A Survey on MLLM：IT，ICL & CoT

Paper Reading by Yiwei Sun

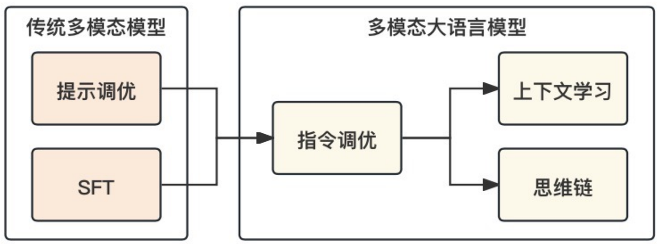
2024.03.12

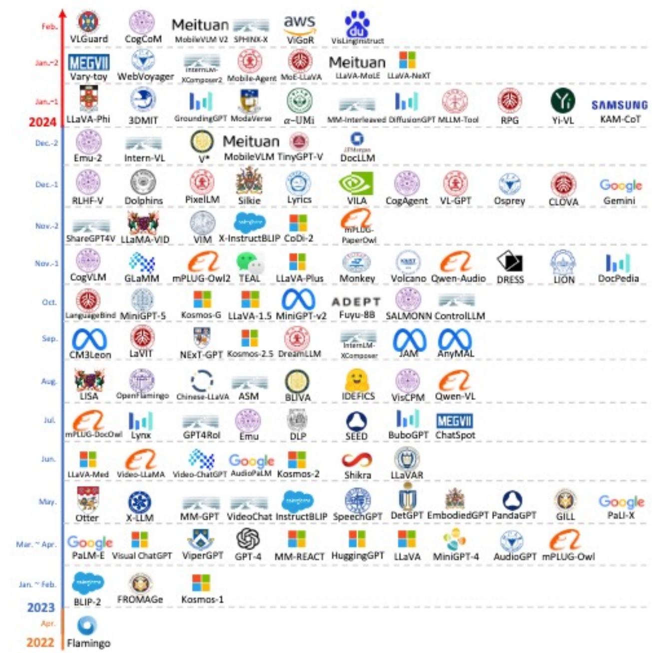
智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

- 研究背景
- 指令调优
- 上下文学习
- 思维链
- 总结反思

研究背景

对于通用多模态模型的需求带来了多模态大语言模型的研究热潮。





提示调优和SFT局限于特定的任务，并不能够赋予零样本学习的能力。因此，将LLM扩展至多模态成为研究的必然。

指令调优，上下文学习以及思维链技术属于扩展通用性的关键技术。其中指令调优是构建MLLM的基础。

https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models

- 研究背景
- 指令调优
- 上下文学习
- 思维链
- 总结反思

回顾LLaVA

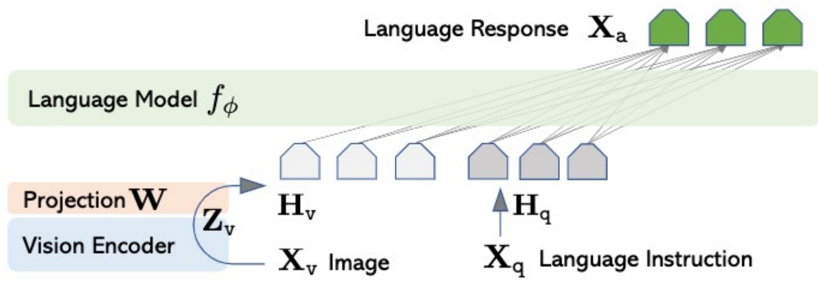


Figure 1: LLaVA network architecture.

Stage 1称为预训练阶段，用于对齐图文表征；
Stage 2称为微调阶段，用于对齐人类意图。

LLaVA的贡献：

1. MLLM的构建技术；
2. 自指令(Self-Instruct)技术

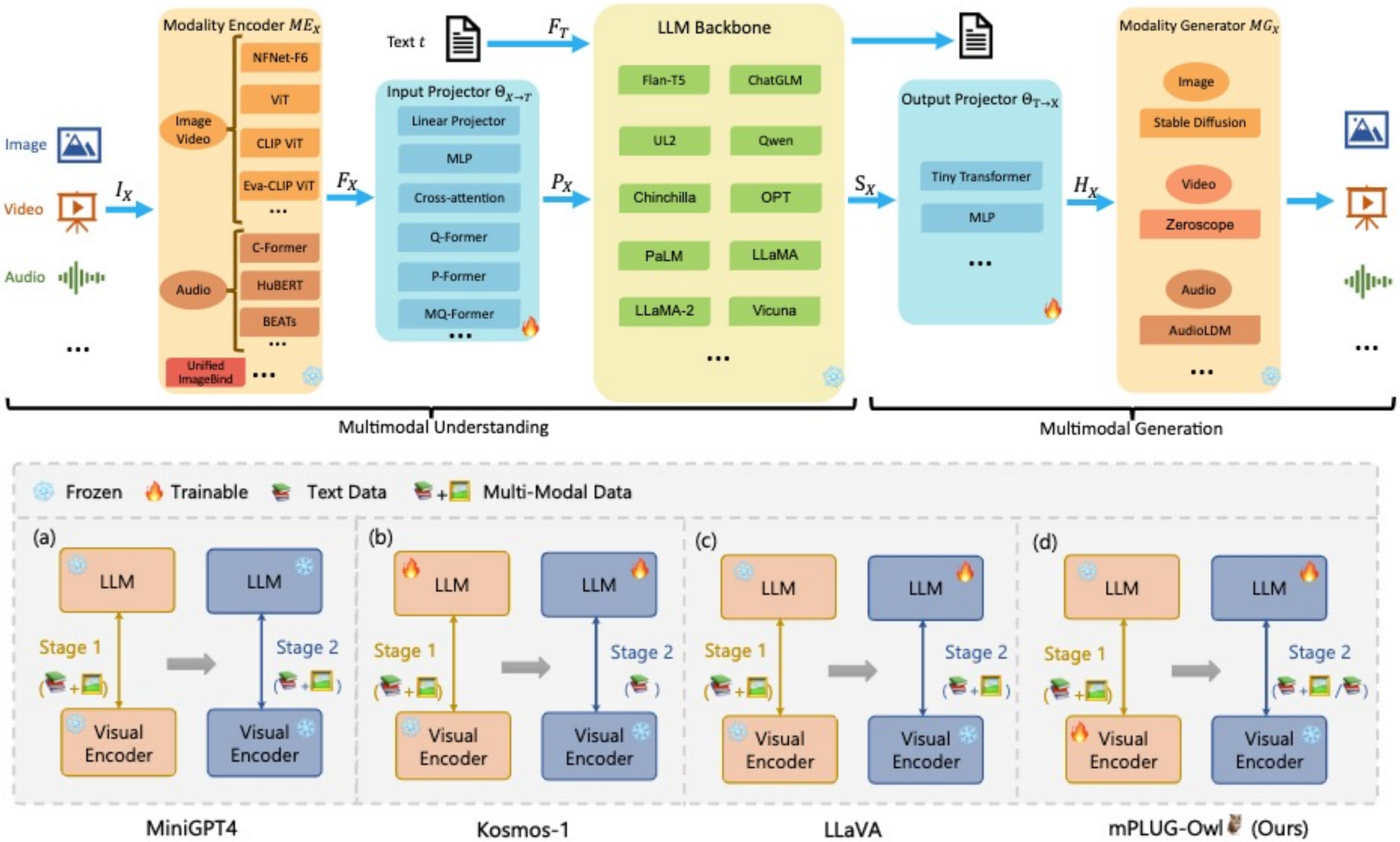
结构		参数
Vision Encoder		CLIP-L-224
Connector		Linear Projection
Language Model		LLaMA
Training Recipe	Stage 1	只微调LP
	Stage 2	同时微调LP和LM
Datasets	Stage 1	CC-595K
	Stage 2	LLaVA-Instruct-158K

SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions

指令微调数据集的两种制作方式

1. 原有数据集的重制：LLaMA-Adapter v2采用COCO Caption数据集微调；
2. 自指令技术：通过模型生成回答，如MiniGPT-4用一阶段的模型生成二阶段的回答或者LLaVA用GPT-4生成指令数据。

MLLM的范式



Modality Encoder ME_X
NFNet-F6
ViT
CLIP ViT
Eva-CLIP ViT
C-Former
HuBERT
BEATs
Unified ImageBind
Image Video
Audio
Image
Video
Audio
I_X
F_X
Text t
F_T
Input Projector $\Theta_{X\to T}$
Linear Projector
MLP
Cross-attention
Q-Former
P-Former
MQ-Former
P_X
LLM Backbone
Flan-T5
ChatGLM
UL2
Qwen
Chinchilla
OPT
PaLM
LLaMA
LLaMA-2
Vicuna
S_X
Output Projector $\Theta_{T\to X}$
Tiny Transformer
MLP
H_X
Modality Generator MG_X
Image
Stable Diffusion
Video
Zeroscope
Audio
AudioLDM
Multimodal Understanding
Multimodal Generation
Frozen
Trainable
Text Data
Multi-Modal Data
(a)
(b)
(c)
(d)
LLM
Stage 1
Stage 2
Visual Encoder
MiniGPT4
Kosmos-1
LLaVA
mPLUG-Owl (Ours)

LLaVA-v1.5：让MLLM更强

结构		参数（LLaVA）	参数（LLaVA-v1.5）
Vision Encoder		CLIP-L-224	CLIP-L-336
Connector		Linear Projection	MLP-2x
Language Model		LLaMA	Vicuna
Training Recipe	Stage 1	只微调LP	只微调LP
	Stage 2	同时微调LP和LM	同时微调LP和LM
Datasets	Stage 1	CC-595K	LCS-558K
	Stage 2	LLaVA-Instruct-158K	LLaVA-1.5-mix-665K

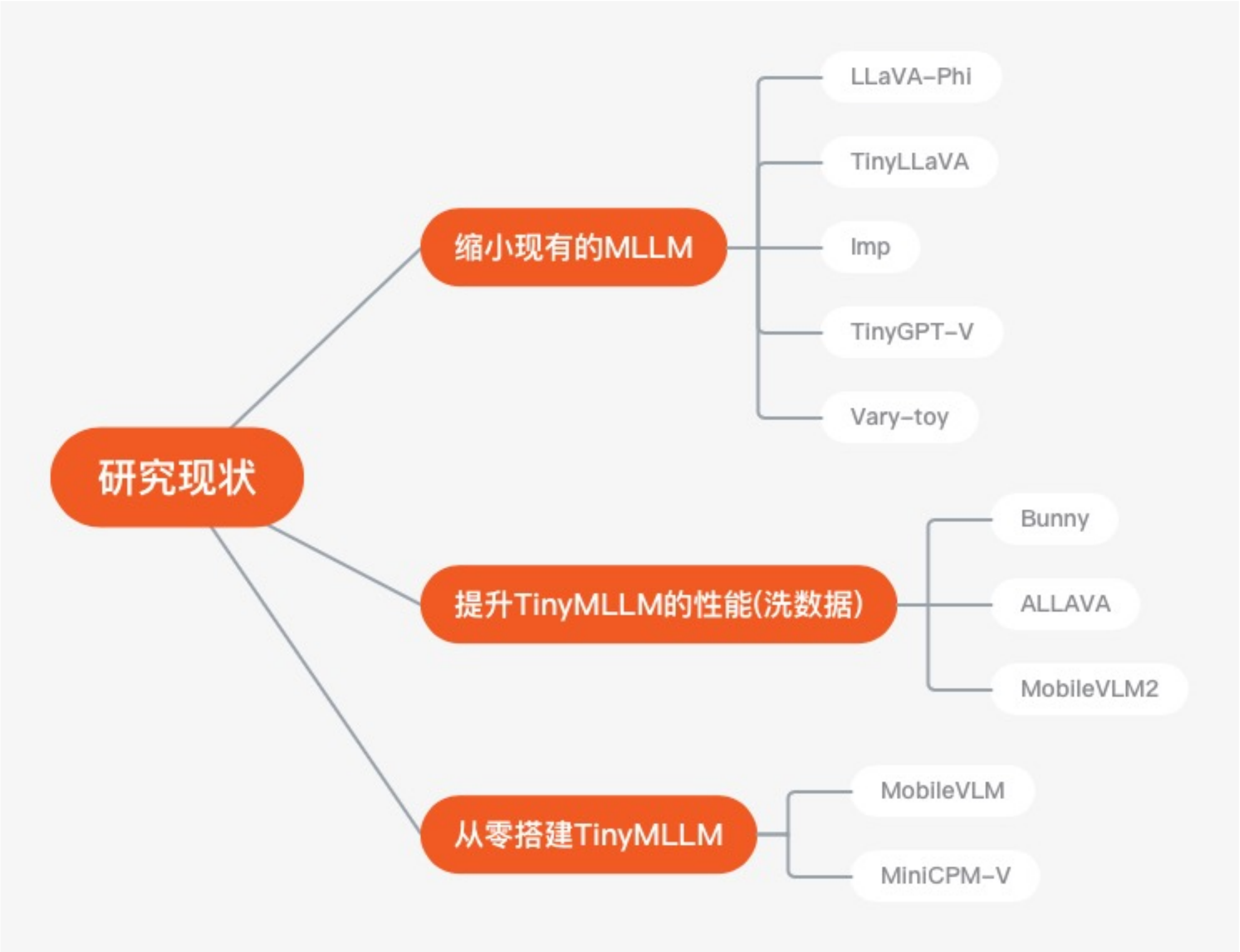
	Method	LLM	Res.	GQA	MME	MM-Vet
	InstructBLIP	14B	224	49.5	1212.8	25.6
Only using a subset of InstructBLIP training data						
0	**LLaVA**	7B	224	–	502.8	23.8
1	+VQA-v2	7B	224	47.0	1197.0	27.7
2	+Format prompt	7B	224	46.8	1323.8	26.3
3	+MLP VL connector	7B	224	47.3	1355.2	27.8
4	+OKVQA/OCR	7B	224	50.0	1377.6	29.6
Additional scaling						
5	+Region-level VQA	7B	224	50.3	1426.5	30.8
6	+Scale up resolution	7B	336	51.4	1450	30.3
7	+GQA	7B	336	62.0*	1469.2	30.7
8	+ShareGPT	7B	336	62.0*	1510.7	30.5
9	+Scale up LLM	13B	336	**63.3***	**1531.3**	**36.3**

Data	Size	Response formatting prompts
LLaVA [28]	158K	–
ShareGPT [38]	40K	–
VQAv2 [12]	83K	Answer the question using a single word or phrase.
GQA [14]	72K	
OKVQA [33]	9K	
OCRVQA [34]	80K	
A-OKVQA [37]	50K	Answer with the option's letter from the given choices directly.
TextCaps [39]	22K	Provide a one-sentence caption for the provided image.
RefCOCO [17, 32]	30K	*Note: randomly choose between the two formats* Provide a short description for this region.
VG [18]	86K	Provide the bounding box coordinate of the region this sentence describes.
Total	665K	

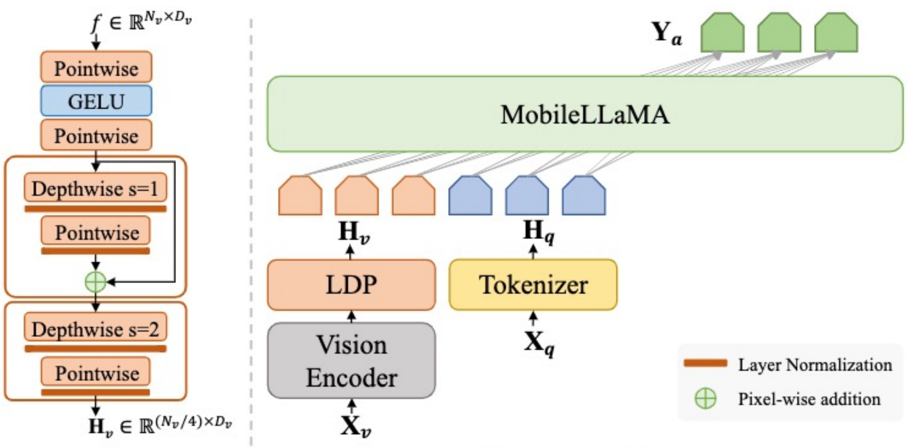
TinyMLLM：将MLLM缩小

名称	LLM Backbone	论文
LLaVA-Phi	Phi-2	LLaVA-phi: Efficient Multi-Modal Assistant with Small Language Model
TinyLLaVA	Phi-2	TinyLLaVA: A Framework of Small-scale Large Multimodal Models
Imp	Phi-2	https://github.com/MILVLG/imp
Bunny	Phi-2	Efficient Multimodal Learning from Data-centric Perspective
TinyGPT-V	Phi-2	TinyGPT-V: Efficient Multimodal Large Language Model via Small Backbones
ALLAVA	Phi-2	HARNESSING GPT4V-SYNTHESIZED DATA FOR A LITE VISION-LANGUAGE MODEL
MobileVLM	MobileLLaMA	MobileVLM: A Fast, Strong and Open Vision Language Assistant for Mobile Devices
MobileVLM2	MobileLLaMA	MobileVLM V2: Faster and Stronger Baseline for Vision Language Model
MiniCPM-V	MiniCPM	MiniCPM: 揭示端侧大语言模型的无限潜力
Vary-toy	Qwen	Small Language Model Meets with Reinforced Vision Vocabulary

TinyMLLM：将MLLM缩小



MobileVLM



LDP   MobileVLM architecture

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2: **Model sizes, architectures, and optimization hyper-parameters.**

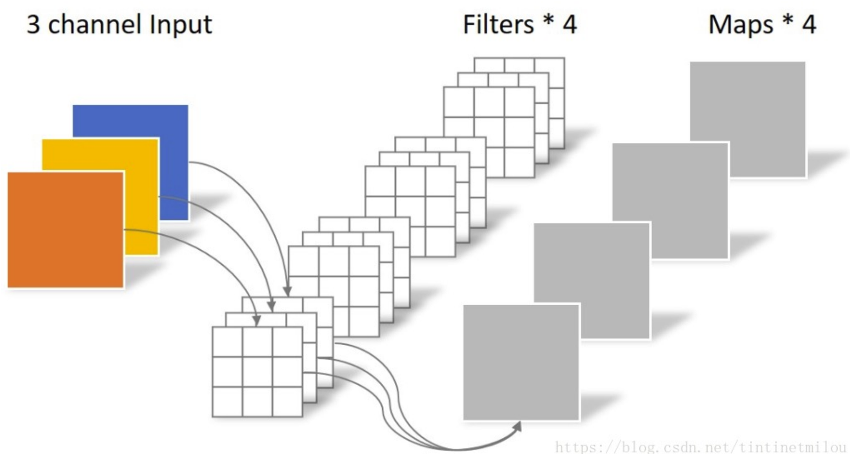
Model	Blocks	Dim	Heads	Context length
MobileLLaMA 1.4B	24	2048	16	2k
MobileLLaMA 2.7B	32	2560	32	2k

结构		参数
Vision Encoder		CLIP-L-336
Connector		LDP
Language Model		MobileLLaMA
Training Recipe	Stage 1	只微调LP
	Stage 2	同时微调LP和LM
Datasets	Stage 1	LCS-558K
	Stage 2	LLaVA-1.5-mix-665K

结构	优势	劣势
Q-Former	控制令牌数量，提取强相关的视觉信息。	丢失令牌的空间位置信息且收敛缓慢。
MLP	保留空间信息。	存在冗余信息，如背景。

MobileVLM

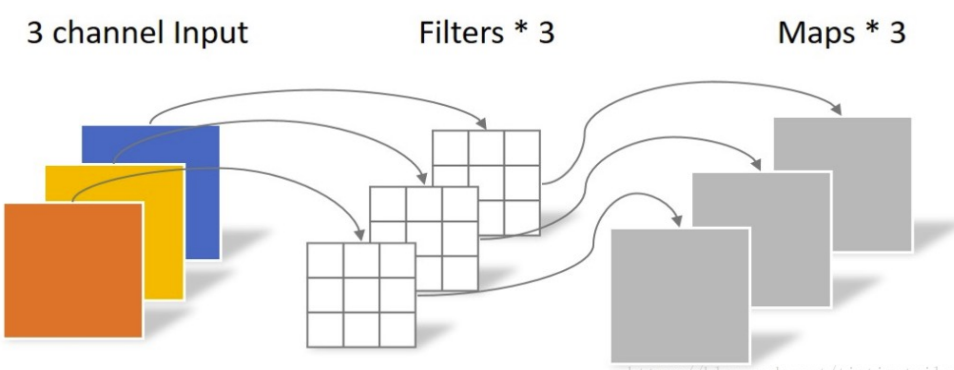
一般的卷积操作：



卷积	参数量
Std	3*(3*3)*4=108
DW	1*(3*3)*3=27
PW	3*(1*1)*4=12

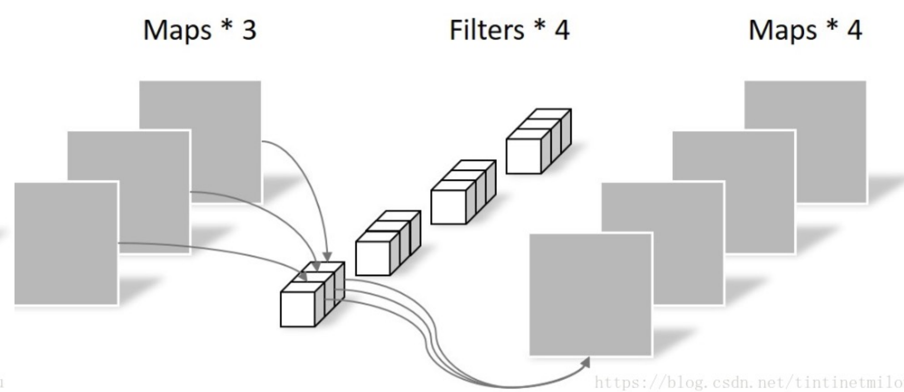
参数量和运算成本低，常用于轻量级网络

DW卷积：



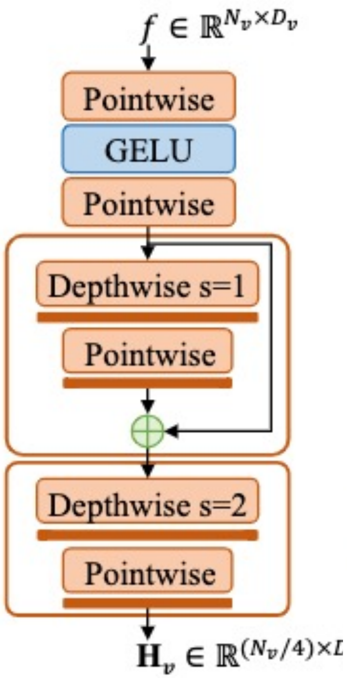
一个卷积核负责一个通道，输入输出通道数相同

PW卷积：



1*1卷积

MobileVLM



LDP

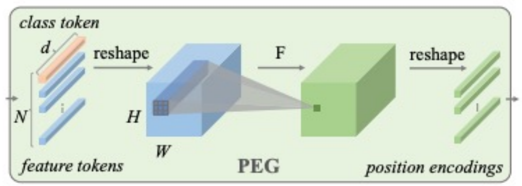


Figure 2. Schematic illustration of Positional Encoding Generator (PEG). Note d is the embedding size, N is the number of tokens.

DW卷积的作用：
1. 缩小令牌带来的资源开销；
2. CNN对于边缘设备十分友好；
3. 能够增强位置信息且鼓励局部的交互。

单纯的PEG结构严重损害了模型在下游任务的性能表现。

VL Projector Architecture Design	Tokens	GQA	SQA^{I}	VQA^{T}	POPE	MME	MMB^{dev}
$[PW]_{\times 2}[DW^{\kappa=1}PW]_{\times 0}[DW^{\kappa=2}PW]_{\times 0}$	576	56.9	53.6	43.7	85.7	1137.7	52.8
$[PW]_{\times 0}[DW^{\kappa=1}PW]_{\times 1}[DW^{\kappa=2}PW]_{\times 1}$	144	54.9	52.9	40.2	84.0	1150.8	50.3
$[PW]_{\times 2}[DW^{\kappa=1}PW]_{\times 1}[DW^{\kappa=2}PW]_{\times 1}$	144	56.1	54.7	41.5	84.5	1196.2	53.2
$[PW]_{\times 2}[DW^{\kappa=1}PW]_{\times 3}[DW^{\kappa=2}PW]_{\times 1}$	144	55.3	53.9	40.8	84.6	1166.3	53.0
$[PW]_{\times 2}[DW^{\kappa=2}PW]_{\times 1}[DW^{\kappa=1}PW]_{\times 1}$	144	55.6	54.3	41.5	84.6	1166.2	52.8

第一行是LLaVA的Connector；第二行在PW之前增加DW，性能下降；第三行额外增加两个PW，增强特征级的交互。

MobileVLM

Method	LLM	Res.	PT	IT	GQA	SQA^I	VQA^T	POPE	MME	MMB^{dev}
Openflamingo [3]	MPT-7B	336	180M	-	–	–	33.6	–	–	4.6
BLIP-2 [66]	Vicuna-13B	224	129M	-	41.0	61.0	42.5	85.3	1293.8	–
MiniGPT-4 [133]	Vicuna-7B	224	5M	5K	32.2	–	–	–	581.7	23.0
InstructBLIP [30]	Vicuna-7B	224	129M	1.2M	49.2	60.5	50.1	–	–	36.0
InstructBLIP [30]	Vicuna-13B	224	129M	1.2M	49.5	63.1	50.7	78.9	1212.8	–
Shikra [15]	Vicuna-13B	224	600K	5.5M	–	–	–	–	–	58.8
mPLUG-Owl [126]	LLaMA-7B	224	2.1M	102K	–	–	–	–	967.3	49.4
IDEFICS-9B [64]	LLaMA-7B	224	353M	1M	38.4	–	25.9	–	–	48.2
IDEFICS-80B [64]	LLaMA-65B	224	353M	1M	45.2	–	30.9	–	–	54.5
Qwen-VL [5]	Qwen-7B	448	1.4B	50M	59.3	67.1	63.8	–	1487.6	38.2
MiniGPT-v2 [14]	LLaMA-7B	448	23M	1M	60.3	–	–	–	–	12.2
LLaVA-1.5 [74]	Vicuna-7B	336	558K	665K	62.0	66.8	58.2	85.9	1510.7	64.3
MobileVLM 1.7B	MobileLLaMA 1.4B	336	558K	665K	56.1	54.7	41.5	84.5	1196.2	53.2
MobileVLM 1.7B w/ LoRA	MobileLLaMA 1.4B	336	558K	665K	57.0	53.1	42.3	86.0	1143.7	50.4
MobileVLM 3B	MobileLLaMA 2.7B	336	558K	665K	59.0	61.0	47.5	84.9	1288.9	59.6
MobileVLM 3B w/ LoRA	MobileLLaMA 2.7B	336	558K	665K	58.4	59.0	46.7	84.6	1296.4	57.0

Bunny

结构		参数
Vision Encoder		SigLIP-384
Connector		MLP 2x
Language Model		Phi-2
Training Recipe	Stage 1	只微调LP
	Stage 2	同时微调LP和LM
Datasets	Stage 1	Bunny-pretrain-LAION-2M
	Stage 2	Bunny-695K

Bunny-pretrain-LAION-2M：

1. a. 执行k-means算法对2B图像嵌入进行聚类；在每个集合中构建一个无向图：如果两个嵌入的余弦相似度高于阈值，则连接；b. 对于每个连通子图，只保留唯一样本，该样本距离簇质心的欧几里得距离在中位数左右；
2. 根据图像嵌入与文本嵌入之间的余弦相似度排序，保留排名在40%~60%之间的样本；
3. 通过图像嵌入和质心嵌入之间的余弦相似度，保留排名在15%~30%的样本。

得到的样本既是多样的又反应类本质。

Bunny-695K: SVIT-mix-665K – ShareGPT-40K + WizardLM-evol-instruct-70K

增加纯文本数据，避免LLM的能力退化。

Bunny

Model	Vision Encoder	LLM	MME^P	MME^C	MMB^T	MMB^D	SEED	$MMMU^V$	$MMMU^T$	VQA^{v2}	GQA	SQA^I	POPE
IDEFICS-80B [43]	OpenCLIP-H (1.0B)	LLaMA-65B	–	–	54.6	54.5	–	–	–	60.0	–	68.9	–
BLIP-2 [24]	EVA01-CLIP-G (1.0B)	Vicuna-13B	–	–	–	–	–	–	–	–	41.0	61.0	–
InstructBLIP [44]	EVA01-CLIP-G (1.0B)	Vicuna-13B	–	–	–	–	–	–	–	–	49.5	63.1	83.7
BLIP-2 [24]	EVA01-CLIP-G (1.0B)	Flan-T5-XXL (11B)	1293.8	290.0	–	–	–	35.4	34.0	65.0	44.6	64.5	–
InstructBLIP [44]	EVA01-CLIP-G (1.0B)	Flan-T5-XXL (11B)	1212.8	291.8	–	–	–	35.7	33.8	–	47.9	70.6	–
Shikra-13B [5]	CLIP-L (0.4B)	Vicuna-13B	–	–	–	–	–	–	–	77.4	–	–	–
LLaVA-v1.5-13B (LoRA) [26]	CLIP-L (0.4B)	Vicuna-13B	1541.7	300.4§	68.4§	68.5	61.3	40.0§	33.2§	80.0	63.3	71.2	86.7
InstructBLIP [44]	EVA01-CLIP-G (1.0B)	Vicuna-7B	–	–	33.9	36.0	53.4	–	–	–	49.2	60.5	–
MiniGPT-v2 [28]	EVA01-CLIP-G (1.0B)	LLaMA2-7B	–	–	–	–	–	–	–	–	60.3	–	–
IDEFICS-9B [43]	OpenCLIP-H (1.0B)	LLaMA-7B	–	–	45.3	48.2	–	–	–	50.9	–	44.2	–
LLaVA-v1.5-7B (LoRA) [26]	CLIP-L (0.4B)	Vicuna-7B	1476.9	267.9§	66.1§	66.1	60.1	34.4§	31.7§	79.1	**63.0**	68.4	86.4
mPLUG-Owl2 [45]	CLIP-L (0.4B)	LLaMA2-7B	1450.2	**313.2**	66.0	66.5	57.8	32.7	32.1	79.4	56.1	68.7	85.8
Shikra-7B [5]	CLIP-L (0.4B)	Vicuna-7B	–	–	60.2	58.8	–	–	–	–	–	–	–
TinyGPT-V [29]	EVA01-CLIP-G (1.0B)	Phi-2 (2.7B)	–	–	–	–	–	–	–	–	33.6	–	–
MobileVLM [15]	CLIP-L (0.4B)	MobileLLaMA (2.7B)	1288.9	–	–	59.6	–	–	–	–	59.0	61.0	84.9
LLaVA-Phi [9]	CLIP-L (0.4B)	Phi-2 (2.7B)	1335.1	–	–	59.8	–	–	–	71.4	–	68.4	85.0
MC-LLaVA [46]	SigLIP-SO (0.4B)	Dolphin 2.6 Phi-2 (2.7B)	–	–	–	–	–	–	–	64.2	49.6	–	80.6
Imp-v1 [10]	SigLIP-SO (0.4B)	Phi-2 (2.7B)	1434.0	–	–	66.5	–	–	–	79.5	58.6	70.0	**88.0**
MiniCPM-V [16]	SigLIP-SO (0.4B)	MiniCPM (2.4B)	1446.0	–	–	67.3	–	34.7	–	–	–	–	–
Moondream1 [47]	SigLIP-SO (0.4B)	Phi-1.5 (1.3B)	–	–	–	–	–	–	–	74.3	56.3	–	–
TinyLLaVA-v1 [48]	CLIP-L (0.4B)	TinyLlama (1.1B)	–	–	–	–	–	–	–	73.4	57.5	59.4	–
Bunny	SigLIP-SO (0.4B)	Phi-2 (2.7B)	**1488.8**	289.3	**69.2**	**68.6**	**62.5**	**38.2**	**33.0**	**79.8**	62.5	**70.9**	86.8

- 研究背景
- 指令调优
- 上下文学习
- 思维链
- 总结反思

上下文学习

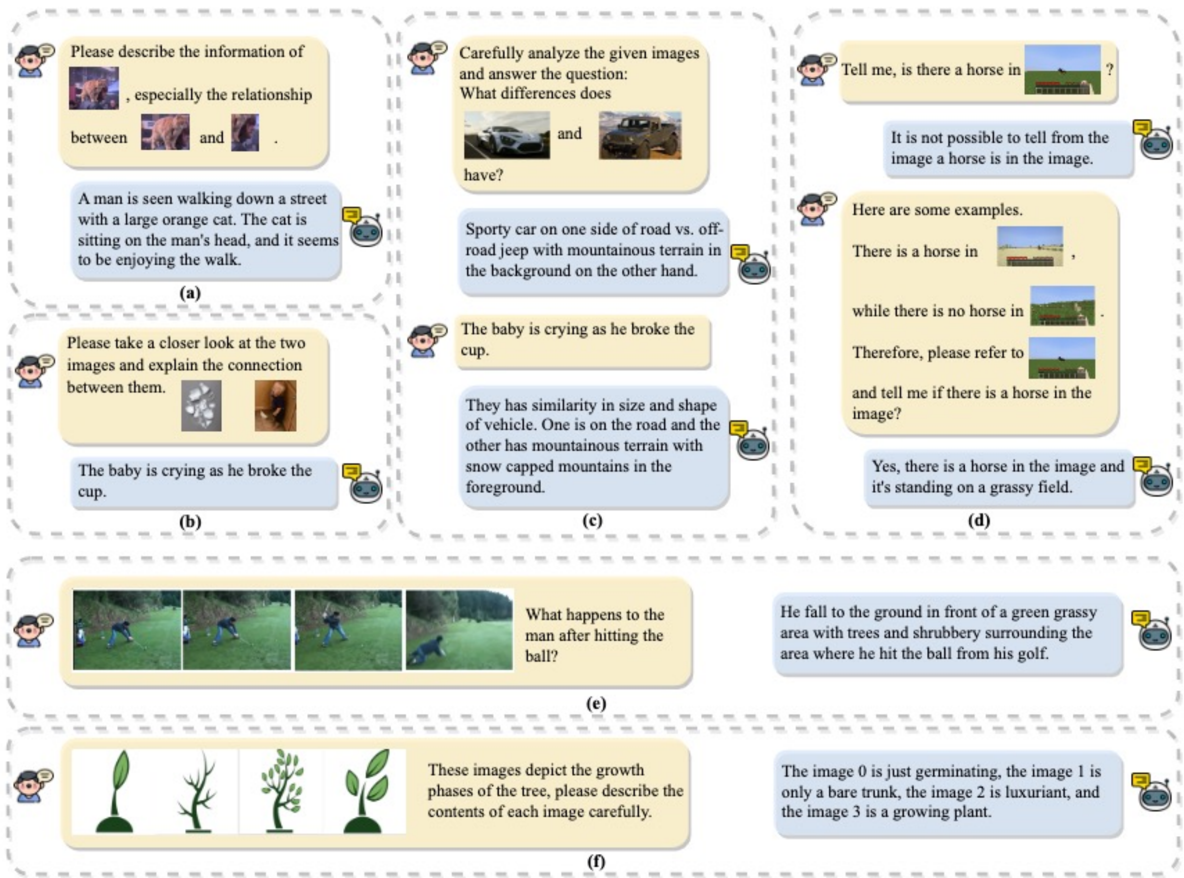
MMICL: EMPOWERING VISION-LANGUAGE MODEL WITH MULTI-MODAL IN-CONTEXT LEARNING

Haozhe Zhao*[1], Zefan Cai*[1], Shuzheng Si*[1], Xiaojian Ma[2], Kaikai An[1], Liang Chen[1], Zixuan Liu[3], Sheng Wang[3], Wenjuan Han†[4], Baobao Chang†[1]

因为缺乏interleaved dataset， 目前大部分的MLLM缺乏对多图提示词的理解。这也导致它们并没有很强的ICL能力。

目前数据集存在的缺陷：

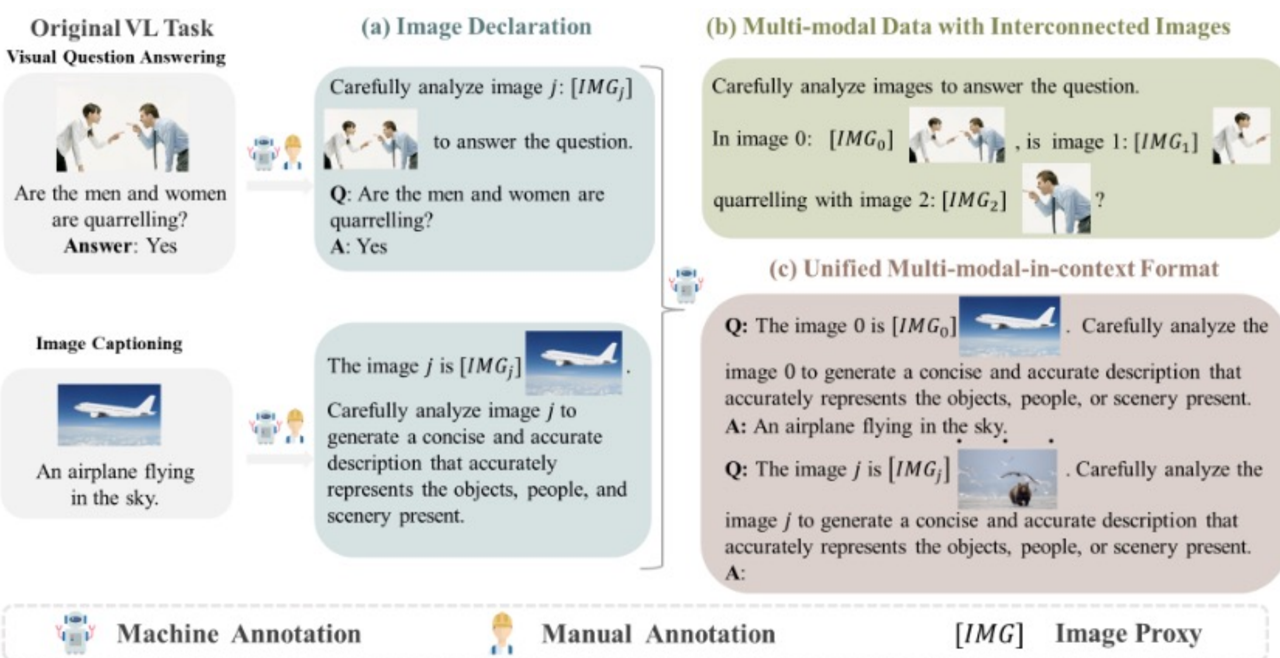
1. 很少涉及文本到图像的参考，即文本与图像之间存在的复杂指称关系。这使得VLM无法处理文本对图像的复杂查询；
2. 缺少多图交互提示词，使得图与图之间存在的空间，时间和逻辑关系被忽略，限制了VLM理解图像之间复杂关系的能力；
3. 缺少高质量的上下文数据集，限制了VLM的上下文学习能力。



上下文学习

MMICL: EMPOWERING VISION-LANGUAGE MODEL WITH MULTI-MODAL IN-CONTEXT LEARNING

Haozhe Zhao*[1], Zefan Cai*[1], Shuzheng Si*[1], Xiaojian Ma[2], Kaikai An[1], Liang Chen[1], Zixuan Liu[3], Sheng Wang[3], Wenjuan Han†[4], Baobao Chang†[1]



a. 创建图像代理[IMGj]用于指代视觉嵌入；增加文本指代。
b. 根据视频数据或者通过从单一图像抠图（目标检测算法）；将文本指称替换为图像。
c. 通过从数据中采样得到上下文。

上下文学习

MMICL: EMPOWERING VISION-LANGUAGE MODEL WITH MULTI-MODAL IN-CONTEXT LEARNING

Haozhe Zhao*[1], Zefan Cai*[1], Shuzheng Si*[1], Xiaojian Ma[2], Kaikai An[1], Liang Chen[1], Zixuan Liu[3], Sheng Wang[3], Wenjuan Han†[4], Baobao Chang†[1]

Templates of Image Captioning (MSCOCO, Flick30k, Nocaps, Diffusiondb)
(1) Carefully analyze image 0: [IMG0] {image} to generate a concise and accurate description that accurately represents the objects, people, and scenery present. (2) Use clear and concise language that accurately describes the content of image 0: [IMG0] {image}. (3) Your caption should provide sufficient information about image 0: [IMG0] {image} so that someone who has not seen the image can understand it. (4) image 0 is [IMG0] {image}. Be specific and detailed in your description of image 0, but also try to capture the essence of image 0 in a succinct way. (5) image 0 is [IMG0] {image}. Based on the image 0, describe what is contained in this photo. Your caption should be no more than a few sentences and should be grammatically correct and free of spelling errors. (6) Include information in your caption that is specific to image 0: [IMG0] {image} and avoid using generic or ambiguous descriptions. (7) image 0 is [IMG0] {image}. Based on the image 0, give a caption about this image. Think about what message or story image 0 is conveying, and try to capture that in your image caption. (8) Based on the image 0, give a caption about this image. Your caption should provide enough detail about image 0: [IMG0] {image} to give the viewer a sense of what is happening in the image. (9) Give a caption about this image. Avoid using overly complex language or jargon in your caption of image 0: [IMG0] {image} that might confuse the viewer. (10) Be creative in your approach to captioning image 0: [IMG0] {image} and try to convey a unique perspective or story.

上下文学习

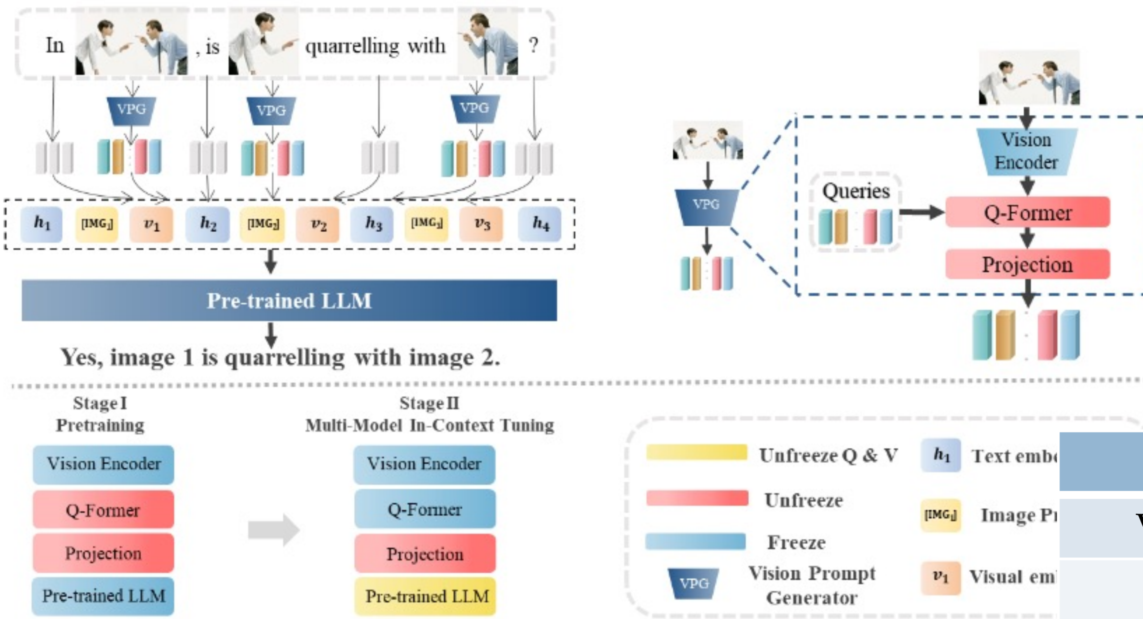
MMICL: EMPOWERING VISION-LANGUAGE MODEL WITH MULTI-MODAL IN-CONTEXT LEARNING

Haozhe Zhao*[1], Zefan Cai*[1], Shuzheng Si*[1], Xiaojian Ma[2], Kaikai An[1], Liang Chen[1], Zixuan Liu[3], Sheng Wang[3], Wenjuan Han†[4], Baobao Chang†[1]

Templates of Video Question Captioning (MSRVTT)
(1) image 0 is [IMG0] {image}. image 1 is [IMG1] {image}. image 2 is [IMG2] {image}. image 3 is [IMG3] {image}. image 4 is [IMG4] {image}. image 5 is [IMG5] {image}. image 6 is [IMG6] {image}. image 7 is [IMG7] {image}. Watch the images carefully and write a detailed description of what you see.
(2) image 0 is [IMG0] {image}. image 1 is [IMG1] {image}. image 2 is [IMG2] {image}. image 3 is [IMG3] {image}. image 4 is [IMG4] {image}. image 5 is [IMG5] {image}. image 6 is [IMG6] {image}. image 7 is [IMG7] {image}. After viewing the images, provide a summary of the main events or key points depicted.
(3) image 0 is [IMG0] {image}. image 1 is [IMG1] {image}. image 2 is [IMG2] {image}. image 3 is [IMG3] {image}. image 4 is [IMG4] {image}. image 5 is [IMG5] {image}. image 6 is [IMG6] {image}. image 7 is [IMG7] {image}. Pay close attention to the details in the images and provide accurate description to the images based on what you see.
(4) image 0 is [IMG0] {image}. image 1 is [IMG1] {image}. image 2 is [IMG2] {image}. image 3 is [IMG3] {image}. image 4 is [IMG4] {image}. image 5 is [IMG5] {image}. image 6 is [IMG6] {image}. image 7 is [IMG7] {image}. Utilize your comprehension skills to describe the context and events depicted in the images.
(5) image 0 is [IMG0] {image}. image 1 is [IMG1] {image}. image 2 is [IMG2] {image}. image 3 is [IMG3] {image}. image 4 is [IMG4] {image}. image 5 is [IMG5] {image}. image 6 is [IMG6] {image}. image 7 is [IMG7] {image}. Reflect on the images's narrative structure and identify any storytelling techniques or narrative devices used. Write a detailed description of what you see.
(6) image 0 is [IMG0] {image}. image 1 is [IMG1] {image}. image 2 is [IMG2] {image}. image 3 is [IMG3] {image}. image 4 is [IMG4] {image}. image 5 is [IMG5] {image}. image 6 is [IMG6] {image}. image 7 is [IMG7] {image}. Consider both the explicit and implicit information conveyed in the images to provide comprehensive description of the images.

Table 13: Instruction templates for task MSRVTT.

上下文学习



结构		参数
Vision Encoder		CLIP-L-224
Connector		Q-former+Linear Proj
Language Model		FlanT5
Training Recipe	Stage 1	微调Connector
	Stage 2	微调LP和部分LLM
Datasets	Stage 1	COCO Captioning LAION-400M
	Stage 2	MIC-5M 10%

上下文学习

Model	Cognition				Perception										Total Avg.
	Comm.	Num.	Text.	Code.	Existen.	Count	Pos.	Color	OCR	Poster	Cele.	Scene	Land.	Art.	
LLaVA	57.14	50.00	57.50	50.00	50.00	50.00	50.00	55.00	50.00	50.00	48.82	50.00	50.00	49.00	51.25
MiniGPT-4	59.29	45.00	0.00	40.00	68.33	55.00	43.33	75.00	57.50	41.84	54.41	71.75	54.00	60.50	51.85
MultiModal-GPT	49.29	62.50	60.00	55.00	61.67	55.00	58.33	68.33	82.50	57.82	73.82	68.00	69.75	59.50	62.97
VisualGLM-6B	39.29	45.00	50.00	47.50	85.00	50.00	48.33	55.00	42.50	65.99	53.24	146.25	83.75	75.25	63.36
VPGTrans	64.29	50.00	77.50	57.50	70.00	85.00	63.33	73.33	77.50	84.01	53.53	141.75	64.75	77.25	74.27
LaVIN	87.14	65.00	47.50	50.00	185.00	88.33	63.33	75.00	107.50	79.59	47.35	136.75	93.50	87.25	86.66
LLaMA-Adapter-V2	81.43	62.50	50.00	55.00	120.00	50.00	48.33	75.00	**125.00**	99.66	86.18	148.50	150.25	69.75	87.26
mPLUG-Owl	78.57	60.00	80.00	57.50	120.00	50.00	50.00	55.00	65.00	136.05	100.29	135.50	159.25	96.25	88.82
InstructBLIP	129.29	40.00	65.00	57.50	185.00	143.33	66.67	153.33	72.50	123.81	101.18	153.00	79.75	134.25	107.47
BLIP-2	110.00	40.00	65.00	75.00	160.00	135.00	73.33	148.33	110.00	141.84	105.59	145.25	138.00	136.50	113.13
Lynx	110.71	17.50	42.50	45.00	195.00	151.67	90.00	170.00	77.50	124.83	118.24	164.50	**162.00**	119.50	113.50
GIT2	99.29	50.00	67.50	45.00	190.00	118.33	**96.67**	158.33	65.00	112.59	145.88	158.50	140.50	**146.25**	113.85
Otter	106.43	72.50	57.50	70.00	**195.00**	88.33	86.67	113.33	72.50	138.78	**172.65**	**158.75**	137.25	129.00	114.19
Cheetor	98.57	77.50	57.50	**87.50**	180.00	96.67	80.00	116.67	100.00	147.28	164.12	156.00	145.73	113.50	115.79
LRV-Instruction	100.71	70.00	85.00	72.50	165.00	111.67	86.67	165.00	110.00	139.04	112.65	147.98	160.53	101.25	116.29
BLIVA	136.43	57.50	77.50	60.00	180.00	138.33	81.67	**180.00**	87.50	**155.10**	140.88	151.50	89.50	133.25	119.23
MMICL	**136.43**	**82.50**	**132.50**	77.50	170.00	**160.00**	81.67	156.67	100.00	146.26	141.76	153.75	136.13	135.50	**129.33**

Table 1: Evaluation results on the MME. Top two scores are highlighted and underlined, respectively.

上下文学习

Model	Flickr 30K	WebSRC	VQAv2	Hateful Memes	VizWiz
Flamingo-3B (Alayrac et al., 2022) (Zero-Shot)	60.60	-	49.20	53.70	28.90
Flamingo-3B (Alayrac et al., 2022) (4-Shot)	72.00	-	53.20	53.60	34.00
Flamingo-9B (Alayrac et al., 2022) (Zero-Shot)	61.50	-	51.80	57.00	28.80
Flamingo-9B (Alayrac et al., 2022) (4-Shot)	72.60	-	56.30	62.70	34.90
KOSMOS-1 (Huang et al., 2023b) (Zero-Shot)	67.10	3.80	51.00	63.90	29.20
KOSMOS-1 (Huang et al., 2023b) (4-Shot)	75.30	-	51.80	-	35.30
Zero-Shot Evaluation					
BLIP-2 (Li et al., 2023d) (FLANT5-XL)	64.51	12.25	58.79	60.00	25.52
BLIP-2 (Li et al., 2023d) (FLANT5-XXL)	60.74	10.10	60.91	62.25	22.50
InstructBLIP (Dai et al., 2023) (FLANT5-XL)	77.16	10.80	36.77	58.54	32.08
InstructBLIP (Dai et al., 2023) (FLANT5-XXL)	73.13	11.50	63.69	61.70	15.11
Zero-Shot Evaluation					
MMICL (FLAN-T5-XL)	60.56	12.55	62.17	60.28	25.04
MMICL (FLAN-T5-XXL)	78.64	18.85	69.99	60.32	29.34
MMICL (Instruct-FLAN-T5-XL)	**78.89**	14.75	69.13	61.12	29.92
MMICL (Instruct-FLAN-T5-XXL)	44.29	17.05	70.30	62.23	24.45
Few-Shot (4-Shot) Evaluation					
MMICL (FLAN-T5-XL)	71.95	12.30	62.63	60.80	50.17
MMICL (FLAN-T5-XXL)	75.37	18.70	69.83	61.12	33.16
MMICL (Instruct-FLAN-T5-XL)	74.27	14.80	69.16	61.12	33.16
MMICL (Instruct-FLAN-T5-XXL)	72.04	**19.65**	**70.56**	**64.60**	**50.28**

ICL能力时而有益时而有害，这是因为示例一方面能带来噪声，另一方面也容易产生幻觉。

- 研究背景
- 指令调优
- 上下文学习
- 思维链
- 总结反思

思维链: Multimodal CoT

Multimodal Chain-of-Thought Reasoning in Language Models

Zhuosheng Zhang [1] Aston Zhang [2] Mu Li [2] Hai Zhao [1] George Karypis [2] Alex Smola [2]

第一篇将CoT技术应用于多模态领域的论文。

根据CoT的学习范式，可以分为finetuning（本文）， few-shot和zero-shot三种类型。

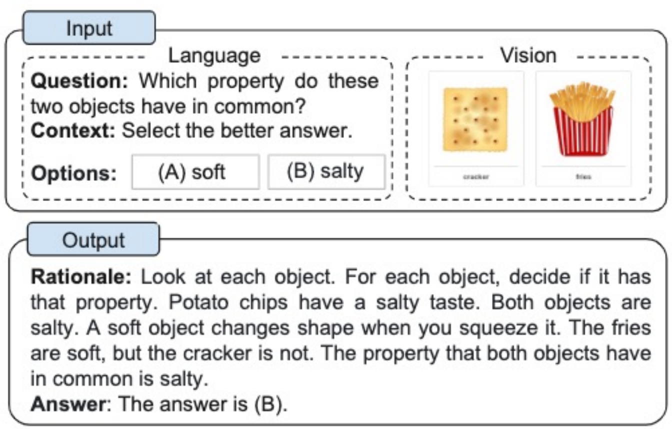


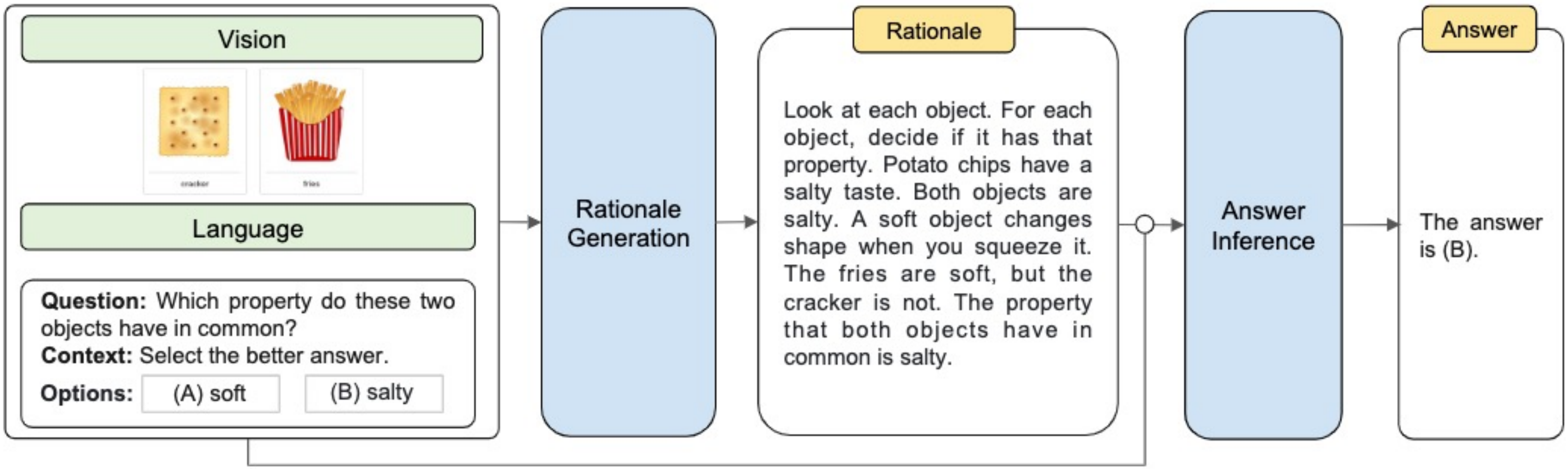
Figure 1. Example of the multimodal CoT task.

Table 2. Effects of CoT in the one-stage setting.

Method	Format	Accuracy
No-CoT	QCM→A	80.40
Reasoning	QCM→RA	67.86
Explanation	QCM→AR	69.77

本文利用UnifiedQA作为baseline，将问题，上下文，选项和图片的文本描述作为输入。当应用CoT技术时，在SQA上的性能有显著下降。

思维链: Multimodal CoT

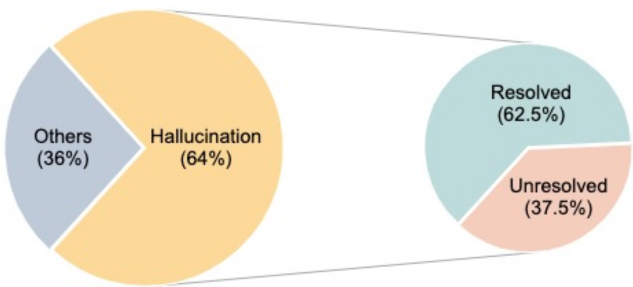


- 为了探究思维链对答案的影响，作者设计了上图的解耦结构，首先生成推理，再根据推理生成答案。
- “无描述的两阶段框架优于带描述的单阶段框架”&“增加描述后性能并没有显著提升”：描述可能存在负面作用。

Table 3. Two-stage setting of (i) rationale generation (RougeL) and (ii) answer inference (Accuracy).

Method	(i) QCM→ R	(ii) QCMR→ A
Two-Stage Framework	91.76	70.53
w/ Captions	91.85	71.12
w/ Vision Features	96.97	84.91

思维链: Multimodal CoT



(a) ratio of hallucination mistakes (b) correction rate w/ vision features

Figure 3. The ratio of hallucination mistakes (a) and correction rate w/ vision features (b).

Table 3. Two-stage setting of (i) rationale generation (RougeL) and (ii) answer inference (Accuracy).

Method	(i) QCM→ R	(ii) QCMR→ A
Two-Stage Framework	91.76	70.53
w/ Captions	91.85	71.12
w/ Vision Features	96.97	84.91

图像描述存在严重的信息丢失，使得不同模态的表征空间中缺乏充足的交互，从而导致模型产生幻觉。

Model	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Human	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
MCAN (Yu et al., 2019)	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down (Anderson et al., 2018)	70M	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN (Kim et al., 2018)	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF (Gao et al., 2019)	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT (Kim et al., 2021)	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM (Lu et al., 2021)	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT (Li et al., 2019)	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA$_{\text{Base}}$ (Khashabi et al., 2020)	223M	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
UnifiedQA$_{\text{Base}}$ w/ CoT (Lu et al., 2022a)	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
GPT-3.5 (Chen et al., 2020)	175B	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT (Lu et al., 2022a)	175B	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
Mutimodal-CoT$_{\text{Base}}$	223M	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
Mutimodal-CoT$_{\text{Large}}$	738M	**95.91**	**82.00**	**90.82**	**95.26**	**88.80**	**92.89**	**92.44**	**90.31**	**91.68**

- CoT技术对于推理能力有非常强大的增益；
- 视觉特征的提取以及与指令之间的交互是解决幻觉问题的关键。
- 两阶段的手工解耦方式提供了一种CoT链构造的思路。

- 作者介绍
- 研究背景
- 解决方法
- 实验效果
- 总结反思

- TinyMLLM是MLLM领域的一个热点问题，也更适合我们组的计算资源；
- 指令跟随能力、上下文学习能力以及思维链技术并不成熟，可以进一步研究；
- 幻觉和灾难性遗忘是困扰MLLM的两大问题，目前的研究并不是很充分。