LLaVA-Plus: Learning to Use Tools for Creating Multimodal Agents

Shilong Liu♠∗, **Hao Cheng**♣, **Haotian Liu**♢∗, **Hao Zhang**♡∗, **Feng Li**♡∗, **Tianhe Ren**▶, **Xueyan Zou**♢∗, **Jianwei Yang**♣, **Hang Su**♠, **Jun Zhu**♠, **Lei Zhang**▶, **Jianfeng Gao**♣, **Chunyuan Li**♣ ⚑

♠Dept. of Comp. Sci. & Tech., Institute for AI, BNRist, Tsinghua University
♣Microsoft Research, Redmond
♢University of Wisconsin-Madison ♡HKUST ▶IDEA Research
∗Work performed during an internship at Microsoft ⚑ Project Lead
https://llava-vl.github.io/llava-plus/

Abstract. This paper presents LLaVA-Plus (**L**arge **L**anguage **a**nd **V**ision **A**ssistants that **P**lug and **L**earn to **U**se **S**kills), a general-purpose multimodal assistant trained using an end-to-end approach that systematically expands the capabilities of large multimodal models (LMMs). LLaVA-Plus maintains a skill repository that contains a wide range of vision and vision-language pre-trained models (tools), and is able to activate relevant tools, given users' multimodal inputs, to compose their execution results on the fly to fulfill many real-world tasks. To acquire the ability of using tools, LLaVA-Plus is trained on multimodal instruction-following data that we have curated. The training data covers many tool use examples of visual understanding, generation, external knowledge retrieval and their compositions. Empirical results show that LLaVA-Plus outperforms LLaVA in existing capabilities, and exhibits many new capabilities. Compared with tool-augmented LLMs, LLaVA-Plus is distinct in that the image query is directly grounded in and actively engaged throughout the entire human-AI interaction sessions, significantly improving tool use performance and enabling new scenarios.

Keywords: Multi-modal large language model · Tool use · Visual-language agent

1 Introduction

A long-standing aspiration in artificial intelligence is to develop general-purpose assistants that can effectively follow users' (multimodal) instructions to complete a wide range of real-world tasks [3, 19]. Recently, the community has witnessed a growing interest in developing foundation models with emergent abilities of multimodal understanding and generation in open-world tasks [10, 20]. While the recipes of using Large Language Models (LLMs) such as ChatGPT [30] to develop

Contact: `liusl20@mails.tsinghua.edu.cn`,`{jfgao,chunyl}@microsoft.com`

general-purpose assistants for natural language tasks have been proved effective in many tasks, the recipes of building general-purpose, multimodal assistants for computer vision and vision-language tasks remain to be explored.

Ongoing efforts of developing multimodal agents can be broadly categorized into two classes [19]: (*i*) *End-to-end training with LLMs*, where image-text data and multimodal instruction-following data are collected to continually train LLMs to acquire the ability of processing visual information, resulting in a series of Large Multimodal Models (LMMs). Impressive visual understanding and reasoning performances have been demonstrated by both proprietary models such as Flamingo [2] and multimodal GPT-4 [32], and open-sourced models such as LLaVA [24] and MiniGPT-4 [56]. Although these end-to-end training methods are effective in helping LMMs to gain emergent abilities (such as in-context learning), it remains challenging to develop a unified architecture that can seamlessly incorporate a wide range of skills, such as image segmentation and generation, which are crucial for real-world multimodal applications. (*ii*) *Tool*[1] *chaining with LLMs*, where the prompts are meticulously crafted to enable LLMs (*e.g.* through LangChain [1]) to invoke different tools (*e.g.* pre-trained vision models) to perform desired (sub-)tasks, without the need of additional model training. Some prominent works include VisProg [12], ViperGPT [43], Visual ChatGPT [45], X-GPT [57], and MM-REACT [47]. The strength of these methods is the ability to perform a broad spectrum of visual tasks through the use of (new) tools, which can be incorporated into an AI agent with very low development cost. However, prompting is neither adaptable nor robust enough to allow multimodal agents to always accurately select and activate appropriate tools (from a large and diverse toolset) and compose their results to generate final answers on the fly for real-world multimodal tasks.

In this paper, we present LLaVA-Plus (**L**arge **L**anguage **a**nd **V**ision **A**ssistants that **P**lug and **L**earn to **U**se **S**kills), a general-purpose multimodal assistant that learns to use tools using an end-to-end training approach that systematically expands the capabilities of LMMs via visual instruction tuning. To the best of our knowledge, this is the first attempt reported to combine the strengths of the end-to-end training and tool chaining methods mentioned above. LLaVA-Plus is equipped with a skill repository that contains a wide range of vision and vision-language tools. The design is an embodiment of the "Society of Mind" scheme [29], where each tool is originally designed for a specific skill and by itself is only useful for specific scenarios, but the combinations of these tools lead to emergent abilities that show signs of higher intelligence. For example, LLaVA-Plus is able to construct a new workflow on the fly, given users' multimodal inputs, select and activate relevant tools from the skill repository, and compose their execution results to fulfill many real-world tasks that are unseen during model training.

LLaVA-Plus can be continually improved by incorporating new skills or tools via instruction tuning. Consider a new multimodal tool that has been developed for a specific scenario or skill. We collect pertinent user instructions that request

[1] The term "tools" in this paper is used to describe the APIs or pre-built models that LMM interfaces with.

Fig. 1: Visual illustration of LLaVA-Plus' capabilities enabled by learning to use skills.

this tool and their execution results (or following) to form instruction-following data for tuning. After instruction tuning, LLaVA-Plus expands its abilities as it learns to use this new tool to deal with the tasks that it cannot handle before. LLaVA-Plus also differs from those existing works on teaching LLMs to use tools [35, 46], where visual signals are only used when the multimodal tools are activated. In contrast, LLaVA-Plus uses the raw visual signals through the entire human-AI interaction sessions to improve LMM's ability of planning (determining the most appropriate tools to use for a given task) and reasoning.

In summary, our paper makes the following contributions:

– *New multimodal instruction-following tool use data.* We present a new pipeline for curating vision-language instruction-following data, dedicated for tool use in human-AI interaction sessions, leveraging ChatGPT and GPT-4 as labeling tools.
– *New large multimodal assistant.* We have developed LLaVA-Plus, a general-purpose multimodal assistant that extends LLaVA [24] by incorporating a large and diverse set of external tools that can be selected, composed, and activated on the fly for performing tasks. As shown in Figure 1, LLaVA-Plus significantly extends LMM's capabilities. Our empirical study validates the effectiveness of LLaVA-Plus with consistently improved results on multiple benchmarks, and in particular, new SoTA on VisiT-Bench with a diverse set of real-life tasks.
– *Open-source.* We will release the following assets to the public: the generated multimodal instruction data, the codebase, the LLaVA-Plus checkpoints, and a visual chat demo.

2 Learning to Use Tools with Visual Instruction Tuning

2.1 Preliminaries: Visual Instruction Tuning in LLaVA

Inspired by the impressive performance of multimodal GPT-4 and the open-source LMMs such as LLaVA/MiniGPT-4, the community has witnessed a surge in developing LMMs and the multimodal instruction-following data, following the instruction tuning paradigm [24, 36]. In this paper, we use LLaVA as a running example. But note that the proposed recipe can be easily applied to other LMMs. Starting with a user input image query $\mathbf{I}_{\texttt{q}}$, existing LMMs such as LLaVA typically accept a natural language instruction input $\mathbf{X}_{\texttt{q}}$ from the user, and output a natural language response $\mathbf{X}_{\texttt{answer}}$. Therefore, we can use a unified scheme to represent multimodal instruction-following data as:

$$\texttt{Human}: \mathbf{I}_{\texttt{q}}\ <\backslash n>\ \mathbf{X}_{\texttt{q}}\texttt{<STOP>}\ \texttt{Assistant}: \mathbf{X}_{\texttt{answer}}\texttt{<STOP>}, \tag{1}$$

where `Human` and `Assistant` are special role tokens, $<\backslash n>$ and `<STOP>` are the line break token and sequence end token, respectively. It naturally covers any multimodal tasks that can be formulated as language-image input and language output, ranging from simple visual understanding tasks such as recognition, captioning, and visual question answering (VQA) to complex visual reasoning tasks. Due to its simplicity, the data pipeline is easy to construct and scale. By training a single Transformer-based model with an auto-regressive objective, the resulting LMM enables a seamless human-assistant interaction, proficiently completing many visual tasks in the wild. However, it is limited in flexibility regarding skill expansion and engagement in human-AI interactions.

2.2 LLaVA-Plus

We propose a modularized system architecture that allows an LMM, working as a planner, to learn to use a wide range of skills at scale, and thus facilitating easy expansion of its capabilities and interface. Specifically, we build a skill repository, where the LMM can leverage a broad range of existing vision and vision-language specialist models as tools for their respective skills when needed, to complete various tasks in the wild. The LMMs in most existing multimodal agents typically perform *user-oriented dialogues*, where the LMMs are required to immediately respond to user instructions based solely on the knowledge encoded in model weights, as shown in equation 1 and the left part of Figure 2. In addition to this, the LMM in LLaVA-Plus also performs *skill-oriented dialogues*, where the LMM initiates requests to call appropriate tools from the skill repository, and subsequently aggregate the tool execution results after applying proper skills, as shown in the right part of Figure 2.

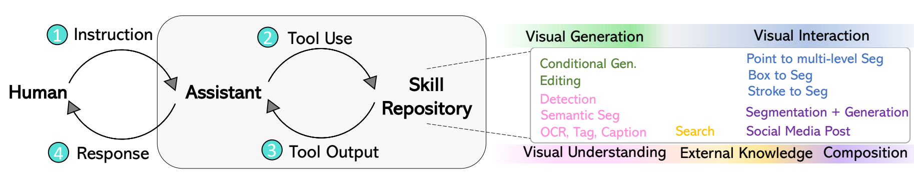


Fig. 2: The four-step LLaVA-Plus pipeline. The skill repository is shown on the right.

A Full Dialogue of LLaVA-Plus. We illustrate how LLaVA-Plus works with a full dialogue session in Figure 2. It proceeds in four steps: ① Humans provide a task instruction $\mathbf{X}_{\texttt{q}}$ related to an image $\mathbf{I}_{\texttt{q}}$. ②The LMM-powered assistant analyzes both $\mathbf{X}_{\mathtt{q}}$ and $\mathbf{I}_{\mathtt{q}}$, and outputs $\mathbf{X}_{\mathtt{skill\_use}}$ that chooses the tool from skill repository and writes the appropriate prompt as the tool argument. ③ By executing the tool, the result $\mathbf{X}_{\mathtt{skill\_result}}$ is returned to the assistant. ④ The assistant aggregates $\mathbf{X}_{\mathtt{skill\_result}}$ with $\mathbf{X}_{\mathtt{q}}$ and $\mathbf{I}_{\mathtt{q}}$, and outputs $\mathbf{X}_{\mathtt{anwser}}$ to humans. The interaction can be represented as:

$$\begin{aligned} &\texttt{Human}: \mathbf{I}_{\mathtt{q}} \ <\backslash n> \ \mathbf{X}_{\mathtt{q}}\texttt{<STOP>}\ \texttt{Assistant}: {\color{green}\mathbf{X}_{\mathtt{skill\_use}}\texttt{<STOP>}} \\ &\texttt{Human}: \mathbf{X}_{\mathtt{skill\_result}}\texttt{<STOP>}\ \texttt{Assistant}: {\color{green}\mathbf{X}_{\mathtt{anwser}}\texttt{<STOP>}} \end{aligned} \quad (2)$$

Compared with equation 1 which is used to train LLaVA, the only newly introduced component for LLaVA-Plus training is the skill-oriented dialogue. Table 1 illustrates one sequence example of calling detection and segmentation skills in human-AI interactions. LLaVA-Plus is trained with an auto-regressive objective on the sequence of equation 2, where only the green sub-sequences (or tokens) are used to compute the loss, and thus the model learns to predict skill use, answers, and when to stop.

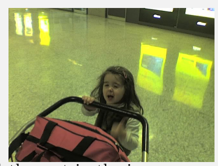
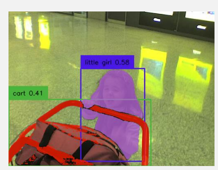
Human ($\mathbf{X}_{\mathtt{q}}$, $\mathbf{I}_{\mathtt{q}}$)	Can you segment the girl and the cart in the image and describe their relationship? *(The source image from COCO with ID 569536)* 
Assistant ($\mathbf{X}_{\mathtt{skill\_use}}$)	**"thoughts"** To accurately segment and identify the girl and the cart in the image, we can leverage the grounding_dino model for open-set detection and the segmentation model SAM. Integrating these models will enable us to generate a comprehensive summary of the segmented subjects. **"actions"** ["API_name": "grounding_dino+sam", "API_params": "caption": "little girl and cart ."] **"value"** I will use grounding_dino+sam to help to answer the question. Please wait for a moment.
Skill Repo.	*Tool execution: Grounding DINO + SAM*
Human ($\mathbf{X}_{\mathtt{skill\_result}}$)	grounding_dino+sam model outputs: 'boxes': [[0.35, 0.37, 0.66, 0.97], [0.0, 0.57, 0.69, 1.0]], 'logits': [0.58, 0.41], 'phrases': ['little girl', 'cart'] Can you segment the girl and the cart in the image and describe their relationship?
Assistant ($\mathbf{X}_{\mathtt{anwser}}$)	**"thoughts"** Thanks to the output of grounding_dino+sam. I can answer the question better. **"actions"** [] **"value"** Sure! I segment and identify the little girl and the cart in the image.  - Centered in the image, a little girl is discernible with bounding box coordinates [0.35, 0.37, 0.66, 0.97]. She's seen actively maneuvering a cart within the airport's ambiance. - To the bottom left of the image, a cart can be pinpointed, bounded by the coordinates [0.0, 0.57, 0.69, 1.0]. It's laden with various luggage items, indicating travel preparations or arrivals. The interaction between the girl and the cart epitomizes a typical airport scenario: a traveler utilizing the cart to facilitate her movement with luggage. This tableau, whether signifying an arrival or departure, encapsulates the vibrancy and expectation inherent to air travel.

Table 1: An example of a LLaVA-Plus workflow that plugs and learns to use the skills of object detection and segmentation, enhanced by a rich region language description. The gray text is not in the training sequence.

Unified Prediction Format from LMMs. Figure 2 shows that the LMM of LLaVA-Plus needs to perform both user-oriented and skill-oriented dialogues. To this end, we use a unified model prediction format to represent dialogues with and without the need of calling the skill repository. Inspired by [48], the format consists of three fields, as illustrated in Table 1: (*i*) `Thought` is a text sequence representing a reasoning process, which determines whether the skill repository is needed to follow the user instruction, and if so, which tools to use. (*ii*) `Action` is a list of function calls for the tools to execute the `thought`. The list is in the JSON format, with each item consisting of two sub-fields: `API_name` to call the tool and `API_params` for the corresponding function arguments if applicable. When `action` is an empty list, no skill is invoked. (*iii*) `Value` is a natural language response that LLaVA-Plus generates by aggregating tool execution results and the human-AI session history. When presented in $\mathbf{X}_{\texttt{skill\_use}}$ of user-oriented dialogues, it is the final response returned to human users. When presented in $\mathbf{X}_{\texttt{anwser}}$ of skill-oriented dialogues, it is a natural language description about tool execution. In the serving stage, we find it important to ensure a good user experience that we only return the content in the `value` field of $\mathbf{X}_{\texttt{anwser}}$ to human users, but hide the entire skill-oriented dialogues unless we need to debug the system.

2.3 Skill Repository: Multimodal Tool Use Instruct Data Generation

The skill repository of LLaVA-Plus consists of multimodal tools of different skills. To allow the LMM to always activate the most appropriate tools to complete a task, the corresponding tool-use multimodal instruction-following data is needed for LMM tuning. In alignment with the LLaVA approach, we input image information into a text-only GPT-4 model, prompting it to generate both questions and responses based on the visual data. Without loss of generality, in this study we want LLaVA-Plus to deal with the scenarios that requires novel skills that LLaVA does not have, *e.g.* the individual skills for visual understanding, generation, and external knowledge retrieval and the compositions of these individual skills, as summarized in Table 2. In what follows, we treat visual understanding skills as core skills and the others as extended skills, and describe the way instruction data is curated.

Core Skills: Understanding Visual understanding skills enable machines to interpret and comprehend visual signals. Existing LMMs have only a limited subset of visual understanding skills, constrained by language inputs and outputs. We expand them to a broader skill set with visual input prompts and visual outputs, including open-set detection and grounding, semantic/instance/interactive segmentation, tagging, captioning, OCR and their compositions, and so on. These understanding skills can be grouped into two categories, depending on whether additional function arguments are required.

Skills with Image-only. The skills without additional function arguments include captioning, tagging, semantic segmentation, caption+grounding, tagging+grounding, and OCR. We have curated training samples for each tool

	Skills		Tools	Source	Size
Individual Skills	Understanding	Detection/Grounding	G-DINO [26]	COCO	13783
		Semantic Segmentation	OpenSeeD [51]	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 [22]	COCO	4064
		Tagging	RAM [54]	COCO	6045
		OCR	EasyOCR [13]	Hiertext	6528
	External Knowledge	Retrieval	CLIP Retrieval [39]	InfoSeek	4087
	Generation	Image Generation	Stable Diffusion [40]	JourneyDB	4694
		Image Editing	Instruct P2P [6]	Instruct P2P	6981
	Visual Prompt	Interactive Segmentation	SAM [14]	COCO	5601
		Multi-granularity	Semantic SAM [21]	COCO	5601
		Example Based Segmentation	SEEM [58]	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 [52]	COCO	5989
	Image Generation + Social Media Post		Stable Diffusion	JourneyDB	4694
	Image Editing + Social Media Post		Instruct P2P [6]	Instruct P2P	5924

Table 2: LLaVA-Plus skill repository and dataset statistics of our created visual instruction-following data for each tool use case. G-DINO indicates Grounding DINO [26]. HierText [27, 28], InfoSeek [8], and JourneyDB [34] are datasets for OCR, external knowledge, and image generation, respectively.

individually. To collect the training samples for a given skill, we fill in the four data variables in equation 2 using different strategies. (*i*) For $\mathbf{X}_{\texttt{q}}$, we use GPT-4 to generate a set of instructions that require the use of tools for proper answers. For each sample, we randomly select a question and rewrite it to enhance data diversity. A rewriting example is shown in Table 9 in Appendix. (*ii*) For $\mathbf{X}_{\texttt{skill\_use}}$, its `thoughts` and `value` are generated by randomly selecting from some preset responses with rewriting. The `actions` is known, so it can be directly assigned. (*iii*) $\mathbf{X}_{\texttt{skill\_result}}$ is generated with a fixed rule: first presenting the tool outputs and then repeating the initial question. (*iv*) For $\mathbf{X}_{\texttt{anwser}}$, its `thoughts` is created in a similar way to `thoughts` in $\mathbf{X}_{\texttt{skill\_use}}$, and `action` is set empty. The `value` of $\mathbf{X}_{\texttt{anwser}}$ is the most important field, as it is the visible response to humans in chat. We feed all previous information, including previous questions, the previous tool outputs, and the context of the image to language-only GPT-4, which then generates responses to form instruction-following data. Inspired by LLaVA, we consider the ground-truth captions, object coordinates, and object categories as image contexts.

Skills with Additional Function Arguments. Visual skills such as object detection and instance segmentation often require humans to provide very specific instructions regarding the concepts of interests. Their instruction-following data is more challenging to create. We use two methods in this study. (*i*) The first method is similar to that in the image-only skill setting, where the initial $\mathbf{X}_{\texttt{q}}$ contains a placeholder `concept`, one or more categories presented in the image are randomly chosen to replace this placeholder, and the final $\mathbf{X}_{\texttt{q}}$ is obtained via rewriting, as shown in Table 9. (*ii*) To allow the LMM to learn more diverse prompts beyond category information, we use GPT-4 to generate questions. Specifically, we manually create two seed samples following the full dialogue in equation 2, send them, together with image contexts, to GPT-4, and ask GPT-4 to generate a full dialogue based on a new image context. An example is shown in Table 10 in Appendix.

Extended Skills The LLaVA-Plus recipe can be applied to any tool to improve the system's capabilities. We demonstrate its versatility by onboarding multimodal tools of different categories. Due to the limited space, we describe the instruction-following data creation process in Section B in Appendix, and summarize the extended skills we have enabled.

- **External Knowledge.** To enable LMMs to use knowledge beyond that encoded in pre-trained model weights, we use the CLIP search API to retrieve external knowledge from LIAON.
- **Generation.** To allow LLaVA-Plus to output images, we use Stable Diffusion (SD) and Instruct-Pix2Pix for image generation and editing, respectively.
- **Visual Prompts.** To better follow human intents, we support various visual prompts for human-AI interaction, such as user-drawn points, sketches and boxes. SAM, Semantic-SAM and SEEM are used for different interactive segmentation tasks.
- **Skill Composition.** To allow LLaVA-Plus to deal with real-world compositional tasks. We curate data for the following scenarios: (*i*) The scenarios where various visual understanding results of the same image in a multi-turn human-AI interaction session are required. We generate instruction data by applying different tools (including detection, segmentation, tagging, and captioning). (*ii*) Interactive Segmentation + Inpainting. By combining the SAM segmentation results from the user pointing and SD, we enable inpainting with visual interaction. (*iii*) Semantic Segmentation + Generation. By combining the spatial layout from OpenSeed semantic segmentation and ControlNet, we enable instructional visual-conditioned generation. (*iv*) Image Generation/Editing + Social Media Post. It is time-consuming for human users to generate posts that contain both images and text. Thus, we use SD to generate an image, or Instruct Pix2Pix to edit an image, then combine the image with its description generated by a pre-trained LMM to create a multimodal post.

2.4 Model Training and Serving

Training. To train LLaVA-Plus, we combine the curated tool use instruction data, as shown in Table 2, with the LLaVA-158K dataset. To convert LLaVA-158K into the unified prediction format as described in Section 2.2, we treat the responses in LLaVA-158K as `value`, and add the fields of `thoughts` and `actions` with templates, as illustrated in the example in Table 8 in Appendix. LLaVA-Plus are built in two settings. (*i*) *LLaVA-Plus (All Tools), where tool use is cast* *as external knowledge.* All visual understanding tools except segmentation in Table 2 are utilized to process the input image, and the extracted recognition results are organized as symbolic sequence representations to enrich the image features in both the training and evaluation stages. (*ii*) *LLaVA-Plus (Fly), where tools are used on the fly.* To reduce the cost of calling all tools, we only provide the execution results of related tools for a given instruction. When reporting quantitative numbers, we train models on the 81K understanding instruction data, because existing benchmarks focus mainly on understanding capabilities. When building demo systems, we train our models on the full dataset.

Serving. LLaVA-Plus is served using the FastChat [44] system, which is composed of web servers that interface with humans, model workers that host the LMM and multiple tools, and a controller to coordinate the web-server and model workers. The 7B LLaVA-Plus and all the tools can be loaded and served in a 80G GPU.

3 Related Works

We summarize the connections and differences between LLaVA-Plus and existing general-purpose multimodal systems in Table 3, where only representative methods are shown due to space constraint. They can be broadly categorized into two classes as discussed below.

Capabilities	**Image Understanding**			**Knowledge Image Gen.**		**Visual Interaction**	**Combined**	**Too Use**	
Input	(Text, Image)					(Point, Box)	All	Allocator	Training
Output	Text	Box	Mask	Text	Image	(Text, Image, Mask)	All		
MM-REACT	✓		✓		✓			LLM	
GPT4Tools	✓	✓	✓		✓			LLM	✓
LLaVA-Plus	✓	✓	✓	✓	✓	✓	✓	LMM	✓
LLaVA/GPT-V	✓								
Kosmos-2	✓	✓							
CM3Leon	✓		✓	✓	✓				

Table 3: Comparison with existing multimodal systems. The empty cells indicate inapplicable. "Allocator" indicates which base model is used to invoke the tools, and "Training" indicates whether model training is needed to enable tool use.

AI Agents with Multimodal Tool Use. There is a growing interest in exploring a paradigm of building general-purpose AI agents that synergistically leverage multiple tools with LLMs to solve sophisticated, open-world problems. The idea is originated in NLP to invoke general tools whose skills are lacked from LLM (*e.g.* ToolFormer [41], ChatGPT-Plugin [31]), and is recently extended to the multimodal space. There are two ways to leverage multimodal tools with the LLM as a planner to determine which tools to invoke: (*i*) tool chaining by prompt engineering and in-context-learning, such as Visual ChatGPT [45], MM-ReAct [47], and (*ii*) instruction tuning of LLM with a focus on multimodal tool use, such as GPT4Tools [46] and Gorilla [35]. Prismer [25] can efficiently aggregate the knowledge of an ensemble of domain experts and adapt it to various vision-language reasoning tasks. LLaVA-Plus represents the first work of utilizing the LMM as the planner for tool use, where image inputs are considered throughout the entire interaction sessions for improved user experience.

Unified Multimodal Models with Versatile Capabilities. Inspired by the success of a unified architecture of LLMs to complete many language tasks, the AI community has witnessed an increasing interest in building unified models with versatile multimodal capabilities. Proprietary models such as Flamingo [2] and multimodal GPT-4 [32] (or GPT-4V [33]) have demonstrated strong multimodal performance on zero-shot task transfer, which quickly inspired their open-source counterparts: LLaVA, MiniGPT-4, Open-Flamingo [4], Otter [17], to name a few. These LMMs can deal with the tasks with image-text input and text output. The capabilities have been extended to support the tasks with image-text output, such as image editing and segmentation, as demonstrated in CM3Leon [49], Emu [42], and GILL [15]. Bounding box outputs for grounding are recently supported, as shown in Kosmos-2 [37], Shikra [7] and DetGPT [38]. GPT4ROI [53] allows users to select regions of interest with bounding boxes for human-AI visual chat. BubaGPT [55] and LISA [16] use an extra referring segmentation model to enable the mask prediction capability. Compared with them, LLaVA-Plus enables a much wider range of multimodal skills and their compositions, as illustrated in Table 3.

4 Experiments

4.1 The Effectiveness of Learning to Use Skills

Tool Use Improves Existing Capabilities. We consider two benchmarks. LLaVA-Bench [24] evaluates the visual chat of LMMs, with three types of questions: conversation, detailed description and visual reasoning. It consists of two datasets: the *COCO* set containing 30 COCO images and 90 chat questions, and the *In-the-Wild* set containing 24 web images with 60 questions. Language GPT-4 (`gpt4-0314`) is used to score the generated answers. The relative scores between the model output and gold response are reported. SEED-Bench [18] evaluates the image-level and instance-level perception and reasoning of LMMs, with 19K multi-choice questions. The results are shown in Table 4. Both LLaVA-Plus variants outperform LLaVA on these two benchmarks, demonstrating the effectiveness of adding visual recognition results of applying new skills in the LMM pipeline. LLaVA-Plus (All Tools) shows superior performance to LLaVA-Plus (Fly) because the former leverages more tools as additional contexts. We further conducted several ablations: (*i*) We tried to directly add the skill execution results in the testing stage of LLaVA, shown as the row of LLaVA (Tools in Test). The degraded performance compared with LLaVA demonstrates the necessity of learning to use skills in training. (*ii*) We removed `thoughts` in the unified data format and observed a performance drop, indicating chain-of-thoughts style data format is beneficial. (*iii*) GPT4Tools trains an LLM for multimodal tool use. Its lower performance indicates that visual instruction tuning of tool use in LLaVA-Plus is important.

	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

(a) LLaVA-Bench.

	Scene	Identity	Attr.	Loc	Count	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

(b) SEED-Bench.

Table 4: LLaVA-Plus variants improves LLaVA on two LMM benchmarks.

	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

Table 5: LLaVA-Bench (Tool Use).

LLaVA-Bench (Tools). To study the novel capabilities enabled by learning to use skills, we create an evaluation set LLavA-Bench (Tools), which measures four capabilities (grounding, tagging, caption, and OCR) with 10, 12, 12, and 10 samples in each. In Table 5, we also compare against the commercial visual chat systems such as Microsoft BingChat and Google Bard. LLaVA-Plus significantly outperforms the others on this benchmark, mainly because the other systems are not equipped with some of these capabilities. By comparing with chaining tools with GPT-4 (row of "All tools + GPT4") and MM-REACT, we demonstrate the advantage of training an open-source LMM as a planner for tool use.

4.2 Comparisons with SoTA LMM systems

MMVet [50] contains 200 images and 218 questions, aiming to evaluate six core vision-language (VL) capabilities and their combinations. For evaluation, an LLM-based evaluator (`gpt4-0613`) is used to score open-ended outputs of different forms. The results are reported in Table 6. LLaVA-Plus consistently outperforms LLaVA on both 7B and 13B model sizes. The categories with most significant improvements are OCR and spatial, indicating the positive impact of the corresponding visual skills on LMM outputs.

Model	Rec	OCR	Knowledge	Generation	Spatial	Math	Total
Results of various open-source LMM on reported in the MM-VET paper [50]							
OpenFlamingo-9B [4]	24.6	14.4	13.0	12.3	18.0	15.0	21.8±0.1
BLIP-2-12B [22]	27.5	11.1	11.8	7.0	16.2	5.8	22.4±0.2
LLaVA-7B [24]	28.0	17.1	16.3	18.9	21.2	11.5	23.8±0.6
MiniGPT-4-14B [56]	29.9	16.1	20.4	22.1	22.2	3.8	24.4±0.4
Otter-9B [17]	28.4	16.4	19.4	20.7	19.3	15.0	24.6±0.2
InstructBLIP-14B [9]	30.8	16.0	9.8	9.0	21.1	10.5	25.6±0.3
MM-ReAct-GPT-3.5 [47]	24.2	31.5	21.5	20.7	32.3	26.2	27.9±0.1
LLaMA-Adapter v2-7B [11]	32.9	20.1	19.0	20.1	22.9	3.9	31.4±0.1
LLaVA-13B (V1.3, 336px) [24]	38.1	22.3	25.2	25.8	31.3	11.2	32.5±0.1
MM-ReAct-GPT-4 [47]	33.1	65.7	29.0	35.0	56.8	69.2	44.6±0.2
Results with our own experiment runs							
LLaVA-7B	30.4	13.3	19.2	20.1	18.7	8.1	24.1±0.0
LLaVA-Plus-7B (All Tools)	30.5	23.6	20.5	22.5	28.5	7.7	27.5±0.3
LLaVA-Plus-13B (All Tools, V1.3, 336px)	37.5	29.4	22.3	24.5	37.3	11.5	**35.0±0.0**
LLaVA-1.5-13B	38.0	25.0	22.6	25.0	30.9	3.8	33.3±0.1
LLaVA-1.5-Plus-13B (All Tools)	40.6	30.9	25.7	29.8	34.5	15.0	**36.8±0.4**

Table 6: Performance of various open-source LMM on MM-VET. Note that MM-ReAct is not a single multimodal model, it is a system built on chaining visual tools via GPT-3.5 or GPT-4, which we append as a reference. Our experiment running on LLaVA-7B yields very similar scores with the same checkpoint reported in MM-VET paper, indicating that our evaluation pipelines are consistent.

VisIT-Bench [5] is a real-world use oriented LMM benchmark, comprising 592 questions and 1,159 public images categorized into 70 instruction families. The results are shown in Table 7, which summarizes the battles between LMMs with GPT-analog human judgment. Elo ratings are computed by treating each pairwise human judgment as a "match". The difference between the Elo ratings of two models provides an estimate for the win probability when pitting model A vs. model B. The "#matches" column indicates the number of total matches in which a particular model participates. Win-rate indicates the win rate of a model against the human-verified reference outputs. LLaVA-Plus significantly outperforms the leading method LLaVA by 100+ ELO score, achieving a new SoTA on the leaderboard.

4.3 Visual Examples of New Capabilities

In Figure 3, we illustrate new capabilities of LLaVA-Plus with visual examples. Please see Section D in Appendix for many other interesting scenarios that demonstrate the versatile capabilities of LLaVA-Plus by learning to use skills and their compositions.

Model	Size	ELO	Matches	Win(#Ratings)
Human Reference		1382	5880	—
LLaVA-Plus	13B	**1203**	678	**35.07%** (134)
LLaVA	13B	1095	5420	18.53% (475)
mPLUG-Owl	7B	1087	5440	15.83% (480)
LlamaAdapter-v2	13B	1066	5469	14.14% (488)
Lynx	8B	1037	787	11.43% (140)
Idefics	9B	1020	794	9.72% (144)
InstructBLIP	13B	1000	5469	14.12% (503)
Otter	8B	962	5443	7.01% (499)
Visual GPT		941	5437	1.57% (510)
MiniGPT-4	11B	926	5448	3.36% (506)
Octopus V2		925	790	8.90% (146)
OpenFlamingo V1		851	5479	2.95% (509)
PandaGPT	13B	775	5465	2.70% (519)
MultimodalGPT		731	5471	0.19% (527)

Table 7: Current ELO rankings on ViSiT-Bench leaderboard as of Sept. 27th, 2023. Some model sizes are skipped due to unrevealed.

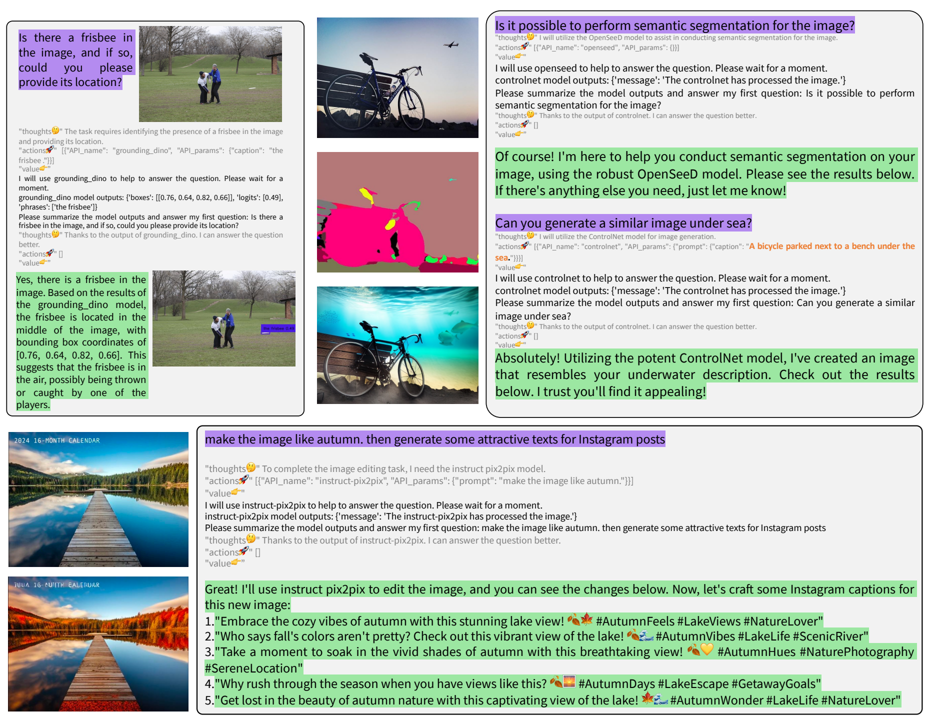


Fig. 3: New capabilities in LLaVA-Plus. Human questions $\mathbf{X}_{\mathtt{q}}$ are in purple, LLaVA-Plus responses $\mathbf{X}_{\mathtt{anwser}}$ are in green. (Left) Object detection and visual chat; (Right) Semantic segmentation and mask-based conditional image generation; (Bottom) Multimodal social media post by editing an image and writing a message.

In the left example, the questions require identifying the precise object locations. LLaVA-Plus can successfully detect the frisbee's coordinates, which help determine its status of flying in the air and thus describe the outdoor scene/activity. The same example is shown to Bard, Bing Chat, MM-REACT and LLaVA in Figure 6 in Appendix. They all fail, revealing the lack of grounding ability.

In the right example, we illustrate an interactive image editing scenario, where users aim to see the spatial layout of the scene first and then generate an image of a similar layout, but with a new "under water" scene. The LMM not only applies the correct skills, but also generates a function argument "A bicycle parked next to a bench under the sea" for conditional image generation. This reveals the appealing property of LMM as a planner, as it can see the raw image, and provide necessary image analysis results throughout the human-AI interaction process. More such examples are in Appendix Figure 11.

In the bottom example, we show that LLaVA-Plus can be used to help create multimodal social media posts. For example, when capturing an image, the user wants to post the same image in an autumn scene and associate the image with some attractive text to post Instagram. LLaVA-Plus can use the editing skills to revise the image, and combine the context of visual images and their related language topics to suggest several caption options. In Appendix Figure 12, we create all four seasons for the same scenarios, and observe that LLaVA-Plus can follow the instruction to easily switch among them while consistently maintaining the original image cue.

5 Conclusion

We have presented LLaVA-Plus, a general-purpose, multimodal assistant which is based on an LMM that plugs and learns to use skills to complete a wide range of vision-language tasks in the wild. The first visual instruction dataset specifically designed for multimodal tool use has been collected for model training. By incorporating the execution results of new skills, LLaVA-Plus consistently outperforms LLaVA across many benchmarks, creates a new SoTA and shows emergent multimodal interaction capabilities. However, LLaVA-Plus is limited due to hallucinations and tool use conflicts in practice. There are interesting problems yet to be addressed in future research on building reliable general-purpose multimodal AI agents.

Reproducibility

To ensure the reproducibility of our research, we will publicly release a comprehensive set of assets including the generated multimodal instruction data, our codebase, the LLaVA-Plus checkpoints, and a visual chat demo. Additionally, we have ensured complete transparency by elaborating on every facet of our training data collection and model training within this paper, as shown in Sec. 2.

References

1. Langchain. `https://github.com/hwchase17/langchain` (2022)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198 (2022)
3. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al.: A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861 (2021)
4. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo (Mar 2023). `https://doi.org/10.5281/zenodo.7733589`, `https://doi.org/10.5281/zenodo.7733589`
5. Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., Schimdt, L.: Visit-bench: A benchmark for vision-language instruction following inspired by real-world use (2023)
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
7. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
8. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can pre-trained vision and language models answer visual information-seeking questions? (Feb 2023)
9. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
10. Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., Gao, J.: Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision (2022)
11. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
12. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. arXiv preprint arXiv:2211.11559 (2022)
13. JaidedAI: Easyocr. `https://github.com/JaidedAI/EasyOCR` (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
15. Koh, J.Y., Fried, D., Salakhutdinov, R.: Generating images with multimodal language models. arXiv preprint arXiv:2305.17216 (2023)
16. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023)
17. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)
18. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
19. Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J.: Multimodal foundation models: From specialists to general-purpose assistants. arXiv preprint arXiv:2309.10020 (2023)

20. Li, C., Liu, H., Li, L.H., Zhang, P., Aneja, J., Yang, J., Jin, P., Hu, H., Liu, Z., Lee, Y.J., Gao, J.: ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. In: NeurIPS Track on Datasets and Benchmarks (2022)
21. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., Gao, J.: Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023)
22. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
25. Liu, S., Fan, L., Johns, E., Yu, Z., Xiao, C., Anandkumar, A.: Prismer: A vision-language model with an ensemble of experts. arXiv preprint arXiv:2303.02506 (2023)
26. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
27. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
28. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Icdar 2023 competition on hierarchical text detection and recognition. arXiv preprint arXiv:2305.09750 (2023)
29. Minsky, M.: Society of mind. Simon and Schuster (1988)
30. OpenAI: ChatGPT. `https://openai.com/blog/chatgpt/` (2023)
31. OpenAI: Chatgpt plugins. `https://openai.com/blog/chatgpt-plugins` (2023)
32. OpenAI: Gpt-4 technical report (2023)
33. OpenAI: Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf` (2023)
34. Pan, J., Sun, K., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., Dai, J., Qiao, Y., Li, H.: Journeydb: A benchmark for generative image understanding (Jul 2023)
35. Patil, S.G., Zhang, T., Wang, X., Gonzalez, J.E.: Gorilla: Large language model connected with massive apis. arXiv preprint arXiv:2305.15334 (2023)
36. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with GPT-4. arXiv preprint arXiv:2304.03277 (2023)
37. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
38. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Zhang, L.K.T.: Detgpt: Detect what you need via reasoning. arXiv preprint arXiv:2305.14167 (2023)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)

41. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761 (2023)
42. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222 (2023)
43. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. arXiv preprint arXiv:2303.08128 (2023)
44. Vicuna: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. `https://vicuna.lmsys.org/` (2023)
45. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
46. Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., Shan, Y.: Gpt4tools: Teaching large language model to use tools via self-instruction. arXiv preprint arXiv:2305.18752 (2023)
47. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381 (2023)
48. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)
49. Yu, L., et al: Scaling autoregressive multi-modal models: Pretraining and instruction tuning (2023)
50. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
51. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Gao, J., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. arXiv preprint arXiv:2303.08131 (2023)
52. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
53. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023)
54. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize anything: A strong image tagging model. arXiv preprint arXiv:2306.03514 (2023)
55. Zhao, Y., Lin, Z., Zhou, D., Huang, Z., Feng, J., Kang, B.: Bubogpt: Enabling visual grounding in multi-modal llms. arXiv preprint arXiv:2307.08581 (2023)
56. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
57. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., Peng, N., Wang, L., Lee, Y.J., Gao, J.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)
58. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718 (2023)