Refer-Agent: A Collaborative Multi-Agent System with Reasoning and Reflection for Referring Video Object Segmentation

Supplementary Material

A. Additional Ablation Studies

Effects of Video Object Segmentation Models. In our primary reasoning pipeline, we employ SAM2 as our video object segmentation model. Here, we replace SAM2 with alternative models (e.g., Xmem++ [1] and Cutie [2]). As shown in Table A1, SAM2 achieves superior performance, surpassing the other two models by 0.9% and 1.1% $\mathcal{J}\&\mathcal{F}$ respectively. Even without SAM2, our framework still outperforms SAM2-based COT-RVS [3] by over 3.2% $\mathcal{J}\&\mathcal{F}$, demonstrating the robustness of our framework.

Table A1. Ablation study of video object segmentation models.

Method	Tracker	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
COT-RVS [3]	SAM2 [4]	65.5	62.4	68.7
Refer-Agent (Ours)	Xmem++ [1]	68.9	65.7	72.1
	Cutie [2]	68.7	66.0	71.3
	SAM2 [4]	69.8	67.0	72.7

Frame Sampling Strategy. During the frame selection stage, we combine CLIP and MLLM to achieve a course-to-fine frame sampling. Here, we compare it with two frame sampling alternatives: 1) Uniform Sampling, which selects N frames uniformly; and 2) Top-N Sampling, which picks the top-N frames based on CLIP scores. As shown in table A2, relying solely on CLIP may result in negative gains, and our approach achieves the best performance through the course-to-fine sampling strategy.

Table A2. Ablation study of coarse frame sampling strategy.

Strategy	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Uniform Sampling	69.1	66.1	72.2
Top-N Sampling	65.2	62.3	68.1
Ours	**69.8**	**67.0**	**72.7**

B. More Details of Frame Selection Refinement

Once the verification during Existence Reflection fails, the system will return the feedback and triggers a new round of frame selection. In the new round, CLIP calculates the semantic similarity based on not only the original query, but also the object descriptions generated by the last-round Intent Analysis Agent:

$$S_{\text{CLIP}} = S_{\text{query}} + \gamma \cdot S_{\text{exp}}, \tag{1}$$

where S_{query} and S_{exp} denote the semantic similarity with each frame for the textual query and the generated descriptions, respectively. And the coefficient γ is set to 3.

C. Additional Qualitative Analysis

More Qualitative Comparisons with SOTAs. In Figure A1, we present more qualitative comparison results. In the first sample, the target exhibits the "squatting" state only in the final few frames. While AL-Ref-SAM2 incorrectly segments another prominent man and GLUS fails to capture the target (outputting empty masks), our method successfully focuses on the end of the video and accurately segments the target. In the second sample, which aims to segment a plastic bottle, the action described in the query appears only briefly in the early part of the video. Our approach effectively captures such localized motion information while rejecting distractors, while competitors incorrectly segment irrelevant objects. In the third sample, designed to evaluate model performance when no corresponding target exists in the video, AL-Ref-SAM2 segments the background and GLUS mistakenly identifies fishes as divers. In contrast, our method can correctly handle such scenarios through alternating reasoning and reflection. These results demonstrate the effectiveness of our approach for segmenting objects in different scenarios.

Details of Reflection Chain. We present a comprehensive illustration of the reflection chain in Figure. A2. This mutual question-response process naturally forms a chain of reflection to verify the correctness of intermediate results. These evidences are ultimately summarized and fed back into the main pipeline, guiding the next round of reasoning. In the left sample, the Existence Reflection agent finds that in the selected keyframe, the two birds are overlapping and a third bird is missed. Hence, it triggers a new round of frame selection, explicitly instructing that "*the keyframe must clearly show all three birds without overlap*". In the right sample, the Consistency Reflection agent successfully identifies that the girl using a vacuum cleaner (an electrical appliance) is a correct target, whereas the man holding a broom (not an electrical appliance) is an incorrect target. In this way, our Refer-Agent effectively mitigates the impact of MLLM hallucinations during the complex reasoning process, enabling robust reasoning segmentation performance.

D. Failure Case

We present a failure case of Refer-Agent in Figure A3. Specifically, for the query "Which animal was used as the party emblem by the American Whig Party?", the Intent Analysis agent incorrectly identifies a skunk as the tar-

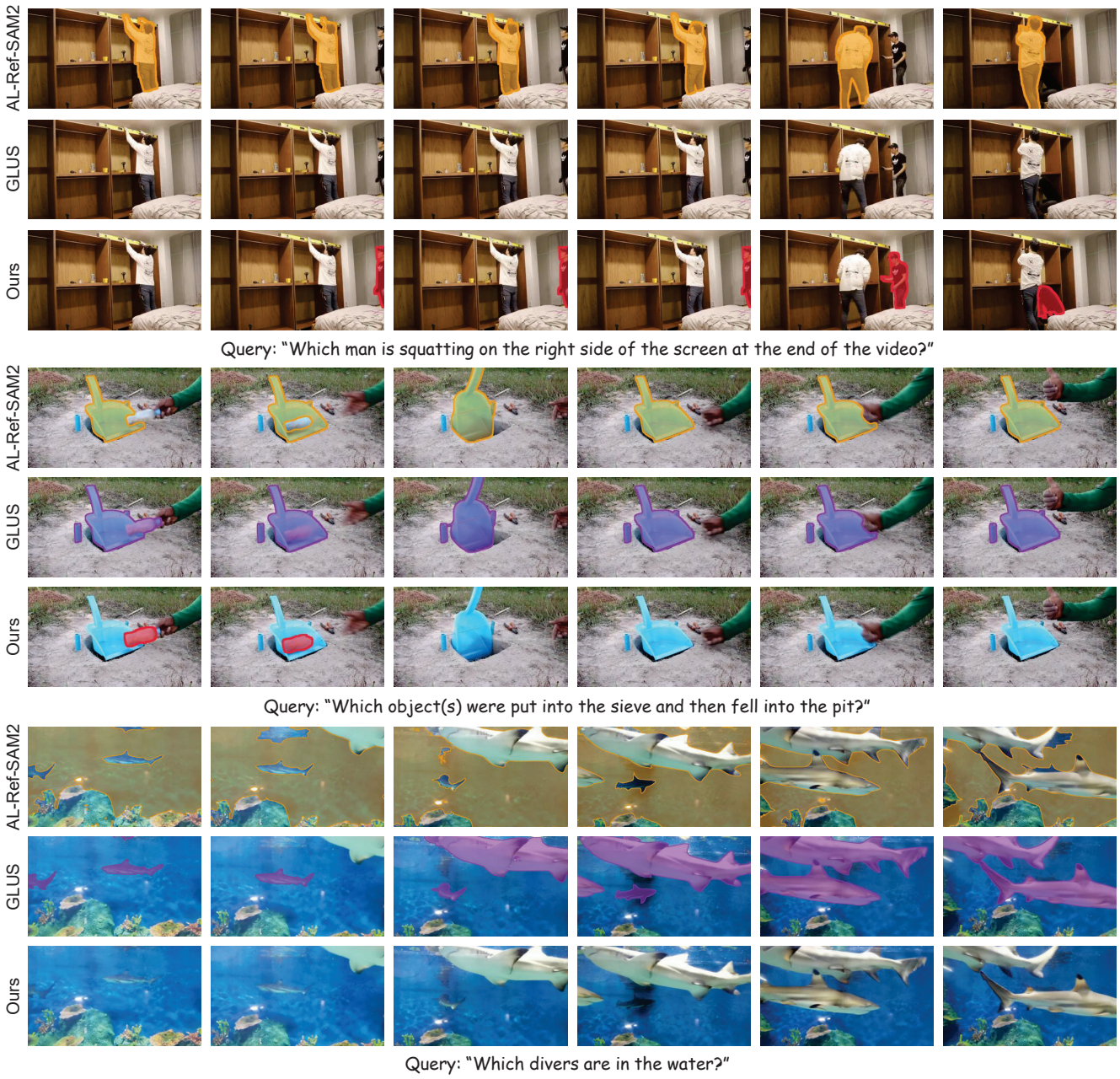

Figure A1. Qualitative comparison of our Refer-Agent with SOTA. In the first sample, our model can focus on the end of the video and segment the target, while AL-Ref-SAM2 segments the prominent object and GLUS outputs empty masks. In the second sample, competitors mistakenly segment the irrelevant objects while our method captures the localized motion information and segments the correct target. In the third sample, our approach successfully outputs empty masks because no corresponding target exists, while competitors fail.

get. This error is failed to be identified by Consistency Reflection, ultimately leading to incorrect segmentation results. Such cases often involve domain-specific knowledge that goes beyond the internal knowledge of general-purpose MLLMs. To address this limitation, Retrieval Augmented Generation (RAG) or search engines could be integrated, which is a promising direction for future enhancements.

E. Prompts or Templates of Refer-Agent

We present the detailed prompts and question templates used in our Refer-Agent in Table A3.

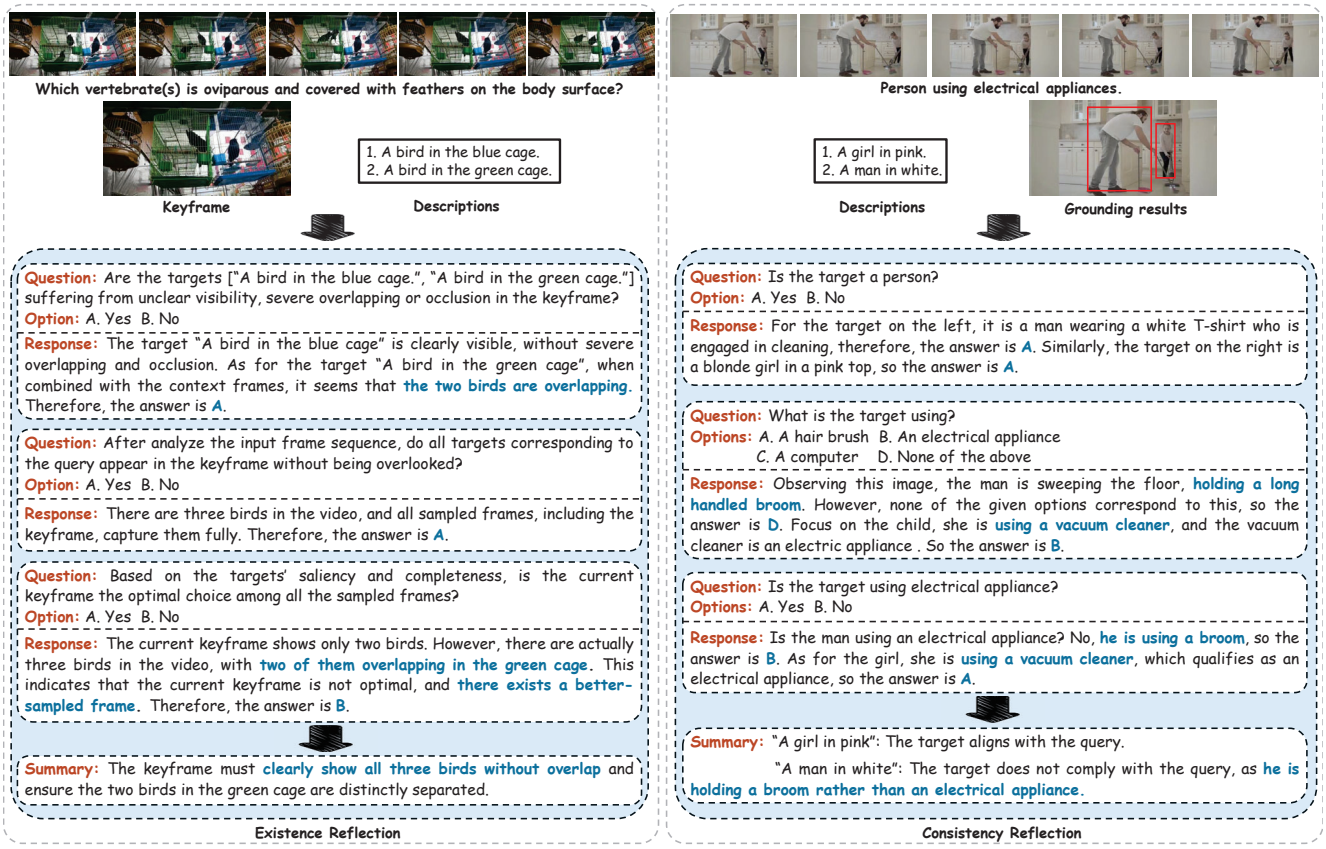

Figure A2. Reflection chain of Existence Reflection and Consistency Reflection. **Left**: Detail process of Existence Reflection. **Right**: Detail process of Consistency Reflection. The reflection chain mechanism effectively assists the model in identifying potential issues.

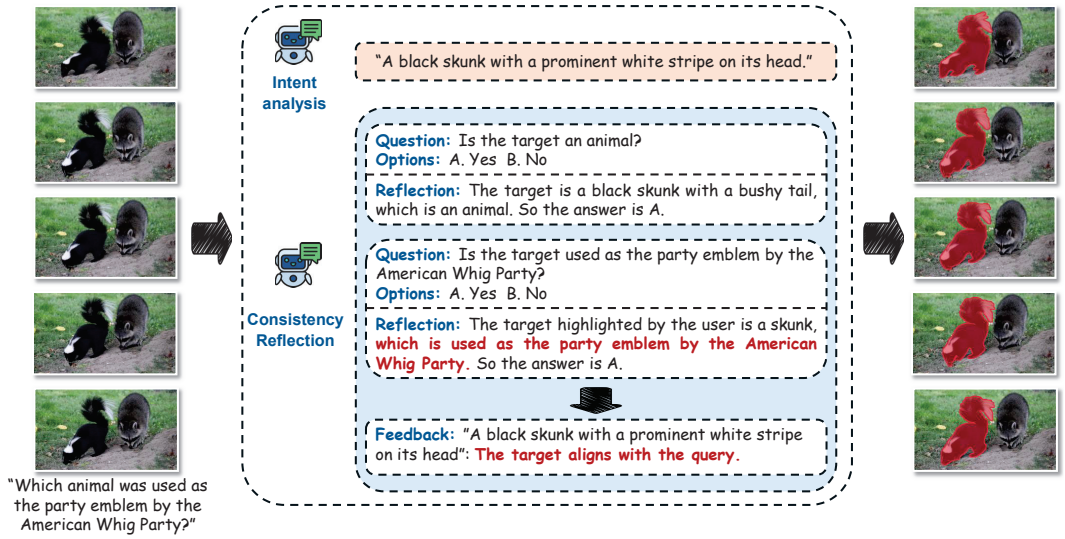

Figure A3. A failure case of Refer-Agent.

Table A3. The prompts and question template used in the key components in Refer-Agent.

Prompt for frame scoring

You are given a query and a sequence of frames, each labeled with a digit as its frame_id indicating its temporal position. As well as an optional attention information for assistance.
1. Describe the events happening in the frame sequence. These frames together form a continuous video segment. You must analyze the frames as a sequence to understand temporal behavior and target dynamics.
2. Score each frame. Integrate frame sequences to understand the query and identify all targets in the video that match the query. Locate all appearing targets that match the query in each frame, and observe their presence in the frame (fully and clearly visible, partially visible, blurred, etc.). Rate each frame on a scale of 1 to 10, where a higher score indicates that the frame is more suitable for displaying all the targets corresponding to the query. If the query primarily describes the state of the target, the frames that effectively exhibit these states should be assigned higher scores.
3. Use the attention information. The attention information (if available) contains guiding information to assist with scoring, primarily indicating what characteristics make a frame suitable or unsuitable as a keyframe. You can use this information to help with your scoring.
The query: {`query`}
The attention information: {`attn_info`}
You must output a valid JSON object with the exact structure:
{"`scores`": a list of numbers indicating each frame's score}

Prompt for generating descriptions

You are given a query and a spliced video frames, each labeled with a digit as its frame_id indicating its temporal position. The keyframe is the largest while the remaining frames are smaller.
1. Describe the events happening in the frame sequence. These frames together form a continuous video segment. You must analyze the frames as a sequence to understand temporal behavior and target dynamics.
2. Determine the targets and provide descriptions. Analyze the given query to determine the main target category. Determine the qualified targets exist in the keyframe and provide corresponding unique and concise descriptions for each target. Make sure each description only describes one target, with the most important distinguishing feature that can uniquely identify the target in keyframe. Static features should be prioritized. Do not mention frame numbers or the phrase like 'at keyframe'. Limit each description to 25 words.
3. Use previous descriptions and feedback. These details (if available) show the alignment and divergence between previous descriptions and query. You can use this information to help with your analysis.
The query: {`query`}
Keyframe ID: {`keyframe_id`}
Previous descriptions and feedback: {`history`}
You must output a valid JSON object with the exact structure:
{"`target_count`": the number of qualified targets, "`descriptions`": ["desc1", "desc2", ...]}

Prompt for object grounding

Find the `<ref>`{`description`}`</ref>` in the image. Compare the difference between objects and find the most closely matched one. Please give the coordinates of the bounding box.

Template for questioner of existence reflection

1. Is the target "{`description`}" suffering from unclear visibility, severe overlapping or occlusion in the keyframe?

2. After analyze the input frame sequence, do all targets corresponding to the query appear in the keyframe without being overlooked?
3. Based on the targets' saliency and completeness, is the current keyframe the optimal choice among all the sampled frames?

Prompt for responder of existence reflection

You are given a query and a spliced video frames, each labeled with a digit as its frame_id indicating its temporal position. The keyframe is the largest while the remaining frames are smaller. You are also given a list of questions about the above information.
1. Analysis Instructions: Analyze the frames as a continuous video sequence to understand temporal behavior and target dynamics. For each question in the list, select the optimal answer option based on visual evidence from the video content. Provide brief, factual reasoning for each answer choice, focusing specifically on what is observed in the video frames.
2. Key Frame Selection Guidance: After answering all questions, provide targeted guidance for keyframe selection that specifically addresses how to best display targets matching the query query. The guidance should: 1) be based on the query, identify which specific targets need to be prominently displayed; 2) provide specific, actionable criteria for selecting frames that best highlight the targets or how to capture the most representative or informative moment for these targets.
The query: {`query`}
Keyframe ID: {`keyframe_id`}
Questions: {`questions`}
You must output a valid JSON object with the exact structure:
{"`answers`": [{"`answer`": The selected option letter, "`reason`": One sentence providing brief visual explanation}, ...], "`guidance`": A concise sentence providing targeted guidance about keyframe selection specifically for displaying query-related targets.}

Prompt for questioner of consistency reflection

You are given a query.
1. If it is an interrogative sentence, convert it into the format of a declarative sentence. If the subject cannot be clearly identified, use "the target" as a substitute.
2. For the processed query, decompose the attributes of the subject into: high-level concepts (object category and state) and low-level details (appearance, shape, and spatial location). Based on these decomposed attributes, design questions (Is, What, Which, etc.) using facts from the query as correct answers and reasonably construct distractors that neither contain nor overlap with the correct answers. For non-"Is" type questions, include an option for "None of the above."
The query: {`query`}
You must output a valid JSON object with the exact structure:
[{"`question`": the designed question1, "`options`": the designed options for this question, "`correct_answer`": the correct option letter for this question}, ...]

Prompt for responder of consistenct reflection

You are given a questions dict and an image with a red bounding box. Carefully analyze the image and visual content within the red bounding box. For each question, select the optimal options that accurately describe the visual attributes observed within the bounded area. Provide specific and short factual reasons grounded in the actual image content.
The questions: {`questions`}
You must output a valid JSON object with the exact structure: {"`answers`": [{"`answer`": The selected option letter, "`reason`": 1 sentence for brief visual explanation.}, ...]}

References

[1] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 635–644, 2023. 1

[2] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3151–3161, 2024. 1

[3] Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Cot-rvs: Zero-shot chain-of-thought reasoning segmentation for videos. *arXiv preprint arXiv:2505.18561*, 2025. 1

[4] 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. 1