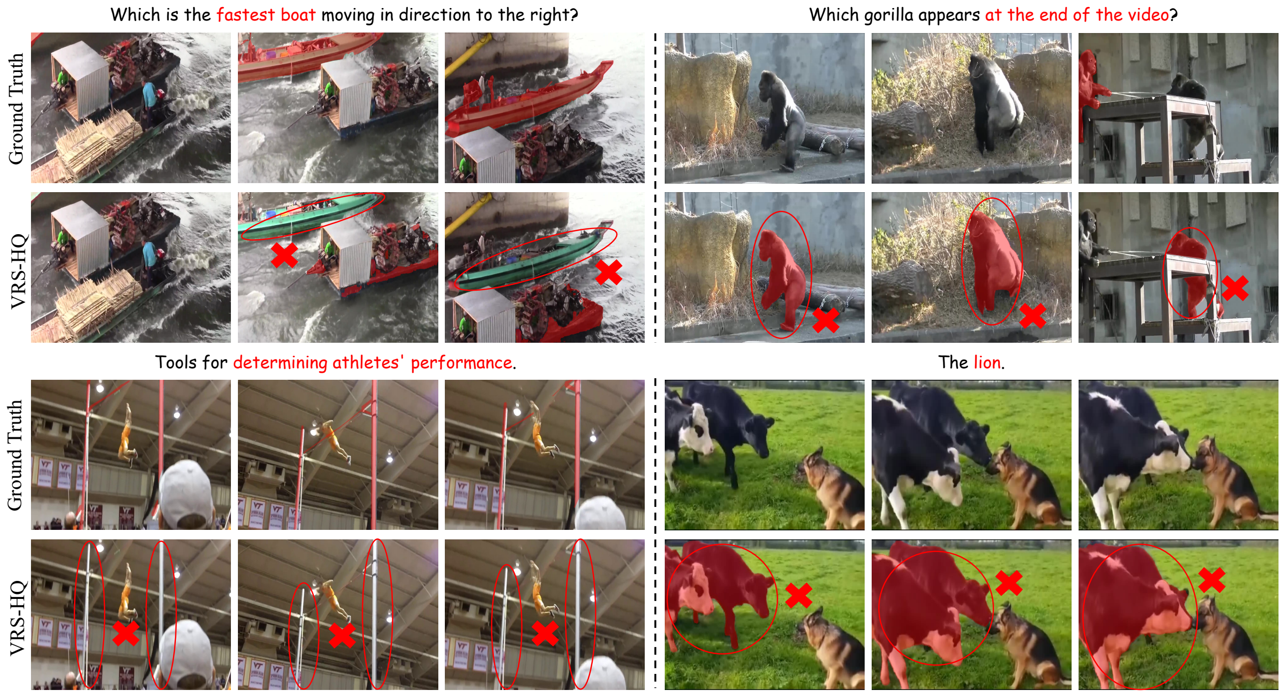

Figure 2. Visualization of failure cases for VRS-HQ. These examples illustrate the model's limitations in scenarios requiring complex world knowledge and temporal reasoning, as well as challenges in processing negative samples.

D. Failure Case Analysis

Fig. 2 presents a detailed analysis of several failure cases, offering a deeper understanding of the limitations of VRS-HQ. The **top row** highlights two specific challenges. **First**, VRS-HQ struggles with keyframe localization when presented with queries based on motion, such as identifying the fastest-moving boat within a video sequence. This suggests a potential weakness in analyzing and interpreting dynamic visual information. **Second**, the model exhibits difficulty segmenting targets with minimal temporal presence, as exemplified by the gorilla visible only in the last two frames of the video. This points to a possible limitation in effectively capturing and utilizing short-duration visual cues. The **bottom row** reveals further limitations. VRS-HQ demonstrates a lack of comprehension when faced with nuanced or implicitly phrased prompts, such as recognizing a "high bar" within the context of gymnastics performance evaluation. This suggests a need for improved understanding of complex semantic relationships within video content. Furthermore, the model occasionally exhibits hallucinatory behavior, generating segmentations for non-existent objects, particularly when dealing with empty targets or scenes where the requested object is absent.

We hypothesize that several strategies could mitigate these limitations. Improving the video comprehension capabilities of the Multimodal Large Language Model (MLLM) could enhance the ability to interpret complex scenes and queries. Enabling the model to process a larger number of sampled frames simultaneously might improve its sensitivity to subtle temporal changes and short-duration events. Finally, designing specialized tokens specifically for representing empty masks could address the observed hallucinations in such scenarios. We leave a thorough investigation of these potential improvements to future research.

E. VQA Task Results

To explore the relationship between dense prediction tasks and multimodal QA, we evaluate VRS-HQ against its foundation MLLM Chat-UniVi, on the POPE [5] benchmark, as illustrated in Tab. 3. VRS-HQ performs better on multimodal QA despite being designed for reasoning segmentation, demonstrating the synergistic relationship between these tasks and the potential for cross-task improvements.

Table 3. Results on the POPE benchmark.

Methods	POPE-R		POPE-P		POPE-A	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Chat-UniVi	85.19	86.05	69.50	74.39	64.97	71.54
VRS-HQ	**87.25**	**87.18**	**75.40**	**77.38**	**70.40**	**73.97**

F. More Qualitative Comparison

In addition to the visual comparisons presented in the main document, we provide further comparisons across more diverse settings in Fig. 3-5 to demonstrate the model's reasoning and segmentation capabilities. As illustrated in Fig. 3, VISA demonstrates reduced sensitivity to color-related expressions (e.g., "white" and "brown") when provided with explicit textual instructions. Furthermore, the example on the left demonstrates VISA's tendency to misidentify visually similar objects with complex spatial variations. In contrast, VRS-HQ effectively aggregates temporal information, capturing inter-frame motion dynamics and leading to improved segmentation accuracy.

Fig. 4 highlights the robust segmentation and reasoning capabilities of VRS-HQ in scenarios with complex temporal dynamics. In the left example, VISA struggles to precisely detect the airplane appearing on the left at the end of the video. Similarly, in the right case, VISA misclassifies the tiger emerging in the lower left corner. In contrast, VRS-HQ leverages the Token-driven Keyframe Selection for more accurate keyframe identification and integrates SAM2 with the temporal token, enriched with both intra-frame spatial and inter-frame temporal relations, resulting in reliable decoding and consistent object tracking.

Fig. 5 presents scenarios requiring general and world knowledge for reasoning. In the first example (left), VISA segments only two koi carp (Cyprinus carpio) correctly, whereas VRS-HQ identifies nearly all the fish present. In the second example (right), VISA fails to associate "dog" with the phrase "common household pet", indicating limitations in its reasoning capabilities. By contrast, VRS-HQ leverages the integration of temporal tokens to achieve a more nuanced semantic understanding, enabling finer control and interpretation.

G. In-the-wild Visualization Results

Fig. 6 and Fig. 7 show qualitative results of VRS-HQ on in-the-wild videos. Fig. 6 shows results on first-person videos from the GTEA dataset [1], using implicit prompts. Even in cluttered kitchen environments with many similar objects, VRS-HQ demonstrates strong generalization capability. It is particularly effective at segmenting smaller targets, such as the spoon and watch shown in the first and third rows, respectively, maintaining robust performance in these challenging scenarios. Fig. 7 shows results on 360-degree panoramic videos from the PanoVOS dataset [10], using more intricate prompts. Notably, VRS-HQ successfully segments individuals even when they are split across the distorted edges of the video (first row), without any task-specific optimizations. Furthermore, it maintains effective tracking performance when the primary subjects within the video are moving dynamically (last two rows).

References

[1] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011.

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[3] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13700–13710, 2024.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Int. Conf. Comput. Vis.*, pages 4015–4026, 2023.

[5] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[6] 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.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763. PMLR, 2021.

[8] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[9] 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 *Eur. Conf. Comput. Vis.*, 2024.

[10] Shilin Yan, Xiaohao Xu, Renrui Zhang, Lingyi Hong, Wenchao Chen, Wenqiang Zhang, and Wei Zhang. Panovos: Bridging non-panoramic and panoramic views with transformer for video segmentation. In *European Conference on Computer Vision*, pages 346–365. Springer, 2025.

CVPR #6328

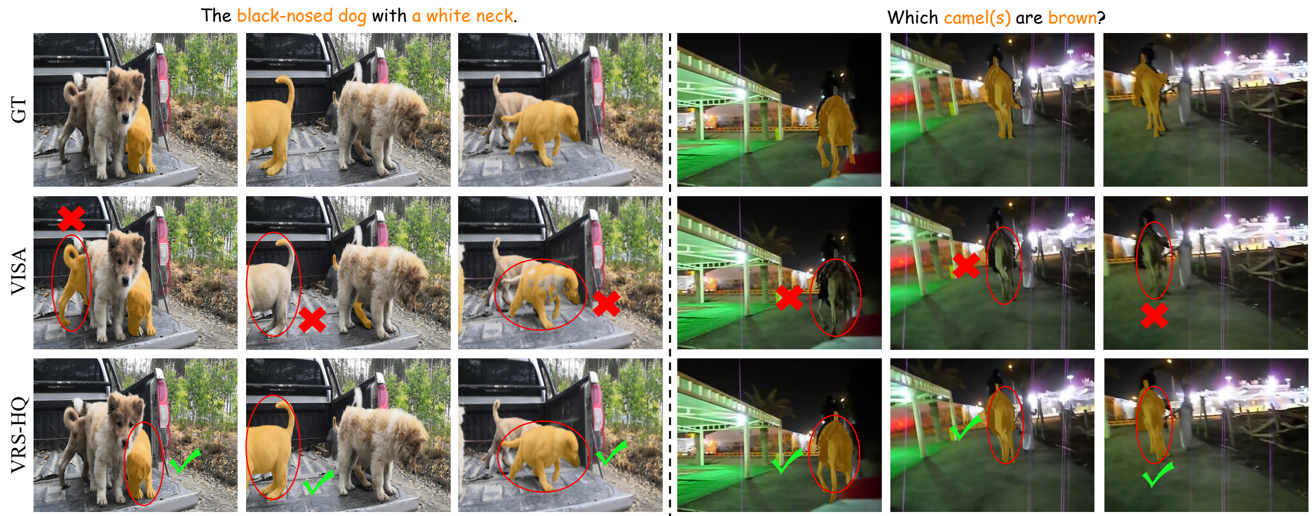

Figure 3. Qualitative comparison of VRS-HQ and VISA in explicit language-based referring scenarios on the ReVOS benchmark.

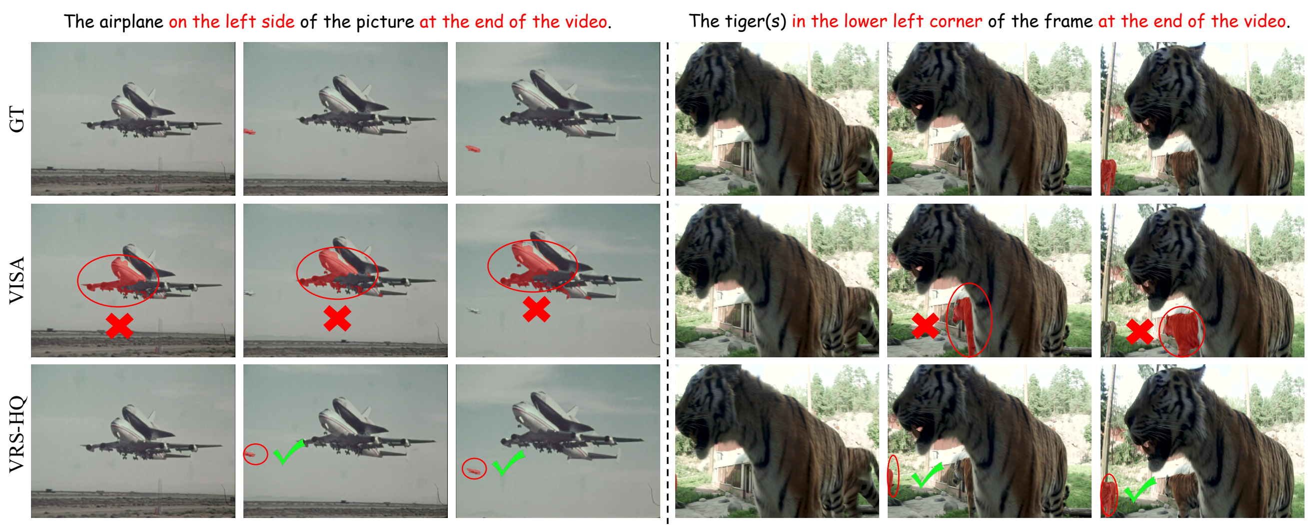

Figure 4. Qualitative comparison of VRS-HQ and VISA in scenarios incorporating complex temporal dynamics on the ReVOS benchmark.

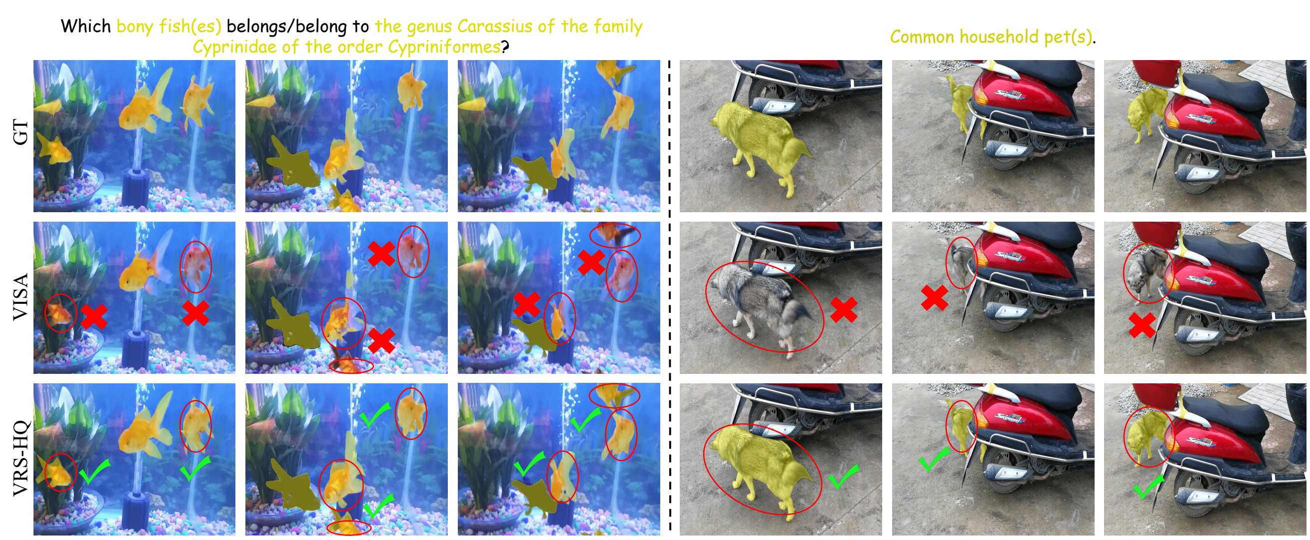

Figure 5. Qualitative comparison of VRS-HQ and VISA in reasoning scenarios that require world knowledge on the ReVOS benchmark.

The video's subject uses a curved tool for scooping and serving liquids or soft solids, which typically features a rounded bowl and a handle, ensuring ease of use and practicality.

A white object with a circular smooth surface, typically placed on a table and unmoved by the videographer, designed to hold and present various types of food, offering functionality in dining settings.

An item used for timekeeping in daily life and offering aesthetic appeal, commonly worn on the wrist and often featuring a circular or rectangular design, occasionally appearing in the video.

Figure 6. Visualization of VRS-HQ utilized in egocentric videos.

A prominent figure subtly highlighted within a commercial promotional video, whose presence and actions serve as the central point for engagement and communication.

The individual demonstrating a more seasoned presence and refined mastery of BMX Freestyle maneuvers, embodying the skill and experience often associated with years of dedication to the sport.

Who displays the most dynamic and expressive range of movement during the dance, transitioning seamlessly between sharp, high-energy motions, captivating attention with their vibrant and energetic performance.

Figure 7. Visualization of VRS-HQ applied to 360-degree panoramic videos.