面向多模态引导的视频对象分割的主动型人工智能

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

Abstract

基于指称的视频对象分割是一个多模态问题，需要通过外部线索来生成细粒度的分割结果。传统方法通常涉及训练专门的模型，这伴随着高计算复杂度和人工标注的工作。近期在视觉-语言基础模型方面的进展为无训练的方法开辟了一个有前景的方向。几项研究已探索利用这些通用模型进行细粒度分割，取得了可以与完全监督的特定任务模型相媲美的性能。然而，现有方法依赖于固定的流程，缺乏适应任务动态变化所需的灵活性。为了解决这一限制，我们提出 M 多 M 态代理（M ^{2} -Agent），这是一个旨在以更灵活和自适应的方式解决这一任务的新颖代理系统。具体而言，M ^{2} -Agent 利用大型语言模型（LLMs）的推理能力，生成针对每个输入定制的动态工作流程。这个自适应过程与一组专门为不同模态的低层次任务设计的工具进行迭代交互，以识别由多模态线索描述的目标对象。我们的代理方法在两个多模态条件的 VOS 任务上展示了明显的改进：RVOS 和 Ref-AVS。

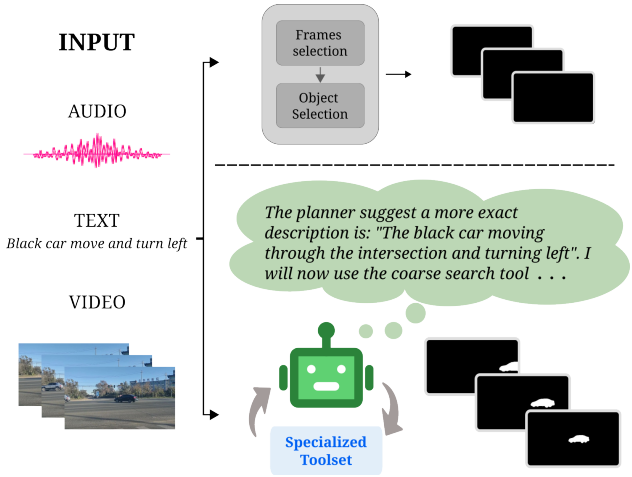

Figure 1. 对比我们提出的 M ^{2} -Agent （右下）和基线 AL-Ref-SAM 2 [6] （右上）。该基线遵循一个固定的两步流程，由帧选择和对象选择组成，无论场景复杂性如何。相比之下，我们提出的方法利用大型语言模型的推理能力来构建特定案例的推理链。M ^{2} -Agent 动态地将任务分解为多个步骤，并在每个步骤中迭代地与专门的工具集进行交互，从而为每种情况产生最佳结果。

1. 介绍

视频对象分割（Video Object Segmentation，简称 VOS）旨在对视频的所有帧中的特定对象进行分割。这是视频理解中的一个基本问题，具有广泛的实际应用。最近的研究努力通过结合额外的模态扩展了这一任务，包括在指定视频对象分割 [2]（RVOS）中结合语言，以及在参考音频-视觉分割 [13]（Ref-AVS）中结合音频，以增强其与现实场景的相关性。这些任务扩展放宽了预定义对象类别的限制，允许通过多模态提示识别任意实例。尽管在实际应用效用上有所提升，这些任务在处理多模态信息时引入了额外的复杂性。

传统方法依赖于完全监督的模型，要求大量计算和像素级标注。此外，虽然在基准性能上表现强劲，但这些特定任务的模型往往难以超越其训练领域以适应多样化的真实场景。大型多模态基础模型的出现使得无需训练的方法应用于多模态视频对象分割成为可能。Ren 等人（[10]）开创了这一方向，使用 GroundingDINO（[8]）进行首帧目标检测，并采用 SAM2（[9]）进行视频分割。然而，专为空间定位设计的 GroundingDINO 在时空参考方面遇到了困难。此外，目标对象总是出现在第一帧中的假设在真实场景中并不成立。AL-Ref-SAM 2（[6]）通过利用 GPT 的时空推理来构建一个两阶段管道，部分解决了这些问题：它首先从一个固定的集合中选择一个最佳的枢轴帧，然后确定最佳捕获目标实例的边界框。虽然 AL-Ref-SAM 2 相对于早期方法有所改进，但其固定的管道对于多样化场景仍然次优秀。如图 1 所示，目标事件仅出现在整个长视频中的一个特定短时间段，并且无法从固定的候选帧列表中检测到，从而导致错误的结果。此外，AL-Ref-SAM 2 中的刚性管道是单程的，没有反馈循环，因此一步失败会导致整体失败。

为了弥合这一差距，我们为 VOS 任务提出了一种灵活的方法，该方法涉及通过与专门的视觉工具交互进行查询特定的规划和多步推理。该方法从一个计划模块开始，为每个输入生成一个定制的执行计划。然后，该计划启动一个多步推理过程，在此过程中，一个大型语言模型动态地与一套音频-视觉工具交互，以解析目标对象。系统不是遵循固定的、手工制作的流程，而

www.xueshuxiangzi.com

是根据上下文情况动态适应其行为。这种自适应框架构成了 M 蜂窝体模态社交代理（ M ^{2} -Agent ），一个能够感知多种输入以解决 VOS 问题的多模态 AI 代理。我们在包括 RVOS 和 Ref-AVS 两项任务上验证了 M ^{2} -Agent 的有效性。在这些任务上，优于先前方法的表现突显了为这一多模态问题提出的代理工作流程的有效性。

2. 方法

2.1. 规划器

现实世界中的视觉理解任务呈现出高度多变的场景和背景，需要 AI 系统动态地适应特定情况的解决方案。在 M ^{2} -Agent 中，一个基于文本的大型语言模型（LLM）作为规划器，根据视频内容和输入查询生成定制的计划。具体来说，我们提取了一个详细的文本叙述，它捕捉了对象在整个视频中的进展。这个叙述是通过一个称为叙述提取器的视觉语言基础模型生成一次。正如图 2 所示，规划器利用叙述来推断给定参考之外的额外背景信息。参考可能指出类别级别（例如，“马”），或者是出现在整个视频中的实例（例如，“棕色马静止站立 ”），或者只出现在特定片段中的实例（例如，“小马正在四处移动 ”）。基于此，它选择合适的计划，从而实现高效的资源利用和特定情况的优化。（请参阅补充材料以获得更多分析，展示其相比于固定基线的优势。）

我们的规划器的另一个优点是能够完善和巩固模糊或未明确指定的查询。例如，像 “那匹静止的马 ” 这样的初始输入可以使用叙述中的上下文提示重新制定为更精确的描述，比如 “那匹静止的成年棕色马 ” 。虽然鉴于大型语言模型的先进文本理解能力，这种查询整合任务对于它们来说是微不足道的，但它极大地促进了后续步骤的精确性。

2.2. 多步推理过程

初始计划为推理提供了一个起点，并且通过在我们的代理系统中的工具交互对实际解决方案进行细化。在开始时， M ^{2} -Agent 读取系统提示，其中包含任务说明和工具描述，并遵循特定格式规范（见附录）。然后，我们应用上下文学习 [14] ，使用几个推理示例来指导代理在独立解决真实查询之前的行为。

M ^{2} -Agent 由一个大型语言模型和一组专用工具组成。多步推理过程作为大型语言模型与工具之间的迭代交互运行。在每一步中，我们遵循 ReAct [15] 来提示决策过程，通过三个不同的阶段来推进：思考、行动和观察。在思考阶段，代理决定下一个行动。如果需要使用工具，代理会按照系统提示的格式来组织输入。然后进入行动阶段，生成用于工具调用的 Python 代码，并在一个独立的进程中执行。大型语言模型仅依赖于工具返回的文本响应进行推理，并不干涉代码执行本身。如果工具失败或返回错误，日志信息会反馈给大型语言模型，使其能够相应地调整计划。与以前的非代理方法相比，这种机制提供了更大的灵活性和鲁棒性。 M ^{2} -Agent 在观察阶段接收工具响应，并将其纳入下一个思考阶段。交互一直进行到达到解决方案或步骤限制为止。如果达到最大步数，系统将默认采用直观策略：先进行框架选择，然后进行盒子选择。图 3 的上半部分显示了多步推理过程的一个例子。

2.3. 专业工具集

我们开发了一套针对任务量身定制的简约而充分的工具集。每个工具都能以确定性的方式执行代码，其输入和输出格式严格遵循预定义的结构。具体来说，该工具集包括：

音频处理：该工具识别来自给定音频输入的声音源。我们使用 BEATs 模型 [1] 将输入音频分类为声源。选出得分最高的五个类别，形成候选列表。

时间搜索：该搜索工具识别与给定描述最匹配的视频片段。它将视频帧序列和文本查询作为输入。为了传达时间顺序，每个帧都叠加了其帧编号（见图 3 的下部）。采样帧与提示一起传递给 Qwen2.5VL 模型，以识别与描述最匹配的帧序列，从而实现最相关片段的检索。我们实施了两种变体：粗略搜索，稀疏采样整个视频的帧以捕捉长时间事件；细粒度搜索，将视频划分为块并密集采样帧，以检测粗略搜索可能错过的短时动作。

实例标识符：标识符工具设计用于根据给定的帧序列、文本描述和对象类别来定位目标实例 ’的 ID。与搜索工具不同，其目标是实例级别的识别。该工具首先检测所有符合指定类别的候选对象，并在帧上覆盖边界框和 ID（见图 3 的下半部分）。带有注释的序列被输入到 Qwen2.5VL 中，并通过文本提示识别最符合描述的对象 ID，从而选择最相关的实例。

视频对象分割与跟踪：给定从多步骤推理过程得到的帧索引和对象 ID，我们使用 SAM-2 [9] 来分割目标实例，并在整个视频中向前和向后传播掩膜。

3. 实验

3.1. 实验设置

任务：我们在两个多模态条件任务上评估 M ^{2} -Agent ：使用 MeViS 数据集的 RVOS 任务 [2] ，以及 Ref-AVS 任务 [13] ，该任务除了文本描述外还包括音频模态。性能使用标准 $\mathcal{J}$ & $\mathcal{F}$ 指标测量。

M ^{2} -Agent 系统的组件：我们使用 Llama 3.3 70B Instruct [4] 进行推理和动作生成。工具集由专门的基础模型组成：用于音频分类的 BEATs [1] ，用于文本引导对象检测的 GroundingDINO [10] ，用于视频分割的 SAM-2 [9] ，以及用于搜索和标识工具的 Qwen2.5VL [11] 。

LLM 被配置为在达到 Observation 标签时停止生成，以使其能够等待工具执行的结果。一旦工具完成操作，响应会被附加到提示中，从而允许 LLM 继续推理过程。

www.xueshuxiangzi.com

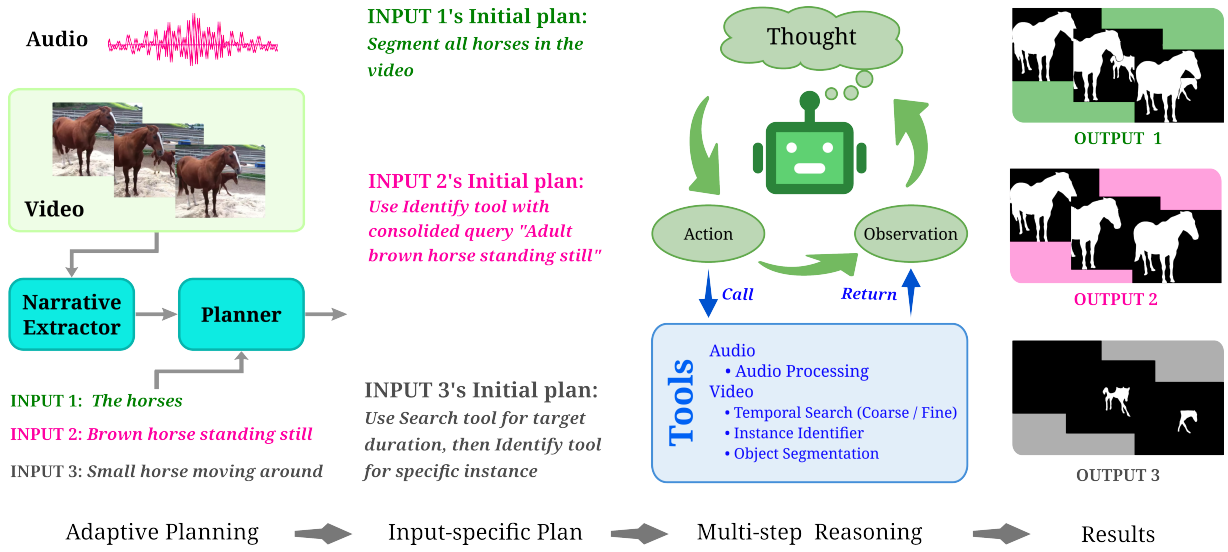

Figure 2. 给定一个视频及包括文本描述和可选音频在内的多模态线索，多模态代理（ M ^{2} -Agent ）执行案例特定的规划和推理以处理输入。该方法首先从视频中提取文本叙述。基于该叙述，规划器为每个输入查询生成一个初步定制的计划，该计划启动多步推理过程以识别指定的实例。通过 LLM 和一组预定义工具之间的交互，即时动态构建推理路径。音频线索由音频处理工具处理以识别声音来源，而视频则由专门的视觉工具分析以定位目标对象，从而制定适合每个案例的最优推理过程。

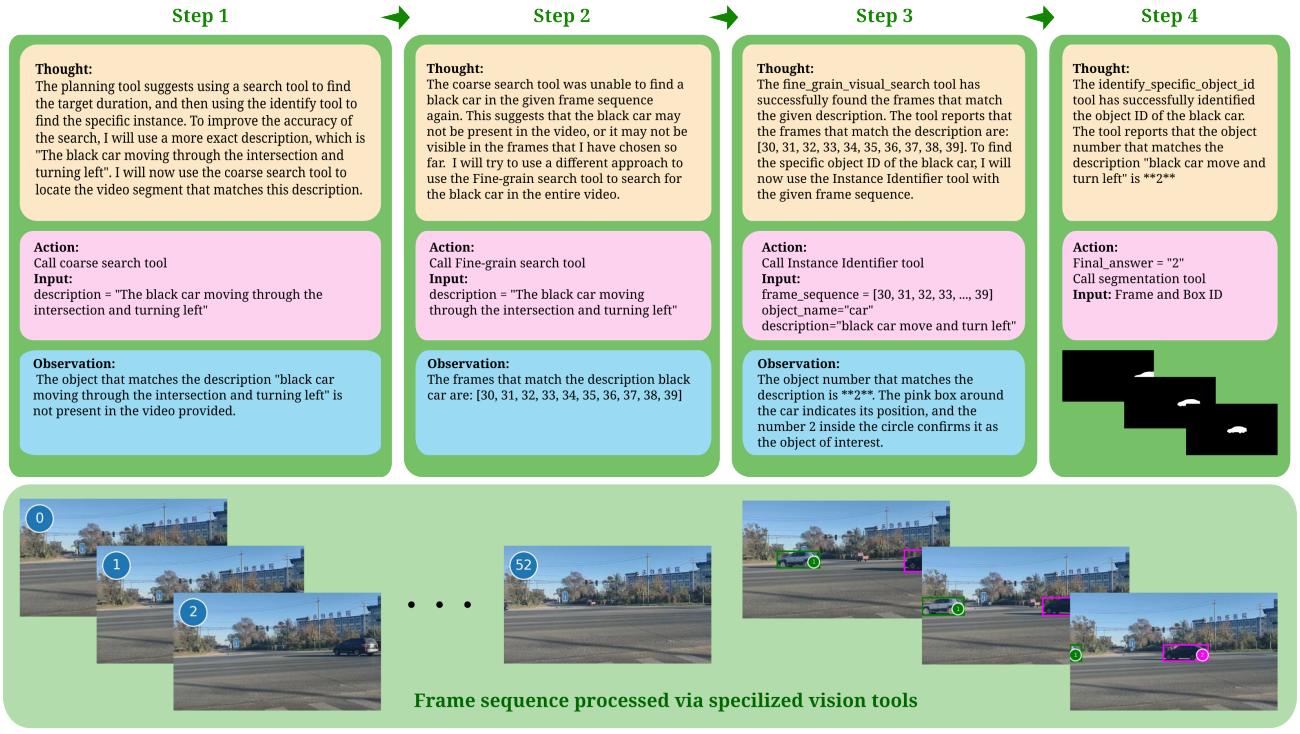

Figure 3. 上图：针对查询 “黑车移动并左转 ”的多步推理过程。智能体通过依次调用工具来解释多模态线索。每一步由三个阶段组成：思考、行动和观察。该过程持续进行，直到获得令人满意的解决方案。下图：专业工具在帧序列上叠加时间索引和对象注释，以构建视觉语言模型的有效提示。

3.2. 主要结果

表 1 将 M ^{2} -Agent 与其他方法在 MeViS 数据集上的表现进行了比较。G-L + SAM 2 [16] 和 Grounded-

www.xueshuxiangzi.com

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. 在 MeViS 数据集上的方法比较。

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. 在 Ref-AVS 数据集上与其他方法的比较。

SAM 2 [10] 为首帧生成初始掩码并使用它来提示 SAM 2。AL-Ref-SAM 2 通过新增帧和框选择的两步管道来改进此方法，但其固定结构限制了适应性。相比之下， M ^{2} -Agent 以更大的灵活性脱颖而出，实现特定案例规划和动态推理。我们的系统构建了适合每种情境的工作流程，产生上下文优化的轨迹。结果表明， M ^{2} -Agent 的表现优于其他无需训练的方法，并且可与最新的全监督 DsHmp [5] 相媲美。

我们在表格 2 中将 M ^{2} -Agent 与目前最先进的方法在 Ref-AVS 上进行了比较。R2VOS [7] 目标是处理 RVOS 任务（视频 + 文本），而 AVSegFormer [3] 和 GAVS [12] 则专注于 AVS（视频 + 音频）。Wang 等人 [13] 对这些模型进行了适配，以同时处理文本和音频，并推出了 EEMC，这是第一个用于 Ref-AVS 的方法。所有这些方法都是通过监督训练来构建多模态表示。相反， M ^{2} -Agent 是无需训练，通过 LLM 指导的工作流程交替进行推理和行动生成。显著的性能提高展示了我们代理方法在多模态理解中的优势。

4. 结论

在这项工作中，我们提出了 M ^{2} -Agent ，一种用于多模态条件视频对象分割的免训练代理。与传统方法不同， M ^{2} -Agent 利用 LLMs 通过迭代推理生成逐步、特定案例的解决方案。在每一步中，它调用专门的工具来处理低级任务，从而实现有效且灵活的执行。与之前的方法相比，更好的性能突出了我们代理方法的优势。

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] 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.

[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] 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.

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

[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.

[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.

[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.

[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.

[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.

[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.

www.xueshuxiangzi.com

[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.

[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.

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

[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.

[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.

www.xueshuxiangzi.com