A Survey and Assessment of Large Models for Visually Impaired Assisting

Yi Zhao, **Rong Xiang**, **Jing Li** *
Hong Kong Polytechnic University
yi-yi-yi.zhao@connect.polyu.hk, {rong-chris.xiang, jing-amelia.li}@polyu.edu.hk

Abstract

One important motivation in computer vision (CV) is assisting visually impaired persons (VIPs). Recently, the advent of Large Models (LMs) has significantly propelled the domains of Natural Language Processing (NLP) and CV, setting unprecedented benchmarks across a spectrum of tasks. However, despite these developments, a significant research gap in LM-based assistance for VIPs exists. This paper aims to address this gap and, to the best of our knowledge, is the first to explore how LMs can aid the visually impaired. An extensive survey of Large Language Models (LLMs), Vision Language Models (VLMs), and Embodied Agents is presented, assessing their prospective roles in facilitating VI assistance. Furthermore, this paper provides an in-depth evaluation of the current state-of-the-art end-to-end VLMs, such as GPT-4, and offers critical insights into their capabilities and limitations in assisting VIPs. The approach undertaken models this task as a visual question answering problem, with outputs being specifically tailored to be grounded tactile guidance to meet the unique needs of VIPs. In summary, the conducted survey and assessment suggest future directions for enhancing VI assistance through Large Models.

1 Introduction

One of the primary motivations for developing computer vision (CV) technologies was to aid visually impaired individuals. Tasks such as Visual Question Answering (VQA) have garnered attention and resulted in advancements that benefit the visually impaired community. VizWiz [Gurari *et al.*, 2018] is the first VQA dataset specially from the visually impaired individuals. Furthering this, Vizwiz-priv [Gurari *et al.*, 2019] is the inaugural privacy-aware VQA dataset originating from this community. [Lasecki *et al.*, 2013] introduced Chorus:View, a system for assisting visually impaired in answering visual questions through the engagement of on-demand crowd-sourced human workers. Similarly, [Ahmetovic *et al.*, 2016] proposed NavCog, a navigation assistant for the visually impaired. While these developments represent significant strides in assistance for visually impaired persons (VIPs), the advent of large-scale models has opened new frontiers. Recent advancements in large-scale models have exhibited capabilities in visual perception, reasoning, planning, decision-making for actions, and interaction with environments. However, at the current time, there is a noticeable research gap in LM-based assistance for visually impaired individuals.

The advancements in LMs are multifaceted. In the area of natural language processing (NLP), models such as GPT-3 [Brown *et al.*, 2020] have established new benchmarks. ChatGPT, a derivative of InstructGPT [Ouyang *et al.*, 2022], has attracted considerable attention due to its proficiency in diverse NLP tasks via dialogic interactions. In addition to the GPT series, other large-scale models such as PaLM [Chowdhery *et al.*, 2022] and LLaMa [Touvron *et al.*, 2023a] have also been developed. Similar to their GPT counterparts, these models demonstrate emergent capabilities [Wei *et al.*, 2022a]. Among the advancements in LLMs, certain methodologies have been particularly noteworthy. Notably, RHFL [Ouyang *et al.*, 2022] has been pivotal in tailoring model functionalities to align with human-centric instructions, and COT [Wei *et al.*, 2022b] sheds light on tapping into the intrinsic reasoning prowess of these models. LLMs demonstrate significant proficiency in reasoning and planning, which could substantially improve decision-making in tasks that involve assisting the visually impaired.

Following advancements in LLMs, VLMs have similarly experienced significant growth, setting new benchmarks in state-of-the-art performance for multimodal tasks that combine visual and linguistic elements, such as Image Captioning and Visual Question Answering. CLIP [Radford *et al.*, 2021] has demonstrated proficiency in zero-shot visual classification, especially for non-predefined categories. Other notable pretrained VLMs include Flamingo [Alayrac *et al.*, 2022], the BLIP series [Li *et al.*, 2022; Li *et al.*, 2023], PaLI-X [Chen *et al.*, 2023b], among others. Especially, GPT-4 [OpenAI, 2023] has showcased human-equivalent performance in complex multi-modal reasoning tasks. VLMs inherit the reasoning and planning capabilities of LLMs, and by integrating a visual module, they acquire enhanced visual perception capacities. This advancement makes them highly suitable as foundational models for developing assistance tools for visu-

*Coresponding author

ally impaired individuals.

In advancing beyond multi-modal models with an aim to progress toward Artificial General Intelligence (AGI), researchers have developed Embodied Agents. These agents possess cognitive abilities such as reasoning and planning, and are capable of taking actions and interacting with environments in both simulated contexts and the real physical world. ReAct [Yao *et al.*, 2022] employs LLMs to interleave the generation of reasoning trajectories and actions for task achievement. This approach demonstrates superior decision-making capabilities compared to methods based on imitation and reinforcement learning. In simulated environments, Ghost in the Minecraft [Zhu *et al.*, 2023b] leverage LLMs to process text-based memory and knowledge. This approach enables it to effectively solve long-horizon tasks and manage uncertainties in the open-world game Minecraft. Voyager [Wang *et al.*, 2023a] represents the first LLM-based embodied agent capable of lifelong learning. It autonomously explores environments, also notably in the Minecraft setting, and develops skills through environmental feedback. In the real physical world, PaLM-E [Driess *et al.*, 2023], built on the 562B parameter LLM PaLM, integrates sensor modalities into language models for physical-world applications. Additionally, RT-2 [Zitkovich *et al.*, 2023], which stands for "Robot Transformer," is trained using both robotic data and large-scale internet visual-language datasets. This approach introduces the concept of the Visual-Language-Robot model.

Since the development of VI assistance demands models that can interact with the physical world, the similar paradigm used for developing embodied agents can be referenced. Nevertheless, it's essential to differentiate between the requirements for developing these embodied robots and assistance tools specifically designed for the visually impaired. We have identified and outlined three primary distinctions:

- The first notable distinction lies in the nature of the guidance. While embodied agents use LMs to generate strictly executable commands, guidance for assisting the visually impaired requires a different approach. Instead of precise instructions, it should be understandable and accessible to those who are blind. Specifically, neuroscience studies have shown that tactile sensation plays an important role for the visually impaired. [Goldreich and Kanics, 2003] demonstrated that VI individuals possess enhanced tactile acuity, and [Ottink *et al.*, 2022] revealed that tactile input aids VI individuals in forming cognitive maps as accurately as their non-VI counterparts. Therefore, we suggest that guidance for VI assistance should incorporate tactile information.

- The second critical distinction lies in the complexity of adapting to environmental feedback. While embodied agents may fail in action execution and subsequently adjust their behavior, the situation is significantly more intricate for the visually impaired. Given the inherent nature of humans not adhering as strictly to instructions as robots, and the risk of inadvertently entering hazardous areas, there is an imperative need for a system capable of both accommodating and promptly correcting deviations. Such a system is crucial for effectively addressing unforeseen errors or incidents, thus ensuring both safety and efficacy.

- The third distinction lies in the interaction dynamics. Unlike embodied agents, assistance for visually impaired individuals necessitates direct human interaction. Therefore, there is a likely need to prioritize encouraging and empathetic responses over impersonal, mechanical instructions. Helping them accomplish tasks not only eases their physical lives but also boosts their self-confidence and fosters a sense of positive well-being.

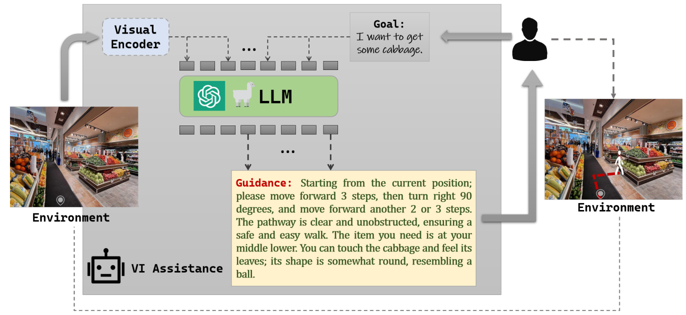

Figure 1: **VIP Assistance Illustration.** This general framework illustrates that the input consists of a visual signal, while the VI user verbally communicates their specific goal to the assistance system. The LLM then undertakes reasoning and planning to achieve this goal, grounded in the physical environment. It is crucial that the guidance provided is understandable to the VI user. Given the significant role of tactile sensation in the lives of visually impaired individuals, we emphasize the necessity of tactile guidance as the desired output. Ultimately, this guidance should enable the VI user to accomplish the intended goal.

Accordingly, Figure 1 depicts a general framework for VI assistance, highlighting the necessity of tactile guidance due to the pivotal role of tactile sensation in the lives of VI individuals. Aligned with this perspective, we have developed a foundational benchmark composed of 200 images, encompassing scenarios in both supermarkets and domestic environments. This benchmark serves to evaluate the capabilities of current large models. Comprehensive assessments were conducted on 6 end-to-end VLMs, including GPT-4, CogVLM [Wang *et al.*, 2023b], Qwen-VL [Bai *et al.*, 2023], LLaVA [Liu *et al.*, 2023], MiniGPT-v2 [Chen *et al.*, 2023a], and BLIVE [Hu *et al.*, 2023a].

The contributions of this work are twofold: (1). Firstly, it encompasses a comprehensive survey of large-scale models, including LLMs, VLMs, and Embodied Agents, along with relevant datasets, benchmarks, and an analysis of how those prior studies can contribute to the development of VI assistance. (2). Secondly, it delivers an in-depth assessment of the performance of end-to-end VLMs, including GPT-4, using a manually crafted benchmark. This offers readers a foundational understanding of these models' abilities to meet current needs and identifies potential directions for future research.

The remainder of this article is structured as follows: Section 2 provides a comprehensive survey of large models, Section 3 offers an in-depth assessment of these models, and the article concludes with Section 4, which outlines potential di-

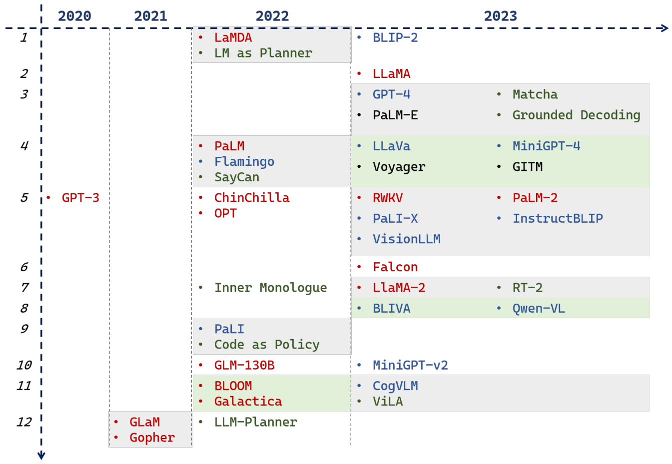

Figure 2: **Timeline of the LMs**

rections for future research.

2 Large Models

2.1 LLMs

Table 1

2.2 VLMs

Table 2

2.3 Embodied Agents

Table 3

3 Assessment

References

[Ahmetovic *et al.*, 2016] Dragan Ahmetovic, Cole Gleason, Chengxiong Ruan, Kris Kitani, Hironobu Takagi, and Chieko Asakawa. Navcog: a navigational cognitive assistant for the blind. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 90–99, 2016.

[Ahn *et al.*, 2022] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[Almazrouei *et al.*, 2023] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. Falcon-40b: an open large language model with state-of-the-art performance. *Findings of the Association for Computational Linguistics: ACL*, 2023:10755–10773, 2023.

[Anil *et al.*, 2023] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 3(1), 2023.

[Brock *et al.*, 2021] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning*, pages 1059–1071. PMLR, 2021.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Chen *et al.*, 2022] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[Chen *et al.*, 2023a] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[Chen *et al.*, 2023b] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[Chowdhery *et al.*, 2022] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Name	**Release Time**	**Pretraining Task**	**Model Architecture**	**# Maximum Parameters**	Instruction-tuned Versions
GPT-3 [Brown *et al.*, 2020]	2020.05	①	①	175B	InstructGPT, GPT-3.5
GLaM [Du *et al.*, 2022]	2021.12	①	②	1.2T	/
Gopher [Rae *et al.*, 2021]	2021.12	①	①	280B	/
LaMDA [Thoppilan *et al.*, 2022]	2022.01	①	①	137B	Bard [Manyika, 2023]
PaLM [Chowdhery *et al.*, 2022]	2022.04	①	①	540B	Flan-PaLM [Chung *et al.*, 2022]
Chinchilla [Hoffmann *et al.*, 2022]	2022.05	①	①	70B	/
OPT [Zhang *et al.*, 2022]	2022.05	①	①	175B	OPT-IML [Iyer *et al.*, 2022]
GLM-130B [Zeng *et al.*, 2022]	2022.10	②	③	130B	ChatGLM
BLOOM [Workshop *et al.*, 2022]	2022.11	①	①	176B	BLOOMZ [Muennighoff *et al.*, 2022]
Galactica [Taylor *et al.*, 2022]	2022.11	①	①	120B	Evol-Instruct
LLaMA [Touvron *et al.*, 2023a]	2023.02	①	①	65B	Alpaca [Taori *et al.*, 2023], WizardLM, Vicuna [Chiang *et al.*, 2023]
RWKV [Peng *et al.*, 2023]	2023.05	①	④	14B	RWKV-4 Raven
PaLM-2 [Anil *et al.*, 2023]	2023.05	③	⑤	/	/
Falcon [Almazrouei *et al.*, 2023]	2023.06	①	①	40B	Falcon-instruct
LLaMA-2 [Touvron *et al.*, 2023b]	2023.07	①	①	70B	LLaMA2-Chat, OpenChat V2

Table 1: Summary of popular pretrained LLMs and their instructed versions, arranged chronologically from the earliest to the most recent releases. For the pretraining task category, symbols ①, ②, and ③ denote language modeling [Radford *et al.*, 2019],autoregressive blank infilling, and mixture of denoisers [Tay *et al.*, 2022], respectively. For the model architecture category, ①, ②, ③, and ④ represent transformer decoder [Radford *et al.*, 2018], mixture-of-experts decoder, bidirectional GLM, RWKV architecture, and transformer, respectively. The symbol "/" is used when specific information is not explicitly available. All these models have a maximum parameter count exceeding 10 billion, with some surpassing 100 billion. The majority of these LLMs were released in the past two years. In addition, numerous instruction-tuned models are based on the LLaMA framework.

[Chung *et al.*, 2022] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[Dehghani *et al.*, 2023] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Driess *et al.*, 2023] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[Du *et al.*, 2022] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.

[Fang *et al.*, 2023] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[Goldreich and Kanics, 2003] Daniel Goldreich and Ingrid M Kanics. Tactile acuity is enhanced in blindness. *Journal of Neuroscience*, 23(8):3439–3445, 2003.

[Gurari *et al.*, 2018] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[Gurari *et al.*, 2019] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2019.

Name	**Release Time**	**Visual Encoder**	**LLM**	**Connector**	**Training/Tuning**
Flamingo [Alayrac *et al.*, 2022]	2022.04	NFNet [Brock *et al.*, 2021]	Chinchilla	Cross-attention dense layers	Pretraining
PaLI [Chen *et al.*, 2022]	2022.09	ViT-e [Zhai *et al.*, 2022]	mT5 [Raffel *et al.*, 2020]	/	Multi-task Pretraining
BLIP-2 [Li *et al.*, 2023]	2023.01	1. ViT-L/14 [Radford *et al.*, 2021], 2. ViT-g/14 [Fang *et al.*, 2023]	1. OPT [Zhang *et al.*, 2022], 2. FlanT5 [Chung *et al.*, 2022]	Q-former	Multi-task Pretraining
GPT-4 [OpenAI, 2023]	2023.03	/	GPT-4	/	/
LLaVa [Liu *et al.*, 2023]	2023.04	ViT-L/14	LLaMA	Projection layer	1. Pretraining, 2. Instruction tuning.
MiniGPT-4 [Zhu *et al.*, 2023a]	2023.04	Blip-2 ViT-g/14 and Q-Former	Vicuna	Linear projection layer	1. Pretraining, 2. Instruction tuning.
PaLI-X [Chen *et al.*, 2023b]	2023.05	ViT-22B [Dehghani *et al.*, 2023]	UL2 [Tay *et al.*, 2022]	Projection layer	1. Multi-task pretraining, 2. Task-specific fine-tuning.
InstructBLIP [Dai *et al.*, 2023]	2023.05	ViT-g/14	1. FlanT5, 2. Vicuna	Q-former	1. Pretraining, 2. Instruction tuning.
VisionLLM [Wang *et al.*, 2023c]	2023.05	1. ResNet, 2. InternImage-H	Alpaca	BERT-Base and Deformable DETR [Zhu *et al.*, 2020]	Multi-task pretraining
BLIVA [Hu *et al.*, 2023a]	2023.08	ViT-g/14	FlanT5	Q-former and projection layer	1. Pretraining, 2. Instruction-tuning.
Qwen-VL [Bai *et al.*, 2023]	2023.08	ViT [Dosovitskiy *et al.*, 2020]	Qwen	Single-layer cross-attention module	1. Pretraining, 2. Multi-task training, 3. Instruction tuning.
MiniGPT-v2 [Chen *et al.*, 2023a]	2023.10	ViT-g/14	LLaMA2-Chat	Linear projection layer	1. Pretraining, 2. Multi-task training, 3. Instruction tuning.
CogVLM [Wang *et al.*, 2023b]	2023.11	Eva-clip ViT [Sun *et al.*, 2023]	Vicuna	MLP adapter	1. Pretraining; 2. Multi-task training

Table 2: Summary of popular pretrained Vision-Language Models (VLMs), arranged chronologically from the earliest to the latest. Typically, a large-scale VLM comprises a Visual Encoder, a LLM, and a vision-language connector. It indicates that ViT is the most favored visual encoder, while LLMs primarily derive from the LLaMA family, including variants like LLaMA, LLaMA2-Chat, Vicuna, and Alpaca. The most common connectors are either a straightforward linear projection layer or cross-attention layers. The prevalent pre-training methodology for VLMs generally encompasses three phases. Initially, extensive image-text pairs from the internet are employed for foundational training. Subsequently, multi-task, fine-grained tuning is applied. The final stage primarily utilizes instructional or conversational data, optimizing the model for interactive language-based user engagement.

Name	**Time**	**LLM**	**Task**	**Environment**	**LM Output**	**Code Interface**	**LM Prompting/Tuning**
LM as Planner [Huang *et al.*, 2022a]	2022.01	GPT3, CodeX	VirtualHome tasks	①	①	①	①
SayCan [Ahn *et al.*, 2022]	2022.04	PaLM	Real-world robotic tasks	②	②	①	①
Code as Policy [Liang *et al.*, 2023]	2022.09	Codex	Real-world robotic tasks	②	②	②	①
Inner Monologue [Huang *et al.*, 2022b]	2022.07	InstructGPT	Ravens tasks and real-world robotic tasks	① + ②	①	①	①
ProgPrompt [Singh *et al.*, 2023]	2022.09	GPT-3	VituralHome tasks and real-world robotic tasks	① + ②	②	②	①
LLM-Planner [Song *et al.*, 2023]	2022.12	GPT-3	Alfred tasks	①	①	①	①
Matcha [Zhao *et al.*, 2023]	2023.03	GPT-3	CoppeliaSim-simulated NICOL robot tasks	①	①	①	①
PaLM-E [Driess *et al.*, 2023]	2023.03	PaLM	TAMP tasks and real-world robotic tasks	②	①	①	①
Grounded Decoding [Huang *et al.*, 2023]	2023.03	InstructGPT, PaLM	Ravens tasks and real-world robotic tasks	① + ②	①	①	①
Voyager [Wang *et al.*, 2023a]	2023.05	GPT-4	Minecraft Tasks	①	②	②	①
GITM [Zhu *et al.*, 2023b]	2023.05	GPT-3.5	Minecraft Tasks	①	②	①	①
RT-2 [Zitkovich *et al.*, 2023]	2023.07	PaLI-X,PaLM-E	Real-world robotic tasks	②	②	①	②
ViLA [Hu *et al.*, 2023b]	2023.11	GPT-4V	Ravens tasks and real-world robotic tasks	① + ②	①	①	①

Table 3: Summary of popular Embodied agents, arranged chronologically from the earliest to the most recent releases. In the Environment category, symbols ① and ② respectively denote simulated environments and physical real-world environments. In the LM (Language Model) output category, symbol ① represents mid-level plans, while ② signifies low-level executable robotic actions. For the Code Interface category, symbol ① indicates an interface without programming code (relying solely on natural language), and symbol ② represents an interface with programming code. In the LM Prompting/Tuning Column, ① corresponds to LM prompting, and ② corresponds to LM tuning. Most Embodied Agents use LM prompting to generate mid-level plans in natural language, which are then translated into low-level robotic actions for execution. Both simulated and real-world physical environments have been studied, with the most common approach being the use of GPT-3/4 with API-based prompting. In addition, embodied agents are trending toward using the VLM-based strategy for improved visual understanding and alignment instead of solely relying on LLMs.

[Hoffmann *et al.*, 2022] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[Hu *et al.*, 2023a] Wenbo Hu, Yifan Xu, Y Li, W Li, Z Chen, and Z Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. *arXiv preprint arXiv:2308.09936*, 2023.

[Hu *et al.*, 2023b] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023.

[Huang *et al.*, 2022a] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.

[Huang *et al.*, 2022b] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[Huang *et al.*, 2023] Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, et al. Grounded decoding: Guiding text generation with grounded models for robot control. *arXiv preprint arXiv:2303.00855*, 2023.

[Iyer *et al.*, 2022] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

[Lasecki *et al.*, 2013] Walter S Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P Bigham. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–8, 2013.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[Liang *et al.*, 2023] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[Manyika, 2023] James Manyika. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2023.

[Muennighoff *et al.*, 2022] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

[OpenAI, 2023] OpenAI. Gpt-4 technical report, 2023.

[Ottink *et al.*, 2022] Loes Ottink, Bram van Raalte, Christian F Doeller, Thea M Van der Geest, and Richard JA Van Wezel. Cognitive map formation through tactile map navigation in visually impaired and sighted persons. *Scientific reports*, 12(1):11567, 2022.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[Peng *et al.*, 2023] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

[Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rae *et al.*, 2021] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer.

The Journal of Machine Learning Research, 21(1):5485–5551, 2020.

[Singh *et al.*, 2023] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023.

[Song *et al.*, 2023] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

[Sun *et al.*, 2023] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[Taori *et al.*, 2023] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[Tay *et al.*, 2022] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2022.

[Taylor *et al.*, 2022] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[Thoppilan *et al.*, 2022] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[Touvron *et al.*, 2023a] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Touvron *et al.*, 2023b] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2023a] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[Wang *et al.*, 2023b] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. 2023.

[Wang *et al.*, 2023c] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023.

[Wei *et al.*, 2022a] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[Wei *et al.*, 2022b] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[Workshop *et al.*, 2022] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[Yao *et al.*, 2022] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[Zeng *et al.*, 2022] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[Zhai *et al.*, 2022] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022.

[Zhang *et al.*, 2022] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[Zhao *et al.*, 2023] Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. Chat with the environment: Interactive multimodal perception using large language models. *arXiv preprint arXiv:2303.08268*, 2023.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr:

Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[Zhu *et al.*, 2023a] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[Zhu *et al.*, 2023b] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.

[Zitkovich *et al.*, 2023] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *7th Annual Conference on Robot Learning*, 2023.