arXiv:2508.10572v1 [cs.CV] 14 Aug 2025

Towards Agentic AI for Multimodal-Guided Video Object Segmentation

Tuyen Tran Thao Minh Le Truyen Tran
Applied Artificial Intelligence Institute, Deakin University, Australia
{t.tran,thao.le,truyen.tran}@deakin.edu.au

Abstract

Referring-based Video Object Segmentation is a multimodal problem that requires producing fine-grained segmentation results guided by external cues. Traditional approaches to this task typically involve training specialized models, which come with high computational complexity and manual annotation effort. Recent advances in vision-language foundation models open a promising direction toward training-free approaches. Several studies have explored leveraging these general-purpose models for fine-grained segmentation, achieving performance comparable to that of fully supervised, task-specific models. However, existing methods rely on fixed pipelines that lack the flexibility needed to adapt to the dynamic nature of the task. To address this limitation, we propose ***M****ulti-****M****odal Agent (*M^2-Agent*), a novel agentic system designed to solve this task in a more flexible and adaptive manner. Specifically,* M^2-Agent *leverages the reasoning capabilities of large language models (LLMs) to generate dynamic workflows tailored to each input. This adaptive procedure iteratively interacts with a set of specialized tools designed for low-level tasks across different modalities to identify the target object described by the multimodal cues. Our agentic approach demonstrates clear improvements over prior methods on two multimodal-conditioned VOS tasks: RVOS and Ref-AVS.*

1. Introduction

Video Object Segmentation (VOS) aims at segmenting specific objects across all frames of a video. This is a fundamental problem in video understanding with a wide range of real-world applications. Recent efforts have extended this task by incorporating additional modalities, including language in Referring Video Object Segmentation [2] (RVOS) and audio in Reference Audio-Visual Segmentation [13] (Ref-AVS), to enhance its relevance to real-world scenarios. These task extensions relax the constraint of pre-defined object categories, allowing the recognition of arbitrary instances described through multimodal cues. Despite improvement in real-world utility, these tasks introduce additional complexity in handling multimodal information.

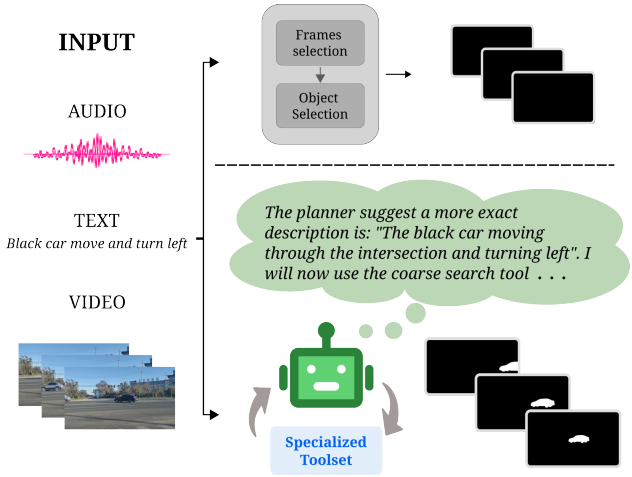


Figure 1. A comparison between the proposed *M^2-Agent* (bottom right) and baseline AL-Ref-SAM 2 [6] (top right). The baseline follows a fixed two-step pipeline composed of frame selection and object selection, regardless of scenario complexity. In contrast, our proposed approach leverages the reasoning capabilities of LLMs to construct a case-specific reasoning chain. *M^2-Agent* dynamically decomposes the task into a variable number of steps and iteratively interacts with a specialized toolset in each step, resulting in optimal results for each situation.

Traditional approaches rely on fully supervised models, requiring heavy computation and pixel-level annotations. Also, while achieving strong benchmark performance, these task-specific models often struggle to generalize beyond their training domains to diverse real-world scenarios. The emergence of large multimodal foundation models has enabled training-free approaches to multimodal VOS. Ren et al. [10] pioneer this direction with a pipeline using GroundingDINO [8] for first-frame object detection and SAM2 [9] for video segmentation. However, GroundingDINO, designed for spatial grounding, struggles with spatiotemporal references. Also, the assumption that the target object always appears in the first frame does not hold in real-world scenarios. AL-Ref-SAM 2 [6] partially addresses these is-

sues by leveraging GPT's spatiotemporal reasoning to construct a two-stage pipeline: it first selects an optimal pivot frame from a fixed set, then identifies the bounding box that best captures the target instance. While AL-Ref-SAM 2 improves upon earlier methods, its fixed pipeline remains suboptimal for diverse scenarios. As shown in Figure 1, the target event only appears in a specific short duration over entire long video and cannot be detected from fixed candidate frame list, resulting in incorrect results. Also, rigid pipeline in AL-Ref-SAM 2 is single-pass without feedback loop, so failure in one step leads to overall failure.

To bridge this gap, we propose a flexible approach for VOS tasks that involves query-specific planning and multi-step reasoning through interaction with specialized vision tools. The approach begins with a Planner module that generates a tailored execution plan for each input. This plan then initiates a multi-step reasoning process, where a large language model dynamically interacts with a suite of audio-vision tools to resolve the target object. Rather than following a fixed, hand-crafted pipeline, the system adapts its behavior on the fly based on contextual situation. This adaptive framework forms **M**ulti-**M**odal Agent (M^2-*Agent*), a multimodal AI agent capable of perceiving diverse inputs to solve the VOS problem. We validate the effectiveness of M^2-*Agent* on two tasks including RVOS and Ref-AVS. The superior performance over prior methods on these tasks highlights the effectiveness of the proposed agentic workflow for this multimodal problem.

2. Method

2.1. Planner

Real-world visual understanding tasks present highly variable scenarios and contexts, requiring AI systems to dynamically adapt with case-specific solutions. In M^2-*Agent*, a text-based LLM acts as a Planner, generating tailored plans based on the video content and input query. Specifically, we extract a detailed textual narrative that captures the object's progression throughout the video. This narrative is generated once using a vision-language foundation model, referred to as the Narrative Extractor. As shown in Figure 2, the Planner leverages the narrative to infer additional context beyond the given reference. The reference may indicate category level (e.g., "the horses"), or an instance appearing throughout the entire the video (e.g., "brown horse standing still"), or one appearing only in a specific segment (e.g., "small horse moving around"). Based on this, it selects an appropriate plan, enabling both efficient resource usage and case-specific optimization. (Refer to supplementary material for additional analysis demonstrating its advantage over rigid baselines.)

Another advantage of our Planner is its ability to refine and consolidate vague or underspecified queries. For example, an initial input like "the horse standing still" can be reformulated using contextual cues from the narrative into a more precise description, such as "the adult brown horse standing still." While this query consolidation task is trivial for LLMs given their advanced text comprehension capabilities, it greatly facilitates the precision of the steps that follow.

2.2. Multi-step Reasoning Process

The initial plan provides a starting point for reasoning and the actual solution is refined through tool interactions in our agentic system. At the start, M^2-*Agent* reads the system prompt, containing task instructions and tool descriptions following specific format specifications (see Supp.). We then apply in-context learning [14] using a few reasoning examples to guide the agent before it independently tackles real queries.

The M^2-*Agent* consists of an LLM and a set of specialized tools. The Multi-step Reasoning Process operates as an iterative interaction between LLM and tools. At each step, we follow ReAct [15] to prompt the decision-making process, progressing through three distinct phases: THOUGHT, ACTION, and OBSERVATION. In the THOUGHT phase, the agent determines the next action. If tool use is required, the agent formats input as specified in the system prompt. It then enters the ACTION phase, generating Python code for tool calling, which is executed externally in a separated process. The LLM relies only on the textual responses returned by the tools for its reasoning and does not intervene in the code execution itself. If the tool fails or returns an error, log messages are fed back to the LLM, allowing it to adjust the plan accordingly. This mechanism offers greater flexibility and robustness compared to previous non-agent approaches. The M^2-*Agent* receives the tool response in OBSERVATION phase and incorporates it in the next THOUGHT phase. The interaction continues until reaching a solution or the step limit. If the maximum number of steps is reached, the system defaults to the intuitive strategy: performing frame selection followed by box selection. An example of the Multi-step Reasoning Process is shown in the upper part of Figure 3.

2.3. Specialized Toolset

We develop a minimal yet sufficient toolset tailored for the task. Each tool executes code deterministically, with input and output formats strictly adhering to a predefined structure. Specifically, the toolset includes:

Audio Processing: This tool identifies the source of sounds from a given audio input. We use the BEATs model [1] to classify the input audio into sound sources. The five highest-scoring categories are selected to form a candidate list.

Temporal Search: The search tool identifies the video

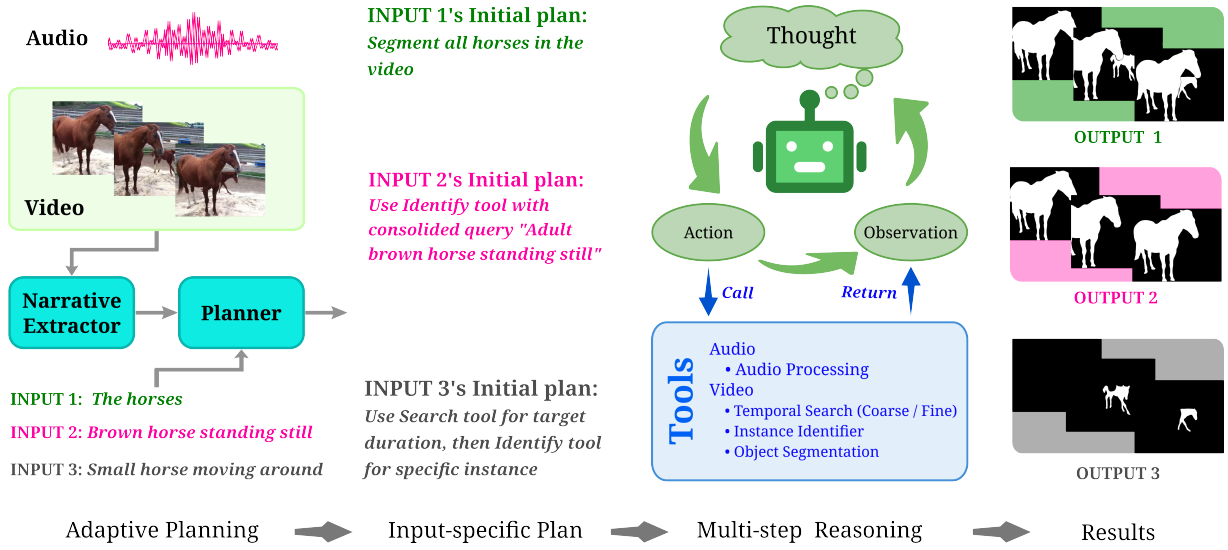


Figure 2. Given a video and multimodal cues including a textual description and optionally audio, the Multi-Modal Agent (M^2-*Agent*) performs case-specific planning and reasoning to process the inputs. The approach begins by extracting a textual narrative from the video. Based on the narrative, the Planner generates an initial tailored plan for each input query, which initiates a multi-step reasoning process to identify the referred instance. This reasoning path is dynamically constructed on the fly through interactions between the LLM and a predefined set of tools. Audio cues are handled by an audio processing tool to identify sound sources, while the video is analyzed by specialized vision tools to locate the target object, resulting in an optimal reasoning trace tailored to each case.

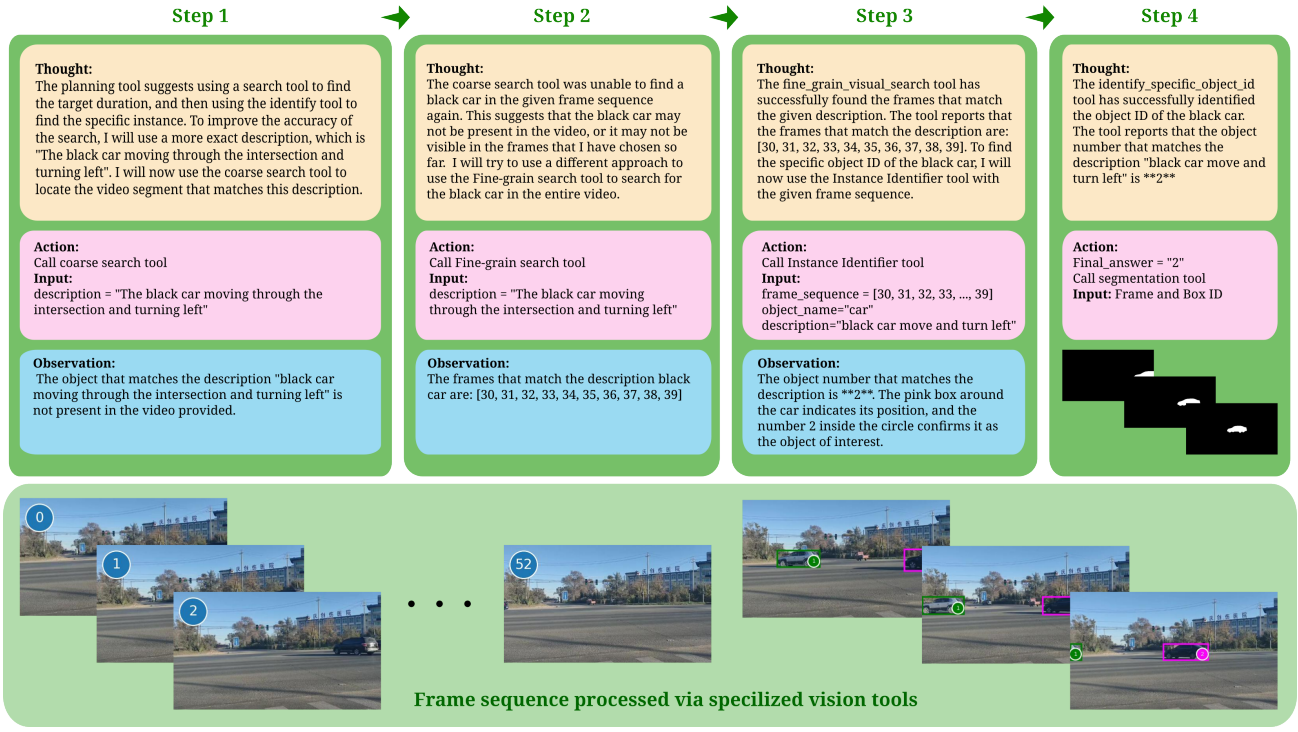


Figure 3. *Top:* A multi-step reasoning process for the query *"Black car move and turn left"*. The agent interprets multimodal cues by sequentially invoking tools. Each step consists of three phases: Thought, Action, and Observation. The process continues until a satisfactory solution is reached. *Bottom:* Specialized tools overlay the frame sequence with temporal indices and object annotations to construct effective prompts for the vision-language model.

segment that best matches a given description. It takes a sequence of video frames and a textual query as input. To convey temporal order, each frame is overlaid with its frame number (see lower part of Figure 3). The sampled frames are passed to the Qwen2.5VL model along with a prompt to identify the frame sequence that best matches the description, enabling retrieval of the most relevant segment. We implement two variants: coarse search, which sparsely samples frames across the entire video to capture long-duration events, and fine-grained search, which divides the video into chunks and densely samples frames to detect short-duration actions that coarse search may miss.

Instance Identifier: The Identifier tool is designed to locate the target instance's ID based on a given frame sequence, textual description, and object category. Unlike the search tool, its objective is instance-level identification. This tool first detects all candidates matching the specified category and overlays bounding boxes and IDs on the frames (see lower part of FigureÂ 3). The annotated sequence is fed to Qwen2.5VL with a textual prompt to identify the object ID that best matches the description, enabling selection of the most relevant instance.

Video Object Segmentation and Tracking: Given the frame index and object ID derived from the Multi-step Reasoning Process, we use SAM-2 [9] to segment the target instance and propagate the mask forward and backward through entire video.

3. Experiments

3.1. Experiment setup

Task: We evaluate *M^2-Agent* on two multimodal-conditioned tasks: the RVOS task using the MeViS dataset [2], and the Ref-AVS task [13], which additionally includes audio modality in addition to textual description. Performance is measured using the standard $\mathcal{J}\&\mathcal{F}$ metric.

Components of *M^2-Agent* system: We use Llama 3.3 70B Instruct [4] for both reasoning and action generation. The toolset consists of specialized foundation models: BEATs [1] for audio classification, GroundingDINO [10] for text-guided object detection, SAM-2 [9] for video segmentation, and Qwen2.5VL [11] used in the Search and Identifier tools.

The LLM is configured to halt generation upon reaching the Observation tag, allowing it to wait for the result from tool execution. Once the tool completes its operation, the response is appended to the prompt, allowing the LLM to continue the reasoning process.

3.2. Main results

Table 1 compares *M^2-Agent* with other methods on the MeViS dataset. G-L + SAM 2 [16] and Grounded-SAM 2 [10] generate an initial mask for the first frame and use it to prompt SAM 2. AL-Ref-SAM 2 improves this by adding a two-step pipeline of frame and box selection, but its rigid structure limits adaptability. In contrast, *M^2-Agent* distinguishes itself with greater flexibility, enabling case-specific planning and dynamic reasoning. Our system constructs tailored workflows that adapt to each scenario, yielding contextually optimized traces. Results show that *M^2-Agent* outperforms other training-free methods and is comparable to the state-of-the-art fully supervised DsHmp [5].

Method	Reference	mIOU
Fully-supervised Fine-tuning		
DsHmp [5]	CVPR24	46.4
Training-free		
G-L + SAM 2 [16]	CVPR23	23.7
G-L (SAM) + SAM 2 [16]	CVPR23	30.5
Grounded-SAM 2 [10]	ICCV23 (Demo)	38.9
AL-Ref-SAM 2 [6]	AAAI25	42.8
M^2-Agent	_	46.1

Table 1. Comparison with other methods on the MeViS dataset.

Method	Reference	mIOU
R2VOS [7]	ICCV23	25.01
AVSegFormer [3]	AAAI24	33.47
GAVS [12]	AAAI24	28.93
EEMC [13]	ECCV24	34.20
M^2-Agent	_	36.26

Table 2. Comparison with other methods on the Ref-AVS dataset.

We compare *M^2-Agent* with state-of-the-art methods on Ref-AVS in Table 2. R2VOS [7] targets RVOS task (video + text), while AVSegFormer [3] and GAVS [12] focus on AVS (video + audio). Wang et al. [13] adapts these models to handle both text and audio and introduce EEMC, the first method for Ref-AVS. All of those methods build multimodal representations through supervised training. In contrast, *M^2-Agent* is training-free, interleaving reasoning and action generation via LLM-guided workflows. The clear performance gains demonstrate the strength of our agentic approach for multimodal understanding.

4. Conclusion

In this work, we propose *M^2-Agent*, a training-free agent for multimodal-conditioned video object segmentation. Unlike conventional methods, *M^2-Agent* leverages LLMs to generate step-by-step, case-specific solutions through iterative reasoning. At each step, it invokes specialized tools for low-level tasks, enabling effective and flexible execution. The improved performance over prior approaches highlights the strength of our agentic approach.

References

[1] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 5178–5193. PMLR, 2023. 2, 4

[2] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 1, 4

[3] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12155–12163, 2024. 4

[4] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 4

[5] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*, 2024. 4

[6] Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3715–3723, 2025. 1, 4

[7] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22236–22245, 2023. 4

[8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 1

[9] 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, 4

[10] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 1, 4

[11] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4

[12] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5669–5677, 2024. 4

[13] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. In *European Conference on Computer Vision*, pages 196–213. Springer, 2024. 1, 4

[14] Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36:36637–36651, 2023. 2

[15] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[16] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19456–19465, 2023. 4