ViLLa: Video Reasoning Segmentation with Large Language Model

Supplementary Material

This supplementary material provides more details about the proposed ViLLa, and our proposed benchmark, VideoReasonSeg. The first part includes discussions about the design of ViLLa and its comparison with previous methods, followed by the implementation details. Then, we provide extra ablation experiments. What's more, we include the data generation pipeline of our VideoReasonSeg dataset, and further data cases demonstrating the variety of our data sample. Finally, we include failure case analysis. The content is organized as follows:

- Discussions of ViLLa's differences with previous methods.
- The implementation details of ViLLa.
- More ablation study experiment of ViLLa.
- The data details and the data generation pipeline of our proposed VideoReasonSeg dataset.
- Failure case analysis of ViLLa.

A. Discussions

ViLLa is a holistically designed framework for Video Reasoning Segmentation (VRS), a burgeoning task demanding spatiotemporal tracking, segmentation, and dynamic reasoning in complex, evolving scenes. While prior works (e.g., LISA, PixelLM, Mask2Former) focus on reasoning static images or tracking through short clips, their architectures inherently mismatch VRS needs for two major reasons: **1) Temporal Reasoning Gap:** Static MLLMs lack mechanisms to model motion, causality, or long-term dependencies in videos, while VOS/VIS methods lacks high-level reasoning beyond tracking. **2) Integration Challenges:** Simply grafting spatial reasoning modules (e.g., ViLLa†) onto temporal trackers incurs additional computational costs and suboptimal performance. ViLLa bridges these barriers via three novel, task-specific components: i) Key Segment Extractor: Identifies critical temporal segments to reduce redundancy. ii) Context Synthesizer: Fuses and condenses long-term spatial-semantic reasoning information with dynamic scene evolution. iii) Hierarchical Temporal Synchronizer: Ensures consistency across long-term dependencies and modeling of complex scenes via multi-scale aggregation with multi-level segmentation tokens. As shown in Tab. 1, ViLLa achieves +3.5/2.4 gains over ViLLa† (adapted from LISA with designs from PixelLM, Mask2Former, and curated data), proving that adaptive integration of spatiotemporal reasoning modules—not direct reuse of static models—is essential and effective for VRS. In short, our VRS-oriented design novelly addresses understudied challenges in VRS.

Table 1. **Comparison** of direct adaptation of former approaches.

Method	VideoReasonSeg		MeViS		
	$\mathcal{J}\&\mathcal{F}$	Accuracy	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ViLLa†	51.9	44.0	47.0	43.9	50.1
ViLLa	**55.4**	**49.9**	**49.4**	**46.5**	**52.3**

Table 2. **Training hyperparameters** for ViLLa.

Config	Value
input resolution	224
max text length	512
optimizer	AdamW
optimizer momentum	β_1, β_2 = 0.9, 0.999
weight decay	0.02
learning rate schedule	cosine decay
learning rate	2e-5
batch size	32
warmup iters	10

B. Implementation Details

Training Details. In the first part, we present the detailed training configuration in our Tab. 2. As for the λ_{txt} and λ_{mask}, they are set to 1.0, and λ_{dice} is 0.5 while λ_{ce} is 2.0.

C. Additional Ablation Experiments

Sampling Strategy. In the key segment extractor, we take the average from the top-K responses to obtain the starting and ending frames of the key segments, where we denote as V_{key}, comprising T_{key} frames. Based on the key segments, we also sample T_{ref} using an adaptive global sampling strategy. 'Global' indicates sampling T_{ref} frames from the whole video apart from the key segments we extract, and 'neighbor' denotes sampling frames from the neighboring frames (both precedent and antecedent) of the key segments. Our 'adaptive' sampling strategy, on the other hand, combines both 'global' and 'local' sampling strategies, which samples $T_{ref}/3$ from the whole video, and $2/3T_{ref}$ from the neighboring frames. As shown in Tab. 3, adaptive sampling slightly outperforms both global and neighbor sampling strategies. As the number T_{ref} increases, the performance gradually improves.

Aggregation Strategy. Tab. 4 shows the results of different aggregation strategies in the segment synchronization decoder. We compare our aggregation between video-level segmentation embeddings with the feature fusion adopted in PixelLM [5]. As shown in the table, our strategy improves the performance on referring VOS dataset, demonstrating the effectiveness of the video-frame aggregation strategy.

Video Instance Segmentation. Tab. 5 presents the results on the video instance segmentation datasets. YouTube-VIS 2019 [8], contains 2.9k videos. The dataset was updated to

Table 3. **Ablation study** on different sampling strategies.

Strategy	T_{ref}	ReasonVideoSeg $\mathcal{J}\&\mathcal{F}$	ReasonVideoSeg Accuracy
Global	0	54.0	47.8
	6	54.3	48.3
	12	54.5	48.7
Neighbor	0	54.0	47.8
	6	54.4	48.5
	12	54.7	49.2
Adaptive	0	54.0	47.8
	6	54.8	49.0
	12	**55.4**	**49.9**

Table 4. **Ablation study** on aggregation strategies.

Strategy	Ref-YouTube-VOS $\mathcal{J}\&\mathcal{F}$	Ref-YouTube-VOS $\mathcal{J}$	Ref-YouTube-VOS $\mathcal{F}$	Ref-DAVIS17 $\mathcal{J}\&\mathcal{F}$	Ref-DAVIS17 $\mathcal{J}$	Ref-DAVIS17 $\mathcal{F}$
Feature Fusion	65.9	64.1	68.5	63.2	60.5	66.1
Embedding Similarity	**66.5**	**64.6**	**68.6**	**64.4**	**61.2**	**67.7**

Table 5. **Comparison** on multiple VIS datasets.

Method	YTVIS-19 AP	YTVIS-21 AP	OVIS AP
SeqFormer [6]	59.3	51.8	-
Mask2Former [1]	61.6	55.3	24.1
VITA [2]	63.0	57.5	27.7
IDOL [7]	64.3	56.1	42.6
ViLLa	**67.6**	**59.9**	**46.5**

YouTube-VIS 2021 with longer videos. OVIS dataset is another resource for video instance segmentation, particularly focusing on scenarios with severe occlusions between objects [4]. It consists of 25 object categories and 607 training videos. Our ViLLa surpasses previous SOTA VIS methods by 3.3, 3.8, and 3.9 points, respectively. The results prove that our model is excelling at modeling temporal relations and segmenting high-quality tracklets.

D. VideoReasonSeg Details

In order to generate multiple-choice QA, we automatically convert the video annotations into this format via LLMs. Specifically, we first use ChatGPT [3] to generate a question for each video. For most questions, we construct the option candidates directly from the ground truth annotations. For example, video segmentation tasks contain masks and instance categories of each video. Then the candidate option for multiple-choices would be the *correct* category, *wrong* category, and a *not-sure* choice. Ultimately, we produce 2 pairs for each of the video. To strengthen the evaluation's robustness, for each question we randomly sample 3 to 5 answer options from the available candidates and shuffle the order of the options. Additionally, to prevent the common issue of answer leakage where longer options tend to be correct, we further use a large language model to ensure that all the answer options for a question are of similar and reasonable lengths.

As for the question and answer pair, we use GPT-4V to construct our dataset. We utilize videos with pre-existing video mask annotations. The video frames, the category names contained in the video, and their related mask annotation are contained in the prompts fed to the GPT-4V. An example of the prompts is shown in Fig. 1. Using carefully crafted prompts, GPT-4V autonomously selects instances to construct question-answer pairs relevant to the video. As illustrated in Fig. 2, we demonstrate two types of questions, both question and multiple choice, from a given video. In this example, we show that our data tests the capability of models of reasoning based on common world-knowledge, and relate 'vehicle carrying passengers' to the white bus on the roadside. In addition, the multiple choice expects the model to distinguish the type of vehicle from other plausible answers, such as 'taxi' and 'bike'. All together these questions are tests of the reasoning capacities of models on both pixel-level and video-level.

During the generation process, we compared GPT-4V with Qwen-2VL and other contemporary models, while other models performed unsatisfactorily in generating well-aligned QAs with long instructions. 2) **Prompting and refinement strategy.** To ensure the instruction-mask alignment, we use GPT-4V to generate data in multiple steps: GPT-4V was explicitly primed with a structured task schema. We also add constraints, such as *"Avoid subjective language"*. After generating instructions, GPT-4V was prompted to score its output on a 1-5 scale, and samples with scores lower than 4 were automatically discarded. Instructions containing vague terms were filtered via keyword rules. 3) **Human evaluation.** After this autonomous process, we invite experts to sample and filter low-quality QA pairs (20%).

Although GPT-4V can efficiently understand the content of the video frame, there are still failure cases in the generated data. One major problem is that questions can be too objective and hard to evaluate. For example, the question "How would you rate the overall difficulty and impressiveness of the skateboarding you observed?" is very objective, and the answers can vary for different people. This requires further prompts and filtering during data generation process.

Video visualizations of additional data of the VideoReasonSeg are further presented in Fig. 3, which shows varied cases that include: a) discrimination from multiple instances; b) multiple instances with fast movement; c) open-world knowledge reasoning.

E. Failure Case Analysis

Even though our ViLLa shows impressive results in video reasoning segmentation, there is still room for improvement. As shown in Fig. 4, ViLLa incorrectly segments the bystander who is observing the two-person talking. We hypothesize that this error arises from the inability of the MLLM to temporally localize the "talking" action, which occurs exclusively in the final three frames (there is even an occlusion in the last frame). Consequently, ViLLa erroneously associates the individual with the motor-riding man in the initial frames and persistently tracks this subject throughout the video.

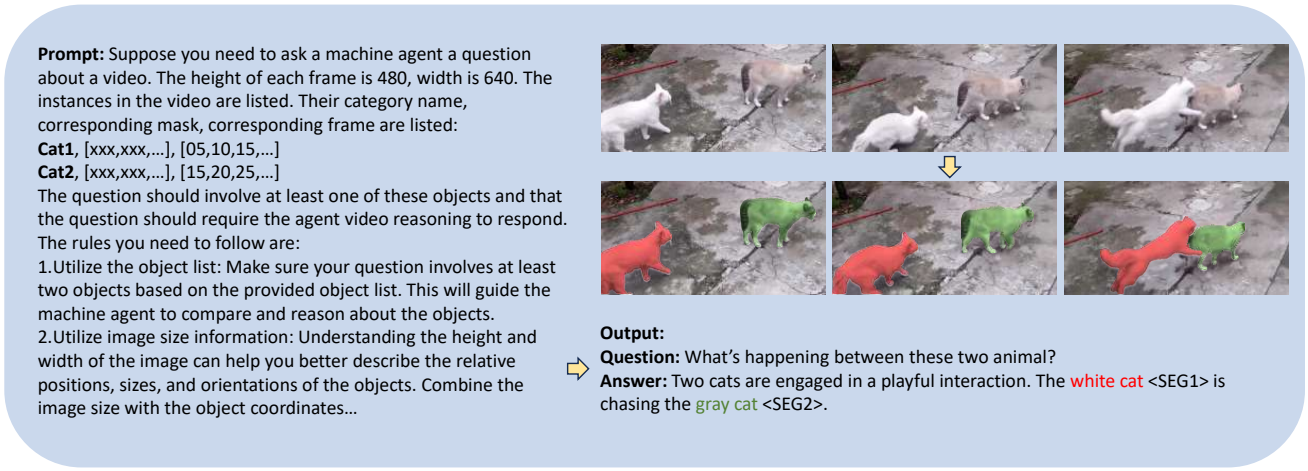

Figure 1. **GPT-4V data generation pipeline.** The right part shows an example of how reasoning segmentation data and multiple choices are generated. The input prompt includes certain rules and the position as well as time localizations to instruct GPT-4V into generating more effective data samples.

Question: <VIDEO> What mode of transportation is depicted in the video that is commonly used for carrying passengers on roads?
Answer: The bus <SEG>.

Question: <VIDEO> What vehicle is shown parked on the roadside in the rural setting depicted in the video?
Choices: A) A taxi cab B) A bike C) A long-distance bus D) A delivery truck

Figure 2. **GPT-4V data generation samples.** The part shows further samples of the generated questions and the multiple choices. The two types of questions can help us better evaluate the model's performance in video reasoning at both pixel-level and video-level.

References

[1] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv:2112.10764*, 2021. 2

[2] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. In *NeurIPS*, 2022. 2

[3] OpenAI. Chatgpt: A language model for conversational ai. Technical report, OpenAI, 2023. 2

[4] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *IJCV*, 2022. 2

[5] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *CVPR*, 2024. 1

[6] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *ECCV*, 2022. 2

[7] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022. 2

[8] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1

(a)

(b)

(c)

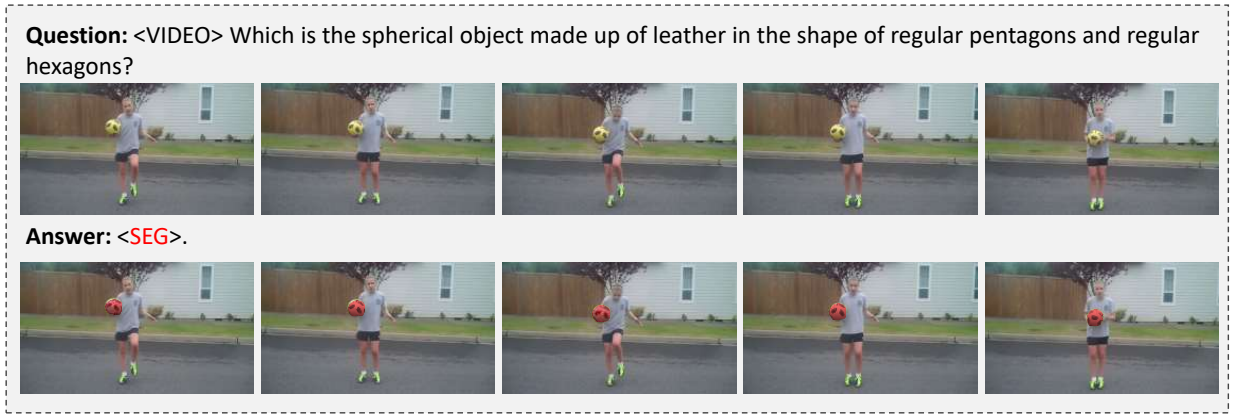

Figure 3. **Data samples** of our proposed VideoReasonSeg.

Figure 4. **Failure case.** ViLLa incorrectly segments the bystander who is observing the two-person talking.