多模态大语言模型综述

周浩榕[1)]

[1)](西安交通大学计算机科学与技术学院 西安 710049)

摘 要 在过去的一年里，多模态大语言模型（Multimodal Large Language Models, MM-LLMs）取得了显著进展，通过经济高效的训练策略，增强了现成的 LLMs 对多模态输入或输出的支持。这些模型不仅保留了 LLMs 固有的推理和决策能力，还增强了对各种多模态任务的处理能力。本文提供了一份全面的调查，旨在促进多模态大型语言模型的进一步研究。首先，我们概述了模型架构和训练流程的一般设计原理。随后，我们引入了一个包含 126 个多模态大型语言模型的分类体系，每个模型都有其特定的公式。此外，我们还回顾了部分多模态大型语言模型在主流基准测试上的表现，并总结了提高多模态大型语言模型效能的关键训练策略。最后，我们探讨了多模态大型语言模型有前景的发展方向。我们希望这份调查能为多模态大型语言模型领域的持续发展做出贡献。

关键词 多模态；大语言模型；分类体系；性能；训练策略；综述

中图法分类号 **** DOI 号

A Survey of Multimodal Large Language Models

ZHOU Hao-Rong[1)]

[1)](School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an , 710049)

Abstract Over the past year, significant progress has been made in multimodal large language models (MM-LLMs), which have enhanced the support for multimodal inputs or outputs in off-the-shelf LLMs through cost-effective training strategies. These models not only retain the inherent reasoning and decision-making capabilities of LLMs but also augment their processing abilities for a variety of multimodal tasks. This paper presents a comprehensive survey aimed at fostering further research in multimodal large language models. Firstly, we outlined the general design principles of model architecture and training procedures. Subsequently, we introduce a taxonomy encompassing 126 multimodal large language models, each with its specific formula. Furthermore, we review the performance of some multimodal large language models on mainstream benchmarks and summarize the key training strategies for enhancing their efficacy. Finally, we discuss promising directions for the development of multimodal large language models. We hope that this survey contributes to the continuous advancement of the field of multimodal large language models.

Key words multimodal; large language models; taxonomy; performance; training strategies; survey

1 引言

近年来，多模态（Multimodal, MM）预训练研究取得了重大进展，不断推动下游任务的性能界限。然而，随着模型和数据集的规模不断扩大，传统的 MM 模型会产生大量的计算成本，特别是从头训练时。认识到 MM 研究是在各种模式的交叉点上进行的，一个合乎逻辑的方法是利用现成的预训练单模基础模型，特别强调强大的大型语言模型（Large Language Models, LLMs）。该策略旨在减少计算费用并提高 MM 预训练的效率，从而出现了一个新的领域：MM-LLMs。

MM-LLMs 利用 LLMs 作为认知引擎来支持各

收稿日期：年-月-日；最终修改稿收到日期：年-月-日.周浩榕(**通信作者**)，女，2003年生，本科生，.E-mail: 1049512158@qq.com.

种 MM 任务。LLMs 提供了理想的特性，如强大的语言生成，零迁移能力和上下文学习（ICL）。同时，其他模式中的基础模型提供了高质量的表示。考虑到不同模式的基础模型是单独预训练的，MM-LLMs 面临的核心挑战是如何有效地将 LLMs 与其他模式的模型连接起来，从而实现协同推理。这个领域的主要焦点是通过 MM 预训练(PT)+MM 指令调整（IT）管道，改进模式之间的对齐并与人类意图对齐。

GPT-4(Vision)和 Gemini[1]的首次亮相，展示了 MM-LLMs 令人印象深刻的 MM 理解和生成能力，引发了众人对 MM-LLMs 的研究热情。最初的研究主要集中在 MM 内容理解和文本生成，包括图像-文本理解等任务，例如 BLIP-2[2]、LLaVA[3]、MiniGPT4[4]和 OpenFlamingo[5]等项目；视频文本理解，如 VideoChat[6]、Video-ChatGPT[7]和 LLaMA-VID[8]等倡议所证明的；以及音频-文本理解，如 QwenAudio[9]等项目。后来，MM-LLMs 的功能得到了扩展，以支持特定的模态输出。这包括具有图像-文本输出的任务，如 GILL[10]、Kosmos -2[11]、Emu[12]和 MiniGPT-5[13]；以及语音/音频-文本输出，如 SpeechGPT[14]和 AudioPaLM[15]等项目。最近的研究努力集中在模仿类似人类的任意形态转换，为人工通用智能的道路提供了光明。一些努力旨在将 LLMs 与外部工具合并，以达到接近任意对任意 MM 的理解和生成，例如 VisualChatGPT[16]、HuggingGPT[17]和 AudioGPT[18]。相反，为了减轻级联系统中的传播误差，NExT-GPT[19]、CoDi-2[20]和 ModaVerse[21]等计划则开发了任意模态的端到端 MM-LLMs。

本文对 MM-LLMs 进行了全面的综述，旨在促进 MM-LLMs 的进一步研究。我们首先从模型架构（第 2 节）和训练管线（第 3 节）描述一般设计原理。我们将一般模型架构分解为五个组件：模态编码器（第 2.1 节）、输入投影仪（第 2.2 节）、LLM 主干（第 2.3 节）、输出投影仪（第 2.4 节）和模态生成器（第 2.5 节）。培训管道阐明了如何增强预训练的纯文本 LLM 以支持 MM 输入或输出，主要由两个阶段组成：MM PT（第 3.1 节）和 MM IT（第 3.2 节）。接下来，我们建立了包含 126 个最先进（SOTA）MM-LLMs 的分类，每个 MM-LLMs 都具有特定的配方特征，并在第 4 节中总结了它们的发展趋势。在第 5 节中，我们全面回顾了主要 MM-LLMs 在主流基准上的表现，并提炼出关键的培训配方，以提高 MM-LLMs 的有效性。在第 6 节中，我们为 MM-LLMs 的研究提供了有希望的方向。最后，我们在第 7 节对全文进行了总结。我们希望我们的调查能够帮助研究者对这一领域有更深入的了解，并为设计更有效的 MM-LLMs 提供启发。

2 模型架构

在本节中，我们将详细概述组成一般模型体系结构的五个组件，以及每个组件的实现选择。强调 MM 理解的 MM-LLMs 只包括前三个组件。在训练期间，模态编码器、LLM 主干和模态生成器通常保持在冻结状态。主要的优化重点是输入和输出投影仪。鉴于投影仪是轻量级组件，MM-LLMs 中可训练参数的比例与总参数数相比明显很小（通常约为 2%）。总体参数数取决于 MM-LLMs 中使用的核心 LLM 的规模。因此，可以有效地训练 MM-LLMs，以授权各种 MM 任务。

2.1 模态编码器

模态编码器（ME）的任务是对来自不同模态 I_X 的输入进行编码，以获得相应的特征 F_X。存在各种预训练的编码器选项用于处理不同的模式，其中 X 可以是图像、视频、音频、3D 等。接下来，我们将按情态进行简要介绍。

对于图像，有各种可选的编码器：NFNet-F6[22]、ViT[23]、CLIP ViT[24]、Eva-CLIP ViT[25]、BEiT-3[26]、OpenCLI[27]、Grounding-DINOT[28]、DINOv2[29]、SAM-HQ[30]、RAM++[31]、InternViT[32]和 VCoder[33]。对于视频，可以统一采样到 5 帧，进行与图像相同的预处理。

音频模态通常由 CFormer[34]、HuBERT[35]、BEATs[36]、Whisper[37]和 CLAP[38]编码。

3D 点云模态通常由 ULIP-2[39]与 Point-BERT[40]主干编码。

此外，为了处理众多异构模态编码器，一些 MM-LLMs，特别是任意对任意的 MM-LLMs，使用 ImageBind[41]，这是一种涵盖六种模态的统一编码器，包括图像/视频、文本、音频、热图、惯性测量单元和深度。

2.2 输入投影仪

输入投影仪的任务是将其他模式 F_X 的编码特征与文本特征空间 T 对齐，然后将对齐的特征作为提示 P_X 与文本特征 F_T 一起馈送到 LLM 主干中。

给定 X-text 数据集 $\{I_X, t\}$，目标是最小化 X 条件文本生成损失。

输入投影仪可以直接通过线性投影仪或多层感知器（MLP）来实现，即几个线性投影仪与非线性激活函数交错。还有更复杂的实现，如交叉注意、Q-Former[42]、P-Former[43]和 MQ-Former[44]。交叉注意（Percepver Resampler）[45]使用一组可训练向量作为查询，编码特征 F_X 作为键，将特征序列压缩到固定长度。然后将压缩的表示直接输入 LLM 或进一步用于 X-text 数据集交叉注意融合。Q-Former 通过可学习的查询从 F_X 中提取相关特征，然后将选中的特征用作提示 P_X。同时，P-Former 产生“参考提示”，对 Q-Former 产生的提示进行对齐约束。MQ-Former 对多尺度视觉和文本信号进行细粒度对齐。然而，Q-、P-、MQ-Former 都需要一个额外的 PT 进程进行初始化。

2.3 LLM主干

拿 LLM 作为核心智能体，MM-LLMs 可以继承一些值得注意的属性，如零次泛化、少次 ICL、思维链（CoT）和指令遵循。LLM 主干处理来自各种模态的表示形式，对输入进行语义理解、推理和决策。它产生直接文本输出 t 和来自其他模态（如果有的话）的信号令牌 S_X。这些信号令牌充当指示，指导生成器是否生成 MM 内容，如果是肯定的，则指定要生成的内容，其中其他模态 P_X 的对齐表示可以被视为 LLM 的软提示调优。此外，一些作品介绍了参数高效微调（PEFT）方法，如 Prefix-tuning[46]、LoRA[47]和 LayerNorm tuning[48]。在这些情况下，额外可训练参数的数量非常少，甚至不到 LLM 参数总数的 0.1%。

2.4 输出投影仪

输出投影仪将 LLM 主干的信号令牌表示 S_X 映射为 H_X 可理解的特征，以下面的模态生成器 MG_X。给定 X-text 数据集 $\{I_X, t\}$，首先将 t 输入 LLM 以生成相应的 S_X，然后将其映射到 H_X。为了方便映射特征 H_X 的对齐，目标是最小化 H_X 和 MG_X 的条件文本表示之间的距离。

2.5 模态生成器

模态生成器 MG_X 的任务是产生不同模态的输出。通常，现有的工作使用现成的潜在扩散模型（LDMs），比如图像合成用 Stable Diffusion[49]，视频合成用 Zeroscope[50]，音频合成用 AudioLDM[51]。由输出投影仪映射的特征 H_X 作为去噪过程中的条件输入来生成 MM 内容。在训练过程中，地面真值内容首先被预训练的 VAE 转化为潜在特征 z_0。然后，将噪声加到 z_0 中，得到噪声潜在特征 z_t。

3 训练管线

MM-LLMs 的训练管线可以划分为两个主要阶段：MM PT 和 MM IT。

3.1 MM PT

在 PT 阶段，通常利用 X-text 数据集，训练输入和输出投影仪通过优化预定义目标来实现各种模式之间的对齐。

X-text 数据集包括图像-文本、视频-文本和音频-文本，其中图像-文本有两种类型：图像-文本对和交错图像-文本语料库。

3.2 MM IT

MM IT 是一种使用指令格式的数据集对预训练的 MM-LLMs 进行微调的方法。通过这个过程，MM-LLMs 可以通过遵守新的指令来推广到未见过的任务，从而提高零射击性能。这简单而有影响力的概念催化了 NLP 领域的后续成功，例如 Instruct-GPT[52]、OPT-IML[53]和 InstructBLIP[54]。

MM IT 包括监督微调（SFT）和人类反馈强化学习（RLHF），旨在与人类意图保持一致，增强 MM-LLMs 的交互能力。SFT 将部分 PT 阶段数据转换为指令感知格式。以可视化问答为例，可以使用各种模板。接下来，它使用相同的优化目标对预训练的 MM-LLMs 进行微调。SFT 数据集可以构建为单回合问答或多回合对话。

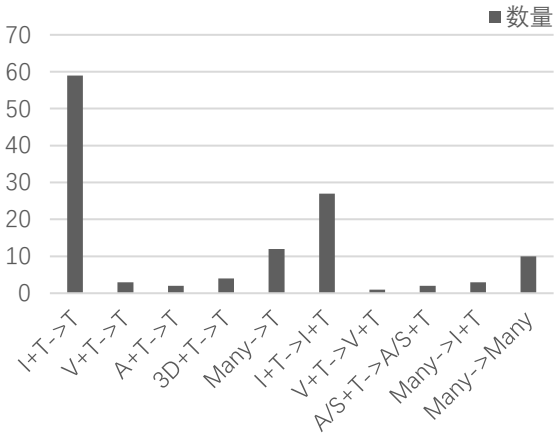
在 SFT 之后，RLHF 涉及模型的进一步微调，依赖于 MM-LLMs 响应的反馈（例如，手动或自动标记的自然语言反馈（NLF））。该过程采用强化学习算法对不可微 NLF 进行有效积分。对模型进行训练以产生以 NLF 为条件的相应响应。

4 SOTA MM-LLMs

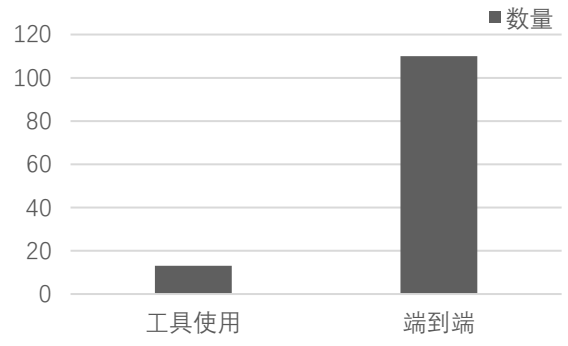
我们从功能和设计两个角度分别对 123 个当前最先进的 MM-LLMs 进行了分类与统计，统计结果如图 1 所示。在设计分类中，“工具使用”表示将 LLM 作为黑盒，通过推理向 LLM 提供访问某些 MM 专家系统以执行特定的 MM 任务；“端到端”表示整个模型以端到端的方式进行联合训练。接下

来，我们将总结它们的发展趋势，并简要介绍一些代表性模型的核心贡献。

现有 MM-LLMs 的趋势：(1)从专注于 MM 的理解到特定模态的生成，并进一步演变为任意到任意模态的转换（例如，MiniGPT-4→MiniGPT-5→NExT-GPT）；(2)从 MM PT 到 SFT 再到 RLHF，训练管线不断完善，力求更好地符合人类意图，增强模型的会话交互能力（例如，BLIP-2→InstructBLIP→DRESS）；(3)采用多样化的模态扩展（例如，BLIP-2→X-LLM 和 InstructBLIP→X-InstructBLIP）；(4)纳入更高质量的训练数据集（例如，LLaVA→LLaVA-1.5）；(5)采用更高效的模型架构，从 BLIP-2 和 DLP 中复杂的 Q-和 P-Former 输入投影机模块过渡到 VILA 中更简单但有效的线性投影机。



(a) 按功能对 MM-LLMs 进行分类的统计结果



(b) 按设计对 MM-LLMs 进行分类的统计结果

图 1 MM-LLMs 的分类统计。I：图像，V：视频，A/S：音频/语音，T：文本。

5 基准及表现

为了提供全面的性能比较，我们编制了一个表，列出了 18 个视觉语言（VL）基准测试中的主要 MM-LLMs，表 1 给出了这些信息。考虑到可用的众多基准，我们将重点放在基于 OKVQA、IconVQA、VQA^{v2} 和 GQA 的不同 MM-LLMs 的评估和比较上。

OKVQA 包括需要用各种知识类型（如常识、世界知识和视觉知识等）进行推理的问题。MiniGPT-v2 和 MiniGPT-v2-chat 在这个基准测试中表现最好，展示了它们出色的推理能力。IconVQA 强调抽象图表理解和整体认知推理在现实世界基于图表的文字问题中的重要性，这既需要敏锐的感知能力，也需要全面的认知推理。MiniGPT-v2 和 MiniGPT-v2-chat 在这个基准测试中也表现出色，突出它们卓越的感知和认知推理能力。VQA^{v2} 是一个更加平衡的 VQA 数据集，其中每个问题都与一系列图像配对。VILA-13B 在这个基准测试中表现最好，证明它在理解多模态信息方面的卓越能力，以及它在获取的知识中对语言偏差的抵抗能力。GQA 是一个专注于图像场景图的 VQA 数据集，提供来自真实世界图像的公正构图问题。每个问题都与其含义的结构化表示和回答所需的详细逻辑步骤相关联。LLaVA-1.5 和 VILA-7B 在该基准测试中表现最好，它们在该领域具有出色的推理能力。

接下来，我们将概述提高 MM-LLMs 有效性的培训配方，并从 SOTA 模型中得到对应。

首先，更高的图像分辨率可以为模型包含更多的视觉细节，有利于需要细粒度细节的任务。例如，LLaVA-1.5 和 VILA 的分辨率为336×336，而 Qwen-VL 和 MiniGPT-v2 的分辨率为448×448。然而，更高的分辨率会导致更长的令牌序列，从而产生额外的训练和推理成本。MiniGPT-v2 通过在嵌入空间中连接 4 个相邻的视觉标记来减少长度来解决这个问题。最近，Monkey 提出了一种无需重新训练高分辨率视觉编码器即可提高输入图像分辨率的解决方案，该方案仅使用低分辨率视觉编码器，支持分辨率高达1300×800。为了增强对富文本图像、表格和文档内容的理解，DocPedia 引入了一种将视觉编码器分辨率提高到2560×2560的方法，克服了开源 ViT 中低分辨率表现不佳的限制。其次，纳入高质量的 SFT 数据可以显著提高特定任务的性能，如将 ShareGPT4V 数据添加到 llva -1.5 和 VILA-13B 中，如表 1 所示。此外，VILA 揭示了几个关键发现：(1)在 LLM 主干上执行 PEFT 可以

表 1 主流 MM-LLMs 在 18 个 VL 基准测试上的比较。红色表示最高结果，蓝色表示第二高结果。‡表示 ShareGPT4 重新实现的测试结果，在基准测试或原始论文中遗漏。*表示在训练过程中观察到训练图像。

Model	LLM Backbone	OKVQA	IconVQA	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}	POPE	MME^{P}	MME^{C}	MMB	MMB^{CN}	SEED^{I}	LLaVA^{W}	MM-Vet	QBench	HM	VSR
Flamingo	Chinchilla-7B	44.7	–	–	–	28.8	–	–	–	–	–	–	–	–	–	–	–	57.0	31.8
BLIP-2	Flan-T5$_{XXL}$(13B)	45.9	40.6	65.0	44.7	19.6	61.0	42.5	85.3	1293.8	290.0	–	–	46.4	38.1	22.4	–	53.7	50.9
LLaVA	Vicuna-13B	54.4	43.0	–	41.3	–	–	38.9	–	–	–	–	–	–	–	–	–	–	51.2
MiniGPT-4	Vicuna-13B	37.5	37.6	–	30.8	–	–	19.4	–	–	–	–	–	–	–	–	–	–	41.6
InstructBLIP	Vicuna-7B	–	–	–	49.2	34.5	60.5	50.1	–	–	–	36.0	23.7	53.4	60.9	26.2	56.7	–	–
InstructBLIP	Vicuna-13B	–	44.8	–	49.5	33.4	63.1	50.7	78.9	1212.8	291.8	–	–	–	58.2	25.6	–	57.5	52.1
Shikra	Vicuna-13B	47.2	–	77.4*	–	–	–	–	–	–	–	58.8	–	–	–	–	54.7	–	–
IDEFICS-9B	LLaMA-7B	–	–	50.9	38.4	35.5	–	25.9	–	–	–	48.2	25.2	–	–	–	–	–	–
IDEFICS-80B	LLaMA-65B	–	–	60.0	45.2	36.0	–	30.9	–	–	–	54.5	38.1	–	–	–	–	–	–
Qwen-VL	Qwen-7B	–	–	78.8*	59.3*	35.2	67.1	63.8	–	–	–	38.2	7.4	56.3	–	–	59.4	–	–
Qwen-VL-Chat	Qwen-7B	–	–	78.2*	57.5*	38.9	68.2	61.5	–	1487.5	360.7	60.6	56.7	58.2	–	–	–	–	–
LLaVA-1.5	Vicuna-1.5-7B	–	–	78.5*	62.0*	50.0	66.8	58.2	85.9	1510.7	316.1‡	64.3	58.3	58.6	63.4	30.5	58.7	–	–
+ShareGPT4V	Vicuna-1.5-7B	–	–	80.6	–	57.2	68.4	–	–	1567.4	376.4	68.8	62.2	69.7	72.6	37.6	63.4	–	–
LLaVA-1.5	Vicuna-1.5-13B	–	–	80.0*	63.3*	53.6	71.6	61.3	85.9	1531.3	295.4‡	67.7	63.6	61.6	70.7	35.4	62.1	–	–
MiniGPT-v2	LLaMA-2-Chat-7B	56.9	47.7	–	60.3	30.3	–	51.9	–	–	–	–	–	–	–	–	–	58.2	60.6
MiniGPT-v2-Chat	LLaMA-2-Chat-7B	55.9	49.4	–	58.8	42.4	–	52.3	–	–	–	–	–	–	–	–	–	59.5	63.3
VILA-7B	LLaMA-2-7B	–	–	79.9*	62.3*	57.8	68.2	64.4	85.5	1533.0	–	68.9	61.7	61.1	69.7	34.9	–	–	–
VILA-13B	LLaMA-2-13B	–	–	80.8*	63.3*	60.6	73.7	66.6	84.2	1570.1	–	70.3	64.3	62.8	73.0	38.8	–	–	–
+ShareGPT4V	LLaMA-2-13B	–	–	80.6*	63.2*	62.4	73.1	65.3	84.8	1556.5	–	70.8	65.4	61.4	78.4	45.7	–	–	–

促进深度嵌入对齐，这对 ICL 至关重要；(2)交错的图像-文本数据被证明是有益的，而单独的图像-文本对是次优的；(3)在 SFT 过程中，将纯文本指令数据（如非自然指令）与图像文本数据重新混合，不仅解决了纯文本任务的退化问题，还提高了 VL 任务的准确性。

6 未来发展方向

在本节中，我们从以下几个方面探讨了 MM-LLMs 的未来发展方向。

6.1 更通用和智能的模型

增强 MM-LLMs 的能力可从以下四个主要途径：(1)扩展模式：目前 MM-LLMs 主要支持图像、视频、音频、3D 和文本等模式。然而，现实世界涉及更广泛的模式。扩展 MM-LLMs 以适应额外的模式（例如，网页、热图和图表）将增加模型的多功能性，使其更普遍适用；(2)LLMs 多样化：纳入不同类型和规模的 LLMs，使使用者能够根据自己的具体要求灵活选择最合适的 LLM；(3)提高 MM IT 数据集质量：当前 MM IT 数据集有很大的改进和扩展空间。多样化的指令范围可以提高 MM-LLMs 理解和执行用户命令的有效性；(4)增强 MM 生成能力：目前大多数 MM-LLMs 主要面向 MM 理解。尽管一些模型包含了生成 MM 的能力，但是生成的响应的质量可能受到 LDMs 能力的限制。探索基于检索的方法的整合在补充生成过程，提高模型的整体性能方面具有重要的前景。

6.2 更具挑战性的基准

现有的基准测试可能不足以挑战 MM-LLMs 的能力，因为许多数据集在 PT 或 IT 集中出现了不同程度的变化。这意味着模型可能在训练期间已经学会了这些任务。此外，目前的基准测试主要集中在 VL 子域。因此，对于 MM-LLMs 的发展来说，构建一个更具挑战性、更大规模的基准是至关重要的，该基准包括更多的模式，并采用统一的评估标准。例如，GOAT-Bench[55]旨在评估各种 MM-LLMs 在识别和响应模因中描述的社会虐待的细微方面的能力。MM-Code[56]评估了在视觉丰富的环境下 MM-LLMs 的算法解决问题的能力。DecodingTrust[57]测量 MM-LLMs 的可信度。MathVista[58]在视觉环境下评估 MM-LLMs 的数学推理能力。此外，MMMU[59]和 CMMMU[60]分别介绍了针对专家人工通用智能的综合多学科 MM 理解和推理基准的英文和中文版本。此外，Fan[61]等人用多面板 VQA 挑战了 MM-LLMs，而 BenchLMM[62]对 MM-LLMs 的跨风格视觉能力进行了基准测试。此外，Liu[63]等人对 MM-LLMs 的光学字符识别能力进行了深入研究。这些努力突出了需要更复杂和多样化的基准来真正衡量 MM-LLMs 的先进能力。

6.3 移动/轻量级的部署

要在资源受限的平台上部署 MM-LLMs，同时实现最优性能，如低功耗移动和物联网设备，轻量级实现至关重要。MobileVLM[64]是该领域的一个显著进步。这种方法战略性地缩小了 LLaMA 的规模，实现了无缝的现成部署。MobileVLM 还推出了一款轻量级下采样投影仪，由不到 2000 万个参数组成，

有助于提高计算速度。最近，有许多类似的轻量化 MM-LLMs 的研究，以相当的性能或最小的损失实现了高效的计算和推理，包括 TinyGPT-V[65]、Vary-toy[66] 、 Mobile-Agent[67] 、 MoE-LLaVA[68] 和 MobileVLM V2[69]。然而，这一途径需要进一步探索，以进一步促进发展。

6.4 嵌入式智能

嵌入式智能旨在通过有效地理解环境、识别相关物体、评估它们的空间关系和制定全面的任务计划来复制人类的感知和与周围环境的互动。嵌入式 AI 任务，如嵌入式规划、嵌入式视觉问答和嵌入式控制，通过利用实时观察，使机器人能够自主执行扩展计划。该领域的典型作品有 PaLM-E[70]和 EmbodiedGPT[71]。PaLM-E 通过 MM-LLM 的培训引入了一种多体代理。除了作为一个具体的决策者之外，PaLM-E 还展示了处理一般 VL 任务的熟练程度。EmbodiedGPT 引入了一种经济有效的方法，以 CoT 方法为特征，增强实体代理与现实世界交互的能力，建立一个将高层次规划与低层次控制连接起来的闭环。基于 MM-LLMs 的嵌入式智能在与机器人集成方面取得了进展，但在增强机器人自主性方面还需进一步探索。

6.5 持续学习

由于其庞大的规模带来了巨大的培训成本，MM-LLMs 不适合频繁的再培训。然而，更新是必要的，以赋予 MM-LLMs 新的技能，并使其与快速发展的人类知识保持同步。因此，需要持续学习（CL）使模型足够灵活，以有效和持续地利用新出现的数据，同时避免重新培训 MM-LLMs 的大量成本。MM-LLMs 的 CL 可分为连续 PT 和连续 IT 两个阶段。最近，已经提出了一个连续的 MM IT 基准，以持续地对 MM-LLMs 进行新的 MM 任务微调，同时在原始 MM IT 阶段学习的任务上保持优越的性能。它引入了两个主要挑战：(1)灾难性遗忘，即模型在学习新任务时忘记了之前的知识；(2)负前向迁移，表明学习新任务时未见任务的表现下降。

6.6 缓解幻觉现象

幻觉需要在没有视觉线索的情况下对不存在的物体产生文本描述，这表现在不同的类别中，例如描述中的误判和不准确。这些幻觉的来源是多方面，包括训练数据中的偏差和注释错误。此外，还存在与段落分隔符相关的语义漂移偏差，在故意插入时会引起幻觉。目前缓解这些幻觉的方法包括利用自我反馈作为视觉线索。然而，挑战仍然存在，需要在准确输出和幻觉输出之间进行细微的区分，以及改进训练方法以提高输出的可靠性。

6.7 偏见和伦理考虑

尽管 MM-LLMs 具有优势，但确保其安全高效的应用仍然至关重要。MM-LLMs 产生的信息可能使刻板印象永久化，并对弱势群体造成伤害。由于 MM-LLMs 从 MM 训练数据中的模式中学习，它们可以重现这些数据中存在的偏差，从而潜在地导致代表性损害。为了解决这个问题，我们可以开发专门用于评估 MM-LLMs 偏差的新基。此外，设计更有效和细粒度的对齐方法是必要的。例如，使用 RLHF[72]可以帮助校准 MM-LLMs，以产生符合人类价值观和愿望的答案。

7 结论

在本文中，我们对 MM-LLMs 的最新进展进行了全面的调查。最初，我们将模型体系结构分为五个组件，提供一般设计公式和培训管道的详细概述。随后，我们介绍了各种 SOTA MM-LLMs，每种 MM-LLMs 都有其特定的配方。我们的调查还揭示了他们在不同 MM 基准上的能力，并展望了这个快速发展领域的未来发展。我们希望这项调查能够为研究者提供一些见解，为 MM-LLMs 领域的持续发展做出贡献。

参 考 文 献

[1] Gemini Team Google, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H.Hoi. 2023e. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, pages 19730–19742.

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023e.Visual Instruction Tuning. In Thirtyseventh Conference on Neural Information Processing Systems.

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

[5] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy,

Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al.2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390.

[6] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023f. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355.

[7] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv preprint arXiv:2306.05424.

[8] Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023j. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. arXiv preprint arXiv:2311.17043.

[9] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023b. Qwen-audio: Advancing universal audio understanding via unified large-scale audiolanguage models. arXiv preprint arXiv:2311.07919.

[10] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov.2023a. Generating images with multimodal language models. In Thirty-seventh Conference on Neural Information Processing Systems.

[11] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei.2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. arXiv preprint arXiv:2306.14824.

[12] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024.Generative pretraining in multimodality. In The Twelfth International Conference on Learning Representations.

[13] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023b.Minigpt-5: Interleaved vision-and-language generation via generative vokens. arXiv preprint arXiv:2310.02239.

[14] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a.SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities.In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pages 15757–15773.

[15] Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al 2023.AudioPaLM: A Large Language Model That Can Speak and Listen. arXiv preprint arXiv:2306.12925.

[16] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a.Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671.

[17] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580.

[18] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al.2023b. Audiogpt: Understanding and generating speech, music, sound, and talking head. arXiv preprint arXiv:2304.12995.

[19] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023d. Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519.

[20] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023c. Any-to-Any Generation via Composable Diffusion. In Thirty-seventh Conference on Neural Information Processing Systems.

[21] Xinyu Wang, Bohan Zhuang, and Qi Wu. 2024d. ModaVerse: Efficiently Transforming Modalities with LLMs. arXiv preprint arXiv:2401.06395.

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

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations.

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

[25] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. 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.

[26] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al 2023d. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19175–19186.

[27] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2818–2829.

[28] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. 2022b. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In The Eleventh International Conference on Learning Representations.

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al 2023. Segment anything. arXiv preprint arXiv:2304.02643.

[31] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al 2023i. Recognize Anything: A Strong Image Tagging Model. arXiv preprint arXiv:2306.03514.

[32] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al 2023j. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238.

[33] Jitesh Jain, Jianwei Yang, and Humphrey Shi. 2023.Vcoder: Versatile vision encoders for multimodal large language models. arXiv preprint arXiv:2312.14233.

[34] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023b. Xllm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv preprint arXiv:2305.04160.

[35] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460.

[36] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023g. BEATs: Audio Pre-Training with Acoustic Tokenizers. In International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, pages 5178–5193.

[37] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, pages 28492–28518.

[38] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023e. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE.

[39] Salesforce. 2022. https://github.com/salesforce/ULIP

[40] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. 2022. Point-bert: Pretraining 3d point cloud transformers with masked point modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19313 – 19322.

[41] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15180–15190.

[42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H.Hoi. 2023e. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, pages 19730–19742.

[43] Yiren Jian, Chongyang Gao, and Soroush Vosoughi.2023. Bootstrapping Vision-Language Learning with Decoupled Language Pre-training. In Thirty-seventh Conference on Neural Information Processing Systems.

[44] Junyu Lu, Ruyi Gan, Dixiang Zhang, Xiaojun Wu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. 2023a. Lyrics: Boosting Finegrained Language-Vision Alignment and Comprehension via Semantic-aware Visual Objects. arXiv preprint arXiv:2312.05278.

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

[46] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4582– 4597.

[47] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al 2021. LoRA: Low-Rank Adaptation of Large Language Models. In International Conference on Learning Representations.

[48] Bingchen Zhao, Haoqin Tu, Chen Wei, and Cihang Xie.2024. Tuning LayerNorm in Attention: Towards Efficient Multimodal LLM Finetuning. In The Twelfth International Conference on Learning Representations.

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. Highresolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695.

[50] Cerspense. 2023. https://huggingface.co/cerspense

[51] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D.Plumbley. 2023b. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In International Conference on Machine Learning, ICML 2023, 2329 July 2023, Honolulu, Hawaii, USA, pages 21450– 21474.

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

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

arXiv:2212.12017.

[54] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi.2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning.In Thirty-seventh Conference on Neural Information Processing Systems.

[55] Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024b. GOAT-Bench: Safety Insights to Large Multimodal Models through Meme-Based Social Abuse. arXiv preprint arXiv:2401.01523.

[56] Kaixin Li, Yuchen Tian, Qisheng Hu, Ziyang Luo, and Jing Ma. 2024a. MMCode: Evaluating Multi-Modal Code Large Language Models with Visually Rich Programming Problems. ArXiv, abs/2404.09486.

[57] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al 2024a. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. Advances in Neural Information Processing Systems, 36.

[58] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, KaiWei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In The Twelfth International Conference on Learning Representations.

[59] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502.

[60] Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, et al 2024a. CMMMU: A Chinese Massive Multi-discipline Multimodal Understanding Benchmark. arXiv preprint arXiv:2401.11944.

[61] Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, and Xin Eric Wang. 2024. Muffin or Chihuahua? Challenging Large Vision-Language Models with Multipanel VQA. arXiv preprint arXiv:2401.15847.

[62] Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. 2023.BenchLMM: Benchmarking cross-style visual capability of large multimodal models. arXiv preprint arXiv:2312.02896.

[63] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al 2023h. On the hidden mystery of ocr in large multimodal models. arXiv preprint arXiv:2305.07895.

[64] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al 2023a. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886.

[65] Zhengqing Yuan, Zhaoxu Li, and Lichao Sun. 2023b.TinyGPT-V: Efficient Multimodal Large Language Model via Small Backbones. arXiv preprint arXiv:2312.16862.

[66] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. Small Language Model Meets with Reinforced Vision Vocabulary. arXiv preprint arXiv:2401.12503.

[67] Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang.2024c. Mobile-Agent: Autonomous Multi-Modal Mobile Device Agent with Visual Perception. arXiv preprint arXiv:2401.16158.

[68] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024a. MoE-LLaVA: Mixture of Experts for Large Vision-Language Models. arXiv preprint arXiv:2401.15947.

[69] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al 2024. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. arXiv preprint arXiv:2402.03766.

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

[71] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Visionlanguage pre-training via embodied chain of thought.In Thirty-seventh Conference on Neural Information Processing Systems.

[72] Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. 2024c. Red teaming visual language models. arXiv preprint arXiv:2401.12915.