LLAVA-Plus (arXiv 202311)

Motivation: 现有基于Prompt的工具调用模型适应性不强，且不够强大，无法让多模态Agent始终准确地选择和激活适当的工具并组合其结果。

整体思路：LLAVA-Plus维护这一个技能存储库，其中包含各种视觉和视觉语言模型（工具），它可以根据用户的多模态输入激活相关的工具，以动态组合其执行结果来实现许多现实世界的任务。

LLAVA-PLUS提出了一种**模块化的系统架构**，整个过程分为**四步**：

(1) 人类体统任务指令和对应的图片；

(2) LMM助手分析人任务指令和对应的图片，并输出需要使用的工具（从工具库中选择工具并将适当的提示写入工具参数）；

(3) 执行工具，将执行结果返回给助手；

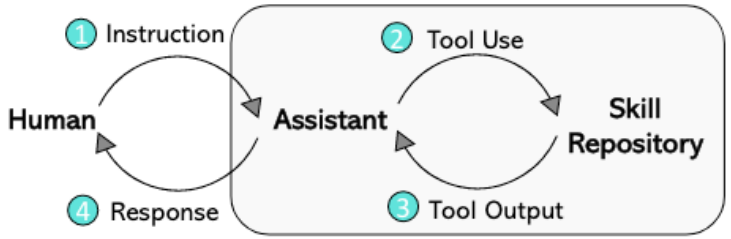
(4) 助手整合工具的执行结果，并将最终结果输出给人类。

交互过程可以表示为：

$$\text{Human}: \mathbf{I}_q <\backslash n> \mathbf{X}_q \text{<STOP>} \; \text{Assistant}: \mathbf{X}_{\text{skill\_use}} \text{<STOP>}$$

$$\text{Human}: \mathbf{X}_{\text{skill\_result}} \text{<STOP>} \; \text{Assistant}: \mathbf{X}_{\text{anwser}} \text{<STOP>}$$

其中，loss只使用绿色的序列来计算。

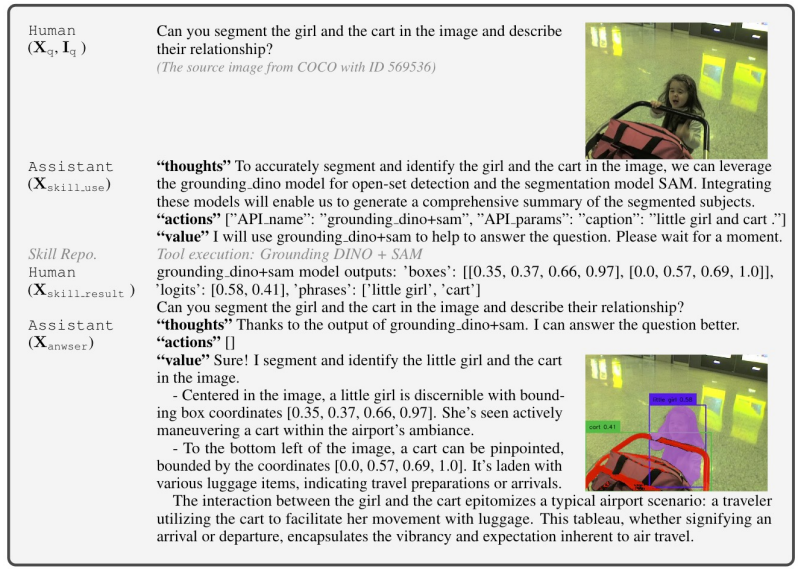


受**ReAct**启发，**将LMM预测格式统一**，这个格式包含三个部分：

（1）**Thought**: 表示推理过程的文本序列，它根据用户指令决定是否需要工具库以及使用哪些工具；

（2）**Action**: Thought中需要调用的工具列表，json格式，两个字段：API_name和API_params；

（3）**Value**: 模型根据工具的调用结果和对话历史信息生成的回复。



多模态工具使用指令数据生成

	Skills		Tools	Source	Size
Individual Skills	Understanding	Detection/Grounding	G-DINO (Liu et al., 2023f)	COCO	13783
		Semantic Segmentation	OpenSeeD (Zhang et al., 2023a)	COCO	5989
		Instance Segmentation	G-DINO+SAM	COCO	5228
		Caption + Grounding	BLIP2+G-DINO	COCO	4037
		Tagging + Grounding	RAM+G-DINO	COCO	4439
		Caption	BLIP2 (Li et al., 2023d)	COCO	4064
		Tagging	RAM (Zhang et al., 2023c)	COCO	6045
		OCR	EasyOCR (JaidedAI, 2023)	Hiertext	6528
	External Knowledge	Retrieval	CLIP Retrieval (Radford et al., 2021)	InfoSeek	4087
	Generation	Image Generation	Stable Diffusion (Rombach et al., 2021)	JourneyDB	4694
		Image Editing	Instruct P2P (Brooks et al., 2023)	Instruct P2P	6981
	Visual Prompt	Interactive Segmentation	SAM (Kirillov et al., 2023)	COCO	5601
		Multi-granularity	Semantic SAM (Li et al., 2023c)	COCO	5601
		Example Based Segmentation	SEEM (Zou et al., 2023b)	COCO	5601
Composed Skills	Mix of Detection, Segmentation, Tagging, Caption		G-DINO, SAM, BLIP2, RAM	COCO	37,431
	Interactive Segmentation + Inpainting		SAM + Stable Diffusion	COCO	3063
	Semantic Segmentation + Generation		OpenSeeD + ControlNet (Zhang et al., 2023f)	COCO	5989
	Image Generation + Social Media Post		Stable Diffusion	JourneyDB	4694
	Image Editing + Social Media Post		Instruct P2P (Brooks et al., 2023)	Instruct P2P	5924

核心技能：视觉理解

仅图像的工具：(i) 使用GPT4生成需要工具才能回答的指令；(ii) Thouht和value从预设定的Response中选择并重写；(iii) Xskill_result由工具结果和原始问题拼接而成；(iv) Xanswer中的Thought与之前方法相似，Action被置为空，value由GPT4生成。

带有附加参数的工具：(i) 与上面方法相似，初始Xq包含占位符，随机选择一个类别替换该占位符，并通过重写得到最终的Xq；(ii) 使用GPT4生成问题，手动创建两个种子样本，将他们与图像上下文一起输入GPT4，并要求GPT4基于新的图像上下文生成完整对话。

扩展工具

(1) 外部知识：使用CLIP搜索API检索LIAON

(2) 生成图像：Stable Diffusion&Instruct-Pix2Pix

(3) 视觉提示：SAM&SEEM

(4) 工具组合：(i) 交互式切割与修复；(ii) 语义分割+生成；(iii) 图像生成/编辑 + 社交媒体帖子

	LLaVA-Bench (COCO)				LLaVA-Bench (In-the-Wild)			
	Conv.	Detail	Reasoning	All	Conv.	Detail	Reasoning	All
LLaVA	82.0	69.1	92.6	81.2	42.6	51.9	68.9	57.1
LLaVA (Tools in Test)	56.2	67.9	53.3	59.1	40.7	48.1	51.2	47.5
LLaVA-Plus (All Tools)	81.6	74.5	95.7	**83.9**	65.5	56.8	79.1	**69.5**
LLaVA-Plus (Fly)	76.2	72.2	92.3	80.4	45.2	50.4	72.6	59.1
LLaVA-Plus (Fly) (no thoughts)	76.6	70.4	90.7	79.4	38.8	39.8	59.8	48.7
GPT4Tools	75.3	53.8	86.9	72.1	31.1	27.1	54.1	40.7

	Scene	Identity	Attribute	Location	Counting	Spatial	Interact.	Reason.	Text	Average
LLaVA	59.50	54.29	56.06	42.54	39.35	33.03	43.30	41.39	30.59	44.45
LLaVA (Tools in Test)	67.13	56.85	45.24	47.24	45.69	40.18	60.82	70.09	30.59	51.54
LLaVA-Plus (All Tools)	68.94	56.80	58.89	47.34	48.14	45.21	60.82	71.30	37.65	55.01
LLaVA-Plus (Fly)	68.43	56.47	59.69	45.40	41.68	44.14	59.79	69.49	34.12	53.25

	Grounding	Tagging	Caption	OCR	All
LLaVA	47.1	87.1	77.0	23.6	58.7
LLaVA (Tools in Test)	41.7	48.5	72.0	31.9	48.5
LLaVA-Plus (All Tools)	89.3	94.4	96.7	48.8	**82.3**
LLaVA-Plus (Fly)	88.6	88.9	90.2	38.4	76.5
Bard (0730)	36.5	105.3	103.3	60.0	76.3
Bing Chat (0730)	56.0	84.0	96.0	44.8	70.2
MM-REACT	30.2	94.7	103.8	77.3	76.5
All Tools + GPT4	77.5	95.6	95.2	39.3	76.9

通过实验对比证明在同量级模型中取得了较优性能。