xGen-MM (BLIP-3): A Family of Open Large Multimodal Models

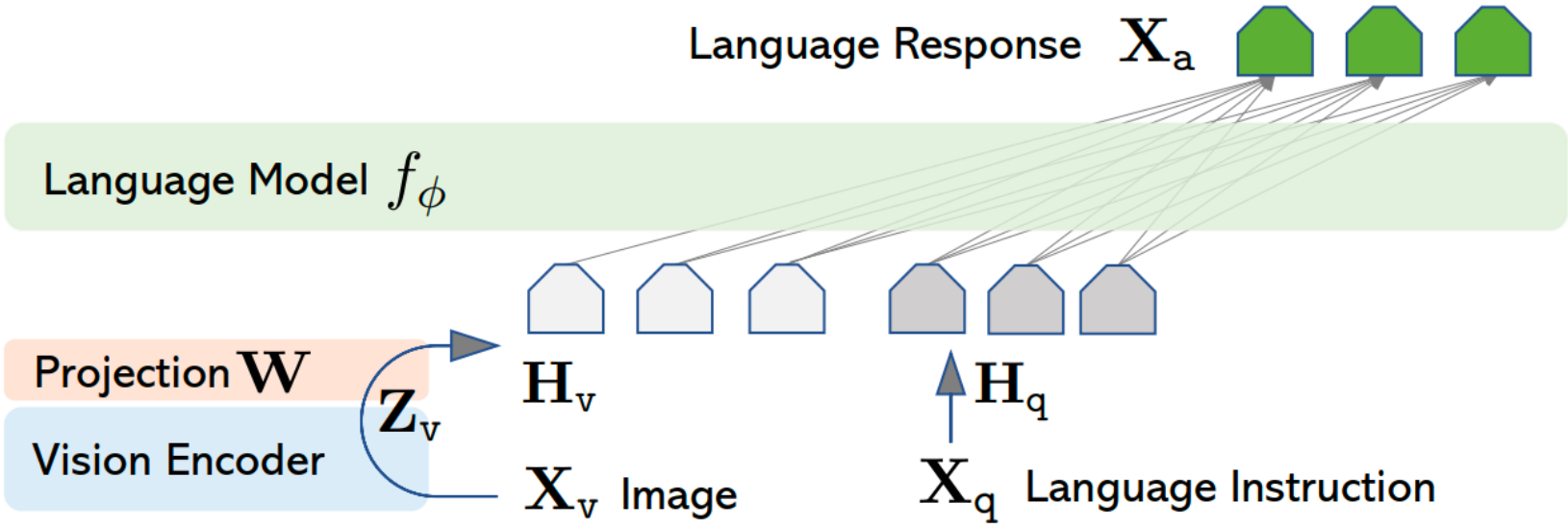
Arxiv 2024

目录

研究背景

多模态大语言模型

- 多模态输入，单模态输出（LLaVa、Qwen-VL、 InternLM、 MiniGPT-4）
- 构建一个LMM，什么是重要的？
  - MM1, Idefics2, Blip3



Improved Baselines with Visual Instruction Tuning. CVPR2024

研究背景

- MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

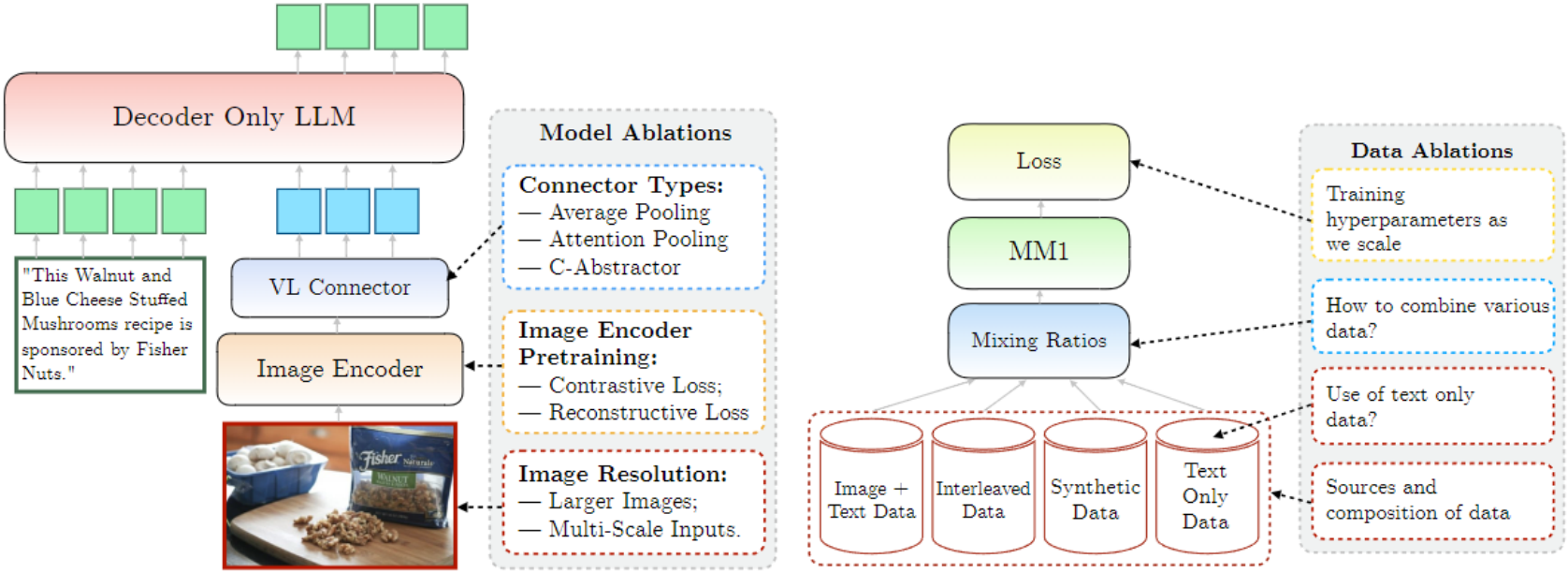


Fig. 3: *Left:* Model ablations: what visual encoder to use, how to feed rich visual data, and how to connect the visual representation to the LLM. *Right:* Data ablations: type of data, and their mixture.

研究背景

- MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

  - **Image Encoder**: A ViT-L/14 [27] model trained with a CLIP loss [91] on DFN-5B [31] and VeCap-300M [57]; images of size 336×336.
  - **Vision-Language Connector**: C-Abstractor [12] with 144 image tokens.
  - **Pre-training Data**: A mix of captioned images (45%), interleaved image-text documents (45%), and text-only (10%) data.
  - **Language Model**: A 1.2B transformer decoder-only language model.

  To evaluate the different design decisions, we use zero-shot and few-shot (4- and 8-shot) performance on a variety of captioning and VQA tasks: COCO Captioning [18], NoCaps [2], TextCaps [103], VQAv2 [38], TextVQA [104], VizWiz [39], GQA [46], and OK-VQA [82].

研究背景

- MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

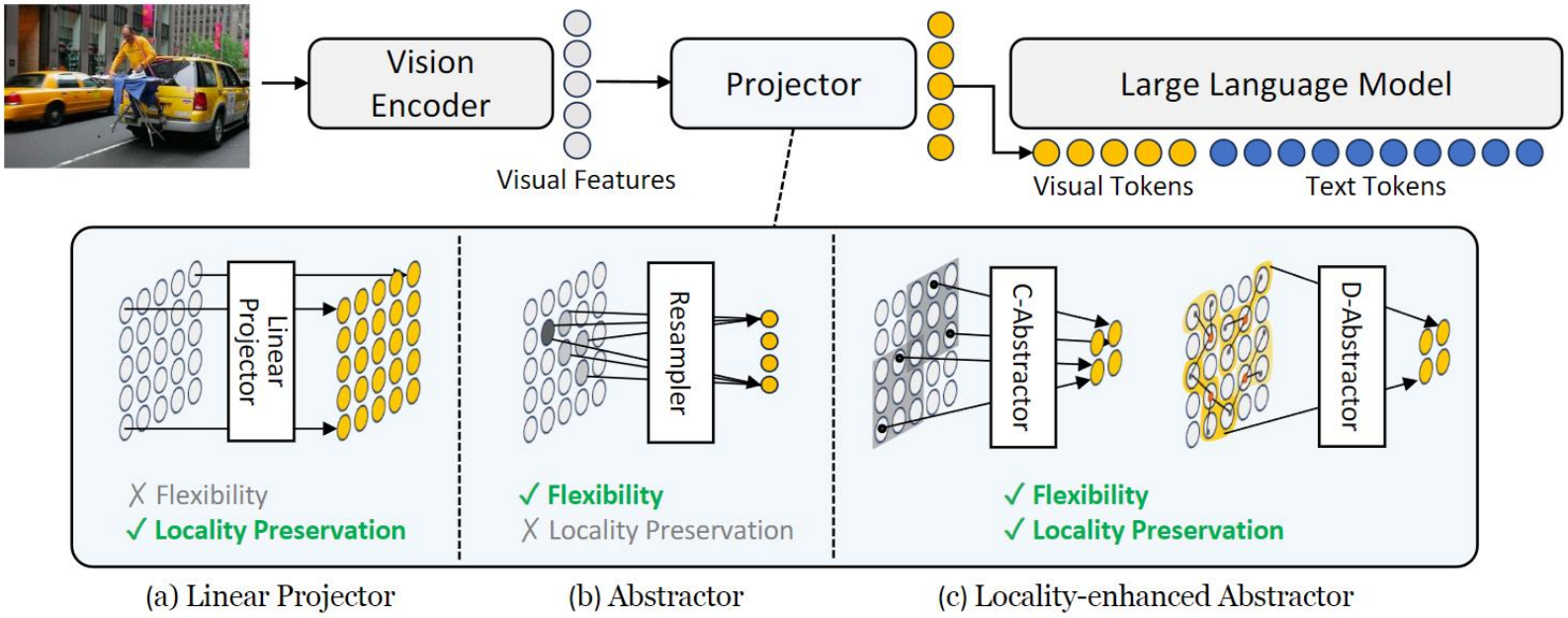


Figure 2. **Conceptual comparison between projectors** in terms of how to convert visual features into visual tokens. (a) Linear projector performs a one-to-one transformation, thus effective in preserving all local contexts of visual features, but limited in flexibility. (b) Abstractor such as resampler offers flexibility by abstracting the visual features into a smaller number of visual tokens but is limited in local context preservation by focusing on salient regions. (c) Our locality-enhanced abstractors can achieve both flexibility and locality preservation.

Honeybee: Locality-enhanced Projector for Multimodal LLM

研究背景

- MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

	Setup				Results		
	Model	Arch.	Image Res.	Data	0-shot	4-shot	8-shot
Recon.	AIM_{600M}	ViT/600M	224	DFN-2B	36.6	56.6	60.7
Recon.	AIM_{1B}	ViT/1B	224	DFN-2B	37.9	59.5	63.3
Recon.	AIM_{3B}	ViT/3B	224	DFN-2B	38.9	60.9	64.9
Contrastive	$CLIP_{DFN+VeCap}$	ViT-L	224	DFN-5B+VeCap	36.9	58.7	62.2
Contrastive	$CLIP_{DFN}$	ViT-H	224	DFN-5B	37.5	57.0	61.4
Contrastive	$CLIP_{DFN+VeCap}$	ViT-H	224	DFN-5B+VeCap	37.5	60.0	63.6
Contrastive	$CLIP_{DFN+VeCap}$	ViT-L	336	DFN-5B+VeCap	39.9	62.4	66.0
Contrastive	$CLIP_{DFN+VeCap}$	ViT-H	336	DFN-5B+VeCap	40.5	**62.6**	66.3
Contrastive	$CLIP_{OpenAI}$	ViT-L	336	ImageText-400M	39.3	62.2	66.1
Contrastive	$CLIP_{DFN}$	ViT-H	378	DFN-5B	**40.9**	62.5	**66.4**

Table 1: MM1 pre-training ablation across different image encoders (with 2.9B LLM). Note that the values in the Data column correspond to the data that was used for the initial training of the image encoder itself, not MM1. Recon.: Reconstructive loss. AIM: [30]; DFN-2/5B: [31]; VeCap: VeCap-300M [57]; OpenAI [91].

Encoder Lesson: Image resolution has the highest impact, followed by model size and training data composition.

研究背景

- MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

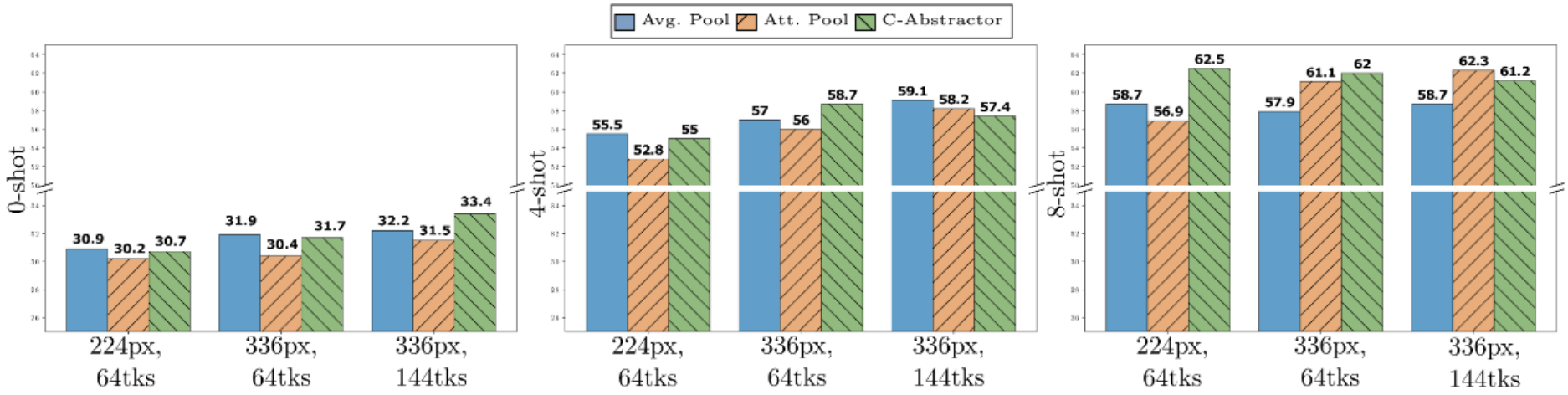


Fig. 4: 0-shot, 4-shot, and 8-shot ablations across different visual-language connectors for two image resolutions, and two image token sizes.

VL Connector Lesson: Number of visual tokens and image resolution matters most, while the type of VL connector has little effect

研究背景

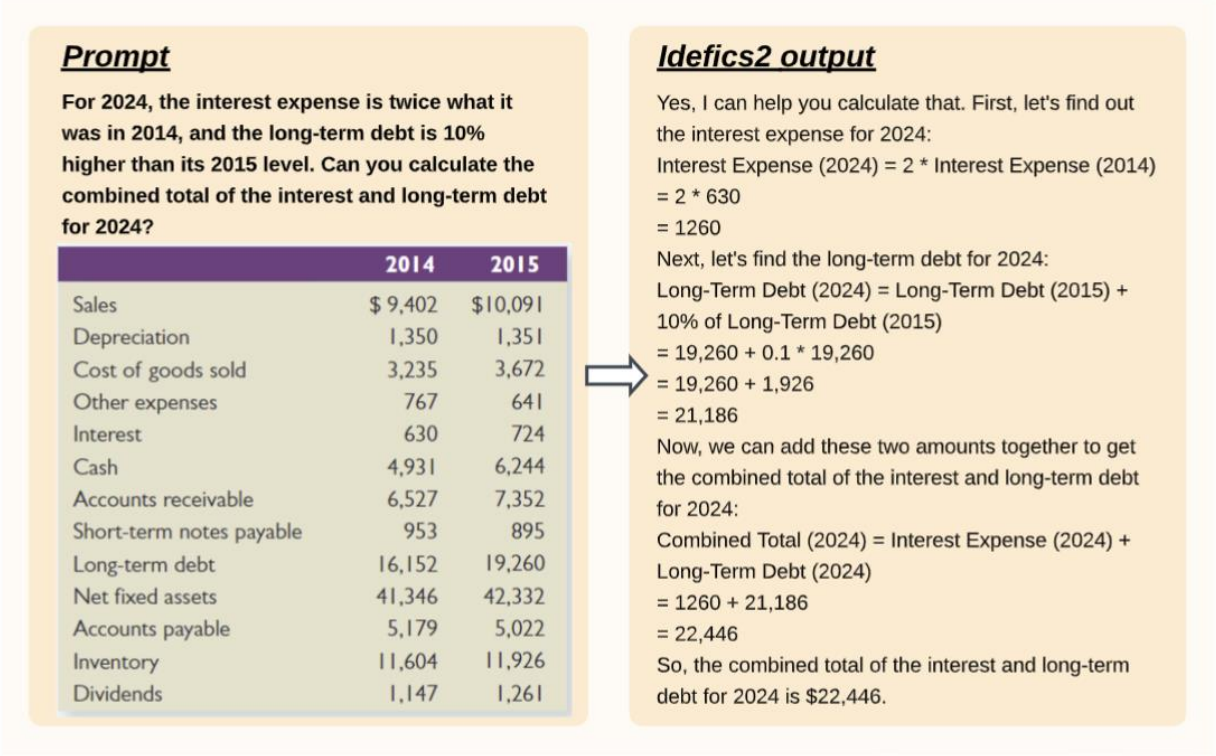
- MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

Data Lesson 1: Interleaved data is instrumental for few-shot and textonly performance, while captioning data lifts zero-shot performance.
Data Lesson 2: Text-only data helps with few-shot and text-only performance.
Data Lesson 3: Careful mixture of image and text data can yield optimal multimodal performance and retain strong text performance.
Data Lesson 4: Synthetic data helps with few-shot learning.

- **Image Encoder**: Motivated by the importance of image resolution, we use a ViT-H [27] model with 378×378 resolution, pre-trained with a CLIP objective on DFN-5B [31].
- **Vision-Language Connector**: As the number of visual tokens is of highest importance, we use a VL connector with 144 tokens. The actual architecture seems to matter less, we opt for C-Abstractor [12].
- **Data**: In order to maintain both zero- and few-shot performance, we use the following careful mix of 45% interleaved image-text documents, 45% image-text pair documents, and 10% text-only documents.

研究背景

- What matters when building vision-language models?
- Idefics2 8B VLM



Prompt

For 2024, the interest expense is twice what it was in 2014, and the long-term debt is 10% higher than its 2015 level. Can you calculate the combined total of the interest and long-term debt for 2024?

	2014	2015
Sales	$ 9,402	$10,091
Depreciation	1,350	1,351
Cost of goods sold	3,235	3,672
Other expenses	767	641
Interest	630	724
Cash	4,931	6,244
Accounts receivable	6,527	7,352
Short-term notes payable	953	895
Long-term debt	16,152	19,260
Net fixed assets	41,346	42,332
Accounts payable	5,179	5,022
Inventory	11,604	11,926
Dividends	1,147	1,261

Idefics2 output

Yes, I can help you calculate that. First, let's find out the interest expense for 2024:
Interest Expense (2024) = 2 * Interest Expense (2014)
= 2 * 630
= 1260
Next, let's find the long-term debt for 2024:
Long-Term Debt (2024) = Long-Term Debt (2015) + 10% of Long-Term Debt (2015)
= 19,260 + 0.1 * 19,260
= 19,260 + 1,926
= 21,186
Now, we can add these two amounts together to get the combined total of the interest and long-term debt for 2024:
Combined Total (2024) = Interest Expense (2024) + Long-Term Debt (2024)
= 1260 + 21,186
= 22,446
So, the combined total of the interest and long-term debt for 2024 is $22,446.

Figure 1: Idefics2-chatty analyzes the table to compute and answer the query.

研究背景

idefics2

- combine the visual inputs and the text input
  - cross-attention : LLM中间层进行交互
  - fully autoregressive ： 直接concat作为输入

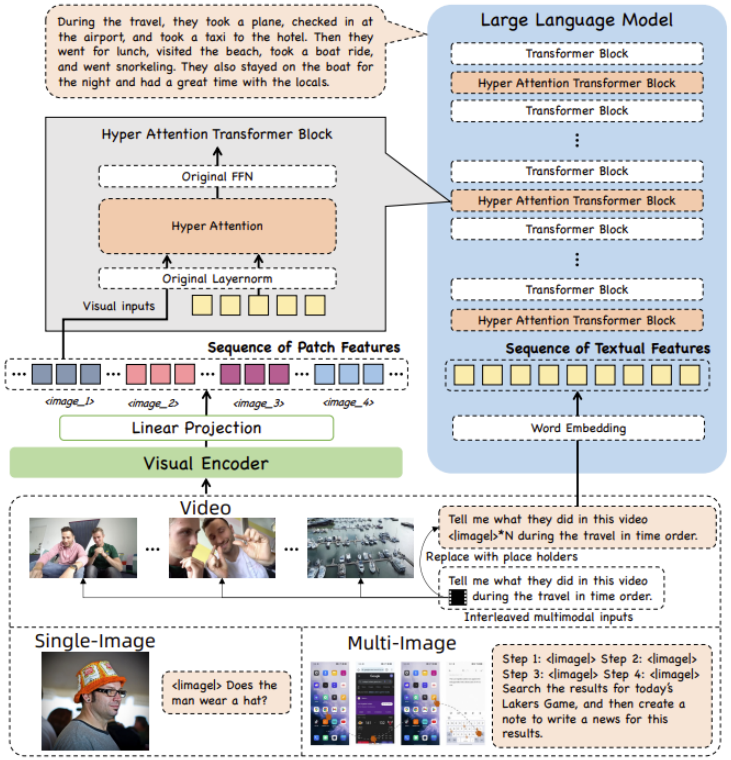


Figure 2: An overview of mPLUG-Owl3.

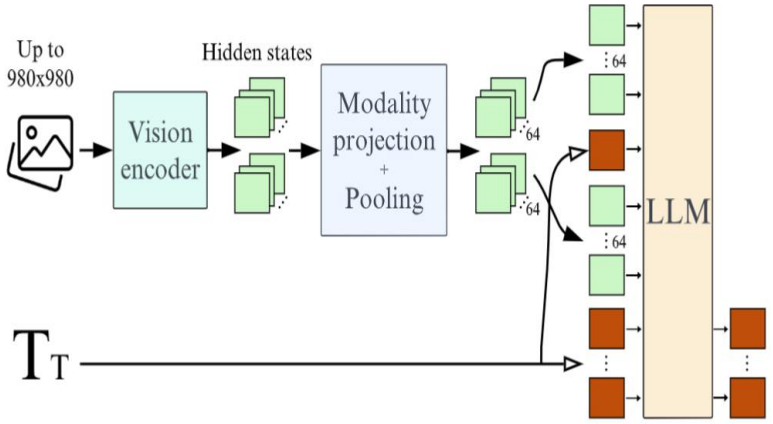


Figure 2: Idefics2 fully-autoregressive architecture: Input images are processed by the Vision encoder. The resulting visual features are mapped (and optionally pooled) to the *LLM* input space to get the visual tokens (64 in our standard configuration). They are concatenated (and potentially interleaved) with the input sequence of text embeddings (green and red column). The concatenated sequence is fed to the language model (*LLM*), which predicts the text tokens output.

- 评测指标：

In this section, we compare recurrent design choices in the vision-language model literature and highlight findings. Unless specified otherwise, we run the ablations for 6'000 steps and report the average score of the 4-shot performance on 4 downstream benchmarks measuring different capabilities: VQAv2 (Goyal et al., 2017) for general visual question answering, TextVQA (Singh et al., 2019) for OCR abilities, OKVQA (Marino et al., 2019) for external knowledge, and COCO (Lin et al., 2014) for captioning.

- 应该用什么backbone
  - LLM

LM backbone	Avg. score
Llama-1-7B	62.5
Mistral-7B	67.6

Table 1: Ablation on the language model backbone.

  - VISION

VE backbone	Res.	Avg. score
CLIP-ViT-H	224	57.4
EVA-CLIP-5B	224	60.2
SigLIP-SO400M	384	60.7

Table 2: Ablation on the vision encoder backbone.

  - 结论：固定参数，LLM影响大于vision encoder

- Fully autoregressive or cross attention?

Architecture	Backbones training	Avg. score
Fully autoreg. no Perceiver	Frozen	51.8
Fully autoreg.	Frozen	60.3
Cross-attention	Frozen	66.7
Cross-attention	LoRA	67.3
Fully autoreg.	LoRA	69.5

Table 3: Ablation for the architecture and method of training.

- 最终采用Fully autoreg.
- LLM和vision都frozen的情况下，cross_attn效果更好
- 需要训练的情况下，full autoreg更高，且参数少推理速度快

- Perceive sampler：降低visual token个数

Pooling	# vis. tok.	Avg. score
Perceiver	128	71.2
Perceiver	64	71.7

Table 4: Ablation on the pooling strategy.

Images	Res.	Avg. score
Square images	768	73.1
AR preserving	378-768	72.1

Table 5: Ablation on the aspect-ratio preserving strategy.

  - Token是可以被减少的，性能变化小
- 保持原有的resolution / ratio 对于结果影响不大
- 4 crops and origin images 对于ocr任务很有效
- Final 模型:open 8B parameters vision-language model: Idefics2.
  - SigLIP-SO400M and Mistral-7B-v0.1

Model	Size	Archi.	# tokens per image	VQAv2	TextVQA	OKVQA	COCO
OpenFlamingo	9B	CA	-	54.8	29.1	41.1	96.3
Idefics1	9B	CA	-	56.4	27.5	47.7	97.0
Flamingo	9B	CA	-	58.0	33.6	50.0	99.0
MM1	7B	FA	144	63.6	46.3	51.4	**116.3**
Idefics2-base	8B	FA	**64**	**70.3**	**57.9**	**54.6**	116.0

Table 8: Performance of Idefics2-base against state-of-the-art base VLMs. The evaluations were done with 8 random in-context examples, and in an open-ended setting for VQA tasks.
FA: fully autoregressive architecture. CA: cross-attention architecture.
(Task, Metric, Split): (VQAv2, VQA acc., testdev), (TextVQA, VQA acc., val), (OKVQA, VQA acc., val), (COCO, CIDEr, test)

Model	Size	# tokens per image	MMMU	MathVista	TextVQA	MMBench
LLaVA-NeXT	13B	2880	36.2/-	35.3	67.1	70.0
DeepSeek-VL	7B	576	36.6/-	36.1	64.4	73.2
MM1-Chat	7B	720	37.0/35.6	35.9	72.8	72.3
Idefics2	8B	**64**	**43.5/37.9**	**51.6**	70.4	**76.8**
Idefics2	8B	320	43.0/37.7	51.4	**73.0**	76.7

Table 9: Performance of Idefics2 against state-of-the-art VLMs up to a size of 14B parameters. The evaluations are done in zero shot. Idefics2 with 64 or 320 tokens per image is the same model (same weights), only the inference differs. The full table is present in Appendix A.3.2.
(Benchmark, Split, Metric): (MMMU, val/test, MMMU score), (MathVista, testmini, MMMU score), (TextVQA, val, VQA acc.), (MMBench, test, accuracy).

- 进一步说明 FA好于CA

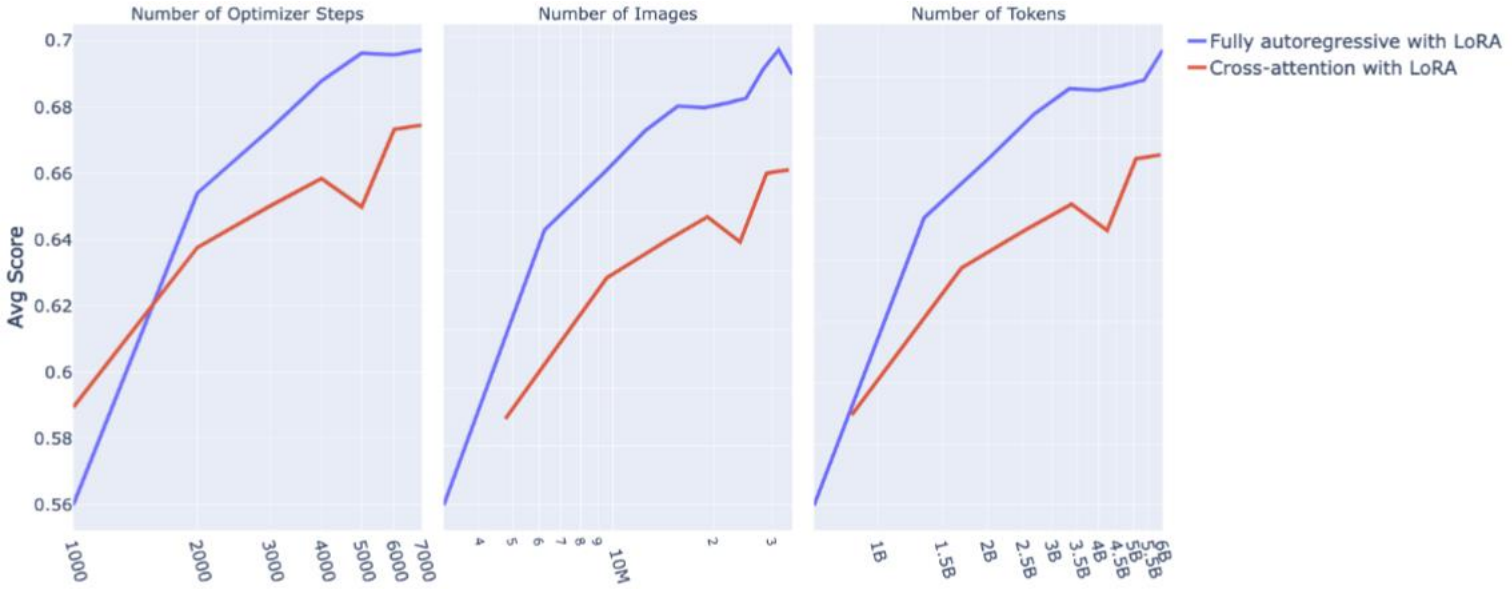


Figure 4: Comparison of the cross-attention and fully autoregressive architectures through the number of steps, the number of images and the number of text tokens.

研究方法

◘ 模型框架

- BLIP2数据不够，与当前LMM能力差距
- BLIP2 用Q-Former 架构来弥合视觉和语言模式，训练目标复杂（ITM、ITC 和 ITG 损失）增加训练难度
- BLIP-2 仅支持单图像输入，而交错的多模态数据格式是多模态数据的自然形式

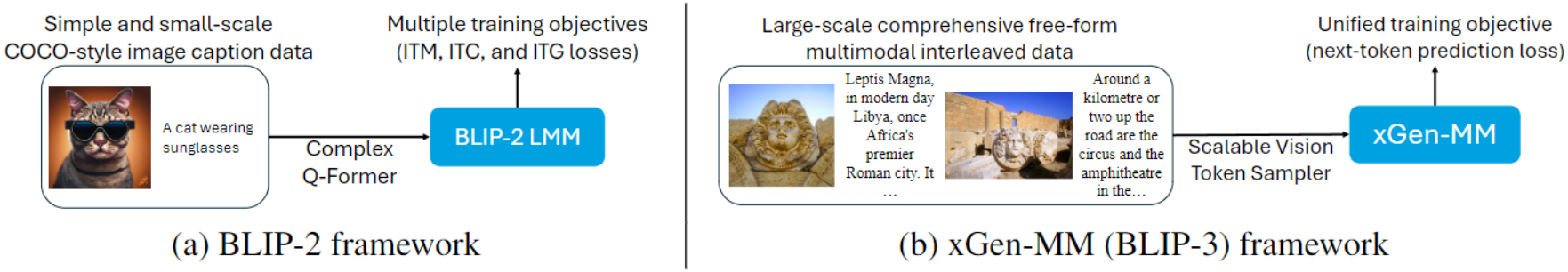


Figure 1: **We introduce xGen-MM (BLIP-3)**, a framework (b) for developing Large Multimodal Models (LMMs). Our framework improves upon BLIP-2 (a) [1] by (1) increasing the richness, scale, and diversity of training data, (2) replacing the Q-Former layers with a more scalable vision token sampler, and (3) simplifying the training process via the unification of the training objectives to a single loss at every training stage. The resulting suite of LMMs can perform various visual language tasks and achieve competitive performance across benchmarks.

研究方法

模型框架

- 用可扩展的视觉标记采样器替换Q-Former，并简化训练目标，只关注多模态上下文中文本标记的自回归损失。
- 引入了两个大规模、高质量的数据集：MINT-1T，一个万亿令牌规模的交错数据集； BLIP3-KALE，一个知识增强的高质量密集字幕数据集。两个额外的专业数据集：BLIP3-OCR-200M，这是一个具有密集 OCR 注释的大规模数据集； BLIP3-GROUNDING-50M，这是一个大规模的视觉基础数据集。
- 带有DPO的安全调优模型，旨在减轻幻觉、提高安全性等有害行为。

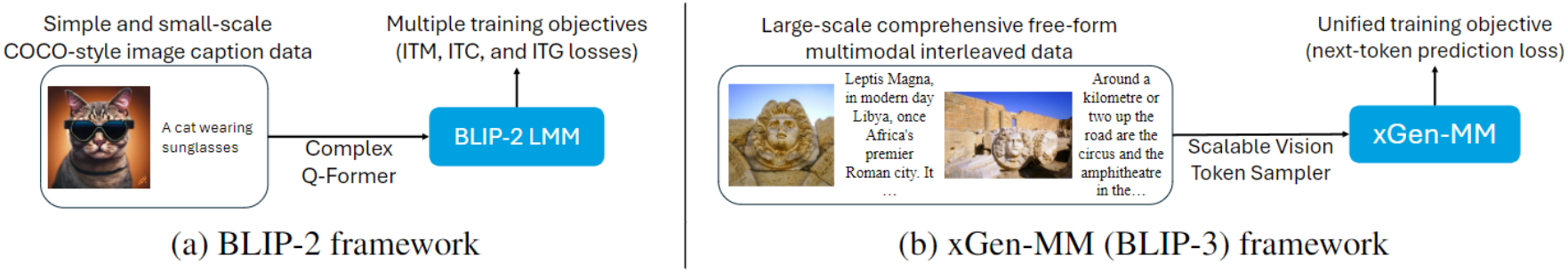


Figure 1: **We introduce xGen-MM (BLIP-3)**, a framework (b) for developing Large Multimodal Models (LMMs). Our framework improves upon BLIP-2 (a) [1] by (1) increasing the richness, scale, and diversity of training data, (2) replacing the Q-Former layers with a more scalable vision token sampler, and (3) simplifying the training process via the unification of the training objectives to a single loss at every training stage. The resulting suite of LMMs can perform various visual language tasks and achieve competitive performance across benchmarks.

Flamingo: a visual language model for few-shot learning

研究方法

模型架构

- VIT / vision token sampler (perceiver resampler) / phi3-mini
- Patchwise encoding / perceiver resampler 降低token个数

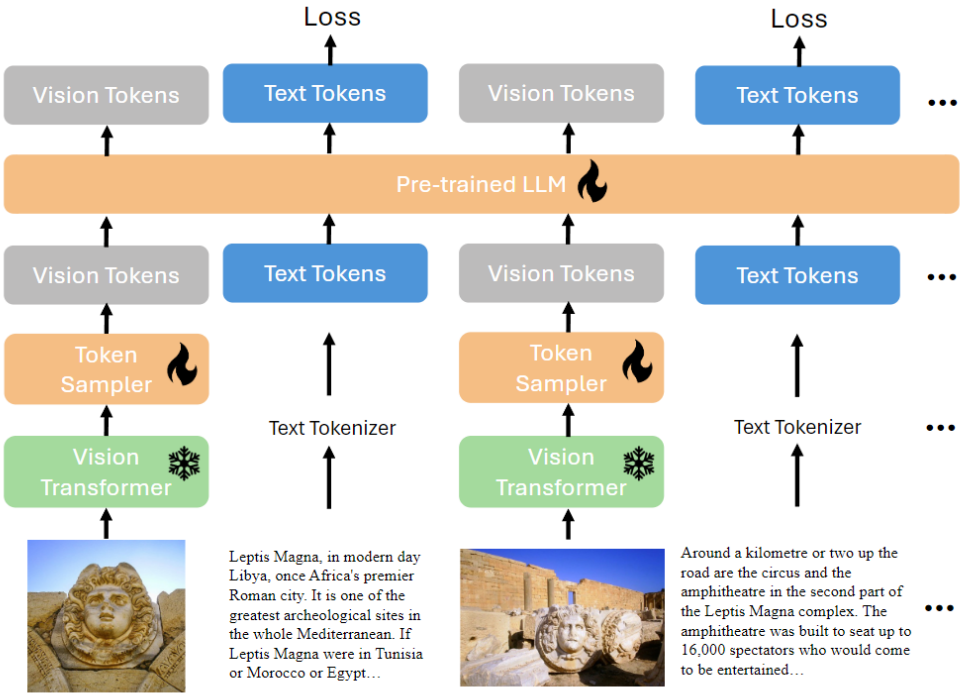
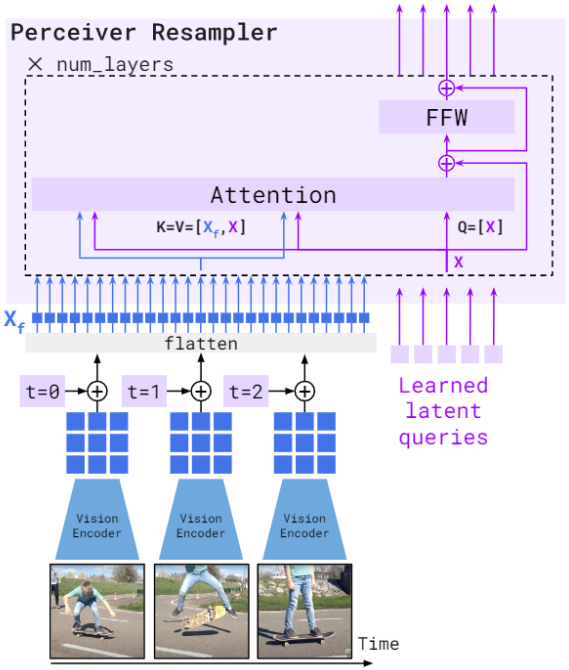


Figure 2: **Overview of the xGen-MM (BLIP-3) framework.** Free-form interleaved images and texts from the ensembled interleaved and caption datasets are input into the framework, with each modality undergoing a separate tokenization process to be fed into the pre-trained LLM in natural order. A standard auto-regressive loss is then applied to the text tokens. The Vision Transformer is kept frozen during training, while all other parameters, including the token sampler and the pre-trained LLM, are trained.

研究方法

损失函数




```
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

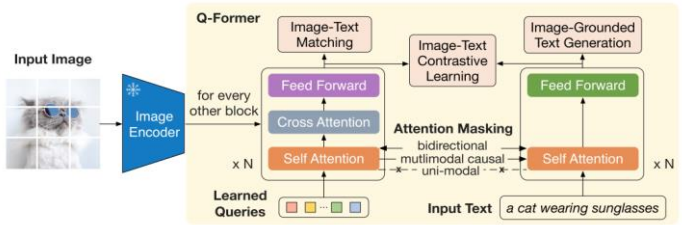


Figure 5: **The Perceiver Resampler** module maps a *variable* size grid of spatio-temporal visual features output by the Vision Encoder to a *fixed* number of output tokens (five in the figure), independently from the input image resolution or the number of input video frames. This transformer has a set of learned latent vectors as queries, and the keys and values are a concatenation of the spatio-temporal visual features with the learned latent vectors.

研究方法

- 训练数据

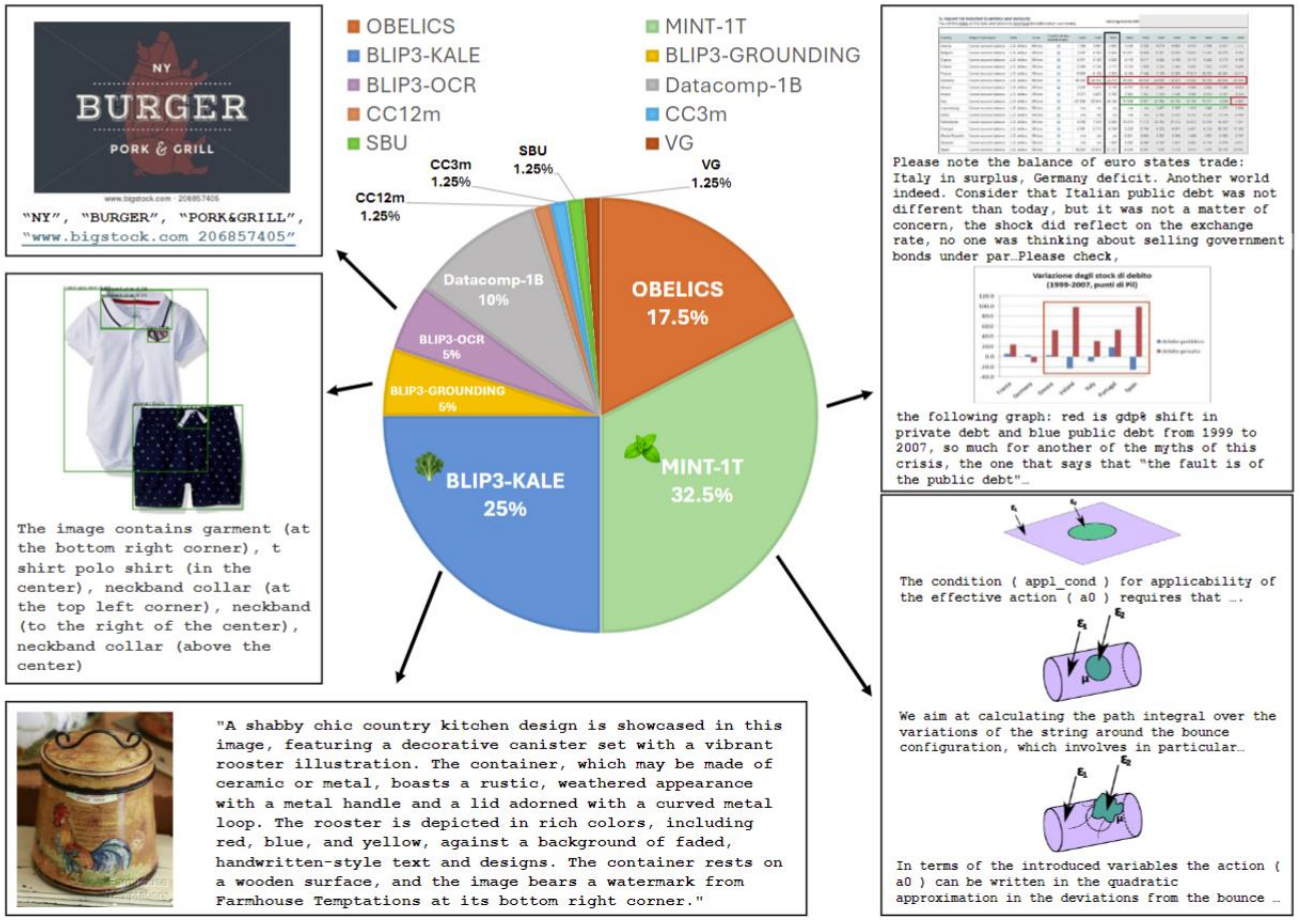


Figure 3: **Overview of xGen-MM (BLIP-3) Pre-training Datasets.**

研究方法

模型训练

- Pre-training.
  - 数据：MINT-1T / OBELICS / 公开数据集合

Model	Shot	Visual Question Answering			Captioning		
		VQAv2	TextVQA	OKVQA	COCO	NoCaps	TextCaps
< 5B Model Comparisons							
Flamingo-3B [22]	0	49.2	30.1	41.2	73.0	–	–
	4	53.2	32.7	43.3	85.0	–	–
	8	55.4	32.4	44.6	90.6	–	–
MM1-3B [9]	0	46.2	29.4	26.1	73.5	55.6	63.3
	4	57.9	45.3	44.6	**112.3**	99.7	84.1
	8	63.6	44.6	48.4	**114.6**	**104.7**	88.8
xGen-MM-base (4B)	0	43.1	34.0	28.0	67.2	82.6	69.5
	4	**66.3**	**54.2**	**48.9**	107.6	**100.8**	**89.9**
	8	**66.9**	**55.3**	**50.1**	109.8	104.6	**94.0**
Larger Models for Reference							
Flamingo-9B [22]	8	58.0	33.6	50.0	99.0	-	-
Idefics-9B [13]	8	56.4	27.5	47.7	97.0	86.8	63.2
MM1-7B [9]	8	63.6	46.3	51.4	116.3	106.6	88.2
Idefics2-8B [11]	8	70.3	57.9	54.6	116.0	-	-

Table 1: **Few-shot Pretraining Evaluation.** Following [9], we randomly sample demonstrations from the training set as few-shot examples. We report CIDEr score for captioning and accuracy for VQA.

研究方法

模型训练

- Pre-training.
  - 数据：MINT-1T / OBELICS / 公开数据集合
- SFT
  - 一百万公开instruction tuning数据集

Model (Size)	SEED -IMG	SEED v2	MMB (dev)	MM Star	MME (norm)	CVB -2D	CVB -3D	RealW QA	MMMU (val)	Math Vista	Sci QA	POPE	Text VQA	Avg. (all)	Avg. (perc.)
Closed-source models															
GPT-4V	72.0	-	80.8	49.7	63.3	64.3	73.8	56.5	53.8	48.2	82.1	75.4	-	-	-
MM1-3B-Chat (3B)	68.8	-	67.8	-	62.9	-	-	-	33.9	-	-	87.4	-	-	-
Open-source models															
HPT-1.5-edge (4B)	**72.3**	-	74.6	45.8	-	-	-	-	42.6	**45.1**	85.4	**91.0**	-	-	-
VILA-1.5-3B (3B)	67.9	-	63.4	-	-	-	-	-	33.3	-	69.0	85.9	-	-	-
VILA-1.5-3B* (3B)	67.9	51.9	62.4	40.3	58.5	50.1	60.3	53.3	34.1	30.6	68.9	86.9	58.1	55.6	59.1
phi-3-vision (4B)	-	-	80.5	-	-	-	-	-	-	44.5	90.8	85.8	70.9	-	-
phi-3-vision* (4B)	71.0	52.7	74.2	47.9	55.3	60.7	68.2	59.1	**46.1**	**45.1**	**90.2**	83.5	**73.3**	63.6	63.6
xGen-MM-inst. (4B)	71.8	53.9	76	46.7	63.8	66.2	**75.4**	**61.6**	42.8	39.2	85.6	87.0	72.0	64.8	66.9
xGen-MM-inst. -interleave (4B)	72.2	**55.5**	**76.8**	**48.1**	**64.4**	**69.3**	72.3	60.5	41.1	39.6	88.3	87.0	71.0	**65.1**	**67.3**

Table 2: **Evaluation on single-image benchmarks.** phi-3-vision* and VILA-1.5-3B* are tested with ou evaluation code[2] for a fair comparison. We also include the GPT-4V (`gpt-4-1106-preview`) performanc (provided by the evaluation codebase) as a reference in the first row.

研究方法

模型训练

- Pre-training.
  - 数据：MINT-1T / OBELICS / 公开数据集合
- SFT
  - 一百万公开instruction tuning数据集

Model	BLINK	QBench-2	Mantis-eval
GPT-4V	51.1	73.4	62.7
VILA-1.5-3B* (3B)	39.8	51.7	41.9
xGen-MM-inst. (4B)	46.6	52.4	42.4
xGen-MM-inst.-interleave (4B)	**49.7**	**75.1**	**56.7**

Table 3: **Evaluation on multi-image benchmarks.** VILA-1.5-3B* results are obtained using the same evaluation code as our models. We include the GPT-4V performance as a reference in the first row.

研究方法

模型训练

- Pre-training.
  - 数据：MINT-1T / OBELICS / 公开数据集合
- SFT
  - 一百万公开instruction tuning数据集
  - 图文交互sft
- Alignment DPO ： 六万 preference sample
- safety fintuning : VLGuard dataset, 2k examples of unsafe images and instructions.

Method	Safety	Hallucination		Helpfulness (Control)			
	VLGuard (↓)	HalBench (↑)	POPE (↑)	SEED-IMG (↑)	MMB-dev (↑)	MME (↑)	MMStar (↑)
xGen-MM-inst. (4B)	56.6	56.3	87.0	71.8	76.0	63.8	46.7
+ DPO	54.9	57.1	87.0	71.9	76.4	63.0	47.1
+ Safety FT	5.2	56.6	86.8	72.1	76.4	64.4	47.1

Table 4: **Post-training results.** We report results on safety and hallucination benchmarks after post-training, as well as on four helpfulness benchmarks as a control. Post-training improves harmlessness without compromising helpfulness.

消融实验

Visual Backbone	Text-VQA	OK-VQA	COCO-Caps	Text-Caps
DFN	41.1 / 41.8	**48.4 / 49.5**	107.2 / 109.4	78.2 / 79.9
SigLIP	**49.1 / 50.5**	48.4 / 48.9	**108.7 / 110.2**	**84.7 / 88.6**

Table 6: Few-shot (4-shot / 8-shot) performance given different visual backbones.

Visual Token	Text-VQA	OK-VQA	COCO-Caps	Text-Caps
128	41.1 / 41.8	**48.4 / 49.5**	107.2 / **109.4**	78.2 / 79.9
64	**41.2 / 42.6**	47.6 / 48.3	**108.0** / 109.3	**79.5 / 81.6**

Table 7: Few-shot (4-shot / 8-shot) performance given the different number of visual tokens.

Text-only SFT data	MMMU (val)	MathVista (mini)	Science QA	MME (norm)	MMStar
Conversation	39.1	37.1	**84.8**	**64.9**	**46.1**
Conversation + Math + Coding	**40.9**	**38.9**	81.4	64.8	45.3

Table 8: **The impact of text-only SFT data.** We compare two choices of text-only SFT data used for the image-text SFT data mixture.

消融实验

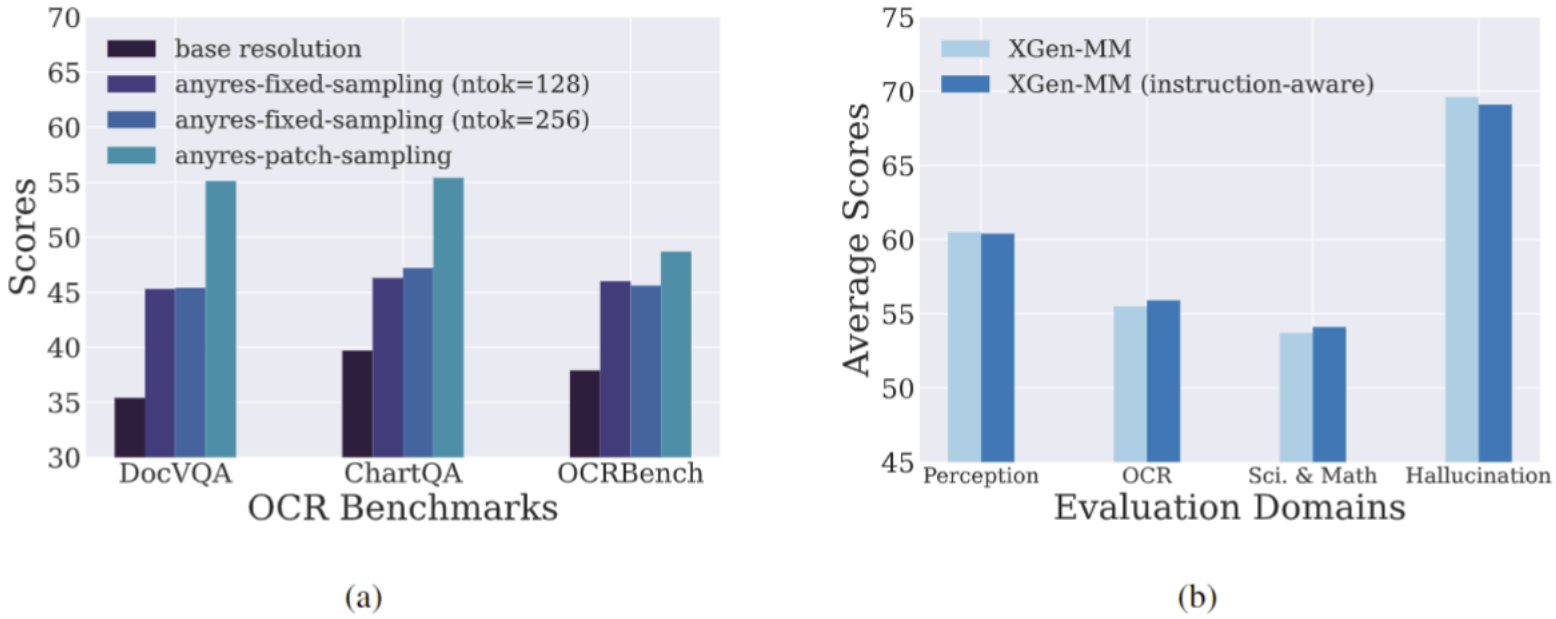


Figure 8: **SFT ablation studies.** (a). Comparison of different vision token sampling strategies on OCR benchmarks. (b). Comparison between our model and its "instruction-aware" alternative. For each evaluation domain in Figure (b), we report the average score on multiple relevant benchmarks.

Instruct aware 可能只在q-former中有用

目录

总结

- 融合形式上，采用fully autoregressive
- Connector方面，不再使用Q-former结构，多采用perceive sampler的方法
- 简单、统一、可扩展的形式
- 数据进一步探索，主要是提出了三个数据集