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

Brandon McKinzie°, Zhe Gan°, Jean-Philippe Fauconnier⋆,
Sam Dodge⋆, Bowen Zhang⋆, Philipp Dufter⋆, Dhruti Shah⋆, Xianzhi Du⋆,
Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang,
Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang,
Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman,
Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch⋆,
Alexander Toshev†, and Yinfei Yang†

Apple
{bmckinzie,zhe.gan,toshev,yinfeiy}@apple.com
°First authors; ⋆Core authors; †Senior authors

Abstract. In this work, we discuss building performant Multimodal Large Language Models (MLLMs). In particular, we study the importance of various architecture components and data choices. Through careful and comprehensive ablations of the image encoder, the vision language connector, and various pre-training data choices, we identified several crucial design lessons. For example, we demonstrate that for large-scale multimodal pre-training using a careful mix of image-caption, interleaved image-text, and text-only data is crucial for achieving state-of-the-art (SOTA) few-shot results across multiple benchmarks, compared to other published multimodal pre-training results. Further, we show that the image encoder together with image resolution and the image token count has substantial impact, while the vision-language connector design is of comparatively negligible importance. By scaling up the presented recipe, we build **MM1**, a family of multimodal models, including both dense variants up to 30B and mixture-of-experts (MoE) variants up to 64B, that are SOTA in pre-training metrics and achieve competitive performance after supervised fine-tuning on a range of established multimodal benchmarks. Thanks to large-scale pre-training, MM1 enjoys appealing properties such as enhanced in-context learning, and multi-image reasoning, enabling few-shot chain-of-thought prompting.

1 Introduction

In recent years, the research community has achieved impressive progress in language modeling and image understanding. Thanks to the availability of large-scale image-text data and compute at scale, we have seen the emergence of highly performant Large Language Models (LLMs) [6, 7, 16, 18, 22, 85, 86, 93, 98, 100, 106, 116] and Vision Foundation Models [35, 82, 84] that have become the *de-facto* standard for the majority of language and image understanding problems.

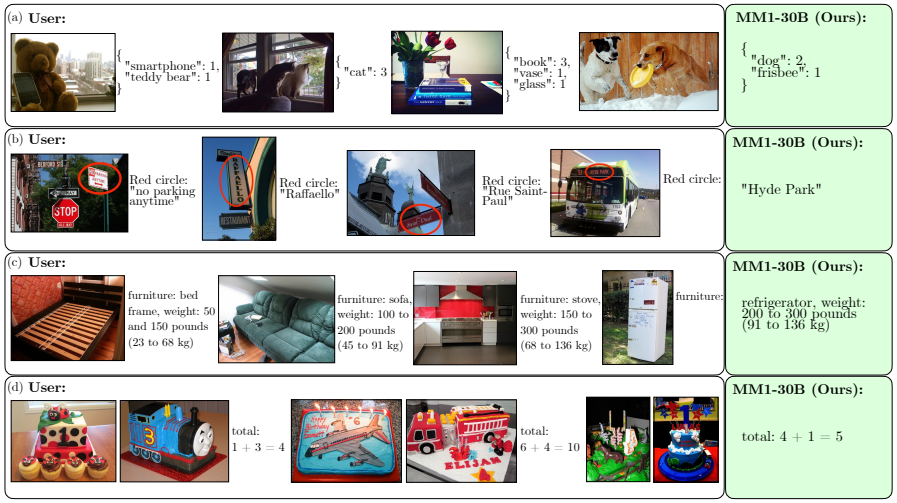


Fig. 1: MM1 can perform in-context predictions thanks to its large-scale multimodal pre-training. This allows MM1 to (a) count objects and follow custom formatting, (b) refer to parts of the images and perform OCR, (c) demonstrate common-sense and word knowledge about everyday objects, and (d) perform basic math functions. Images are from the COCO 2014 validation set [66].

Given the above developments, an area of multimodal foundation models has emerged that marries the above advances into a single model achieving superior capabilities. In particular, Multimodal Large Language Models (MLLMs) are large-scale foundation models that consume image and text data and produce text [24, 61, 73, 101]. After the rise of LLMs, MLLMs are emerging as the next frontier in foundation models.

When it comes to transparency, existing MLLMs fall into two categories: closed models [1, 97] and open models [3–5, 71, 83]. In the former category, the models might be available for use, but little to nothing is known about the data, model architecture, and training details. In the latter category, the model parameters might be released together with a detailed description of data, model, and training configurations, thus allowing the community to build upon. However, most of the works, both open and closed, release close to nothing about the process they have undergone to arrive at their algorithmic design choices, especially regarding multimodal pre-training.

To further research in this area, we believe it is imperative to distill principles and lessons of how to build such models that might outlive concrete component implementations. Thus, in this paper, we document the MLLM building process and attempt to formulate design lessons, that we hope are of use to the community.

In particular, our contributions are as follows. First, we perform ablations at small scale across (1) model architecture decisions and (2) pre-training data

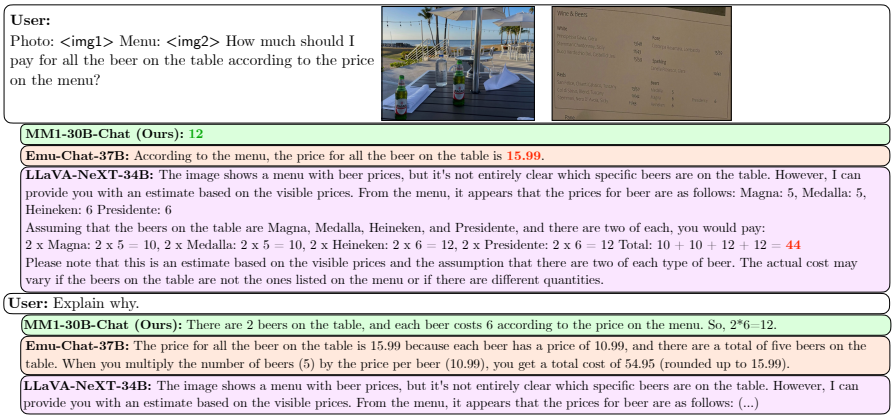


Fig. 2: MM1 can follow instructions and reason across images. Example and images from VILA [65]; VILA answers correctly when prompted with chain-of-thought.

choices. We identify several interesting trends. On the modeling side, we see that design aspects are in the following order of importance: image resolution, visual encoder loss and capacity, and visual encoder pre-training data. Surprisingly, though, we find little evidence that architectural decisions of how visual data is fed into the LLM matter.

Further, we use three different types of multimodal pre-training data: image-caption, interleaved image-text, and text-only data. We see that when it comes to few-shot and text-only performance, interleaved and text-only training data is of paramount importance, while for zero-shot performance, caption data matters most. We demonstrate that these trends hold after Supervised Fine-Tuning (SFT), both on the evaluations used in the pre-training as well as on further benchmarks. This shows that capabilities and modeling decisions discovered during pre-training are retained after fine-tuning.

Finally, we scale up our model by using larger LLMs, from 3B, 7B, to 30B, and by exploring mixture-of-experts (MoE) models, from 3B with 64 experts to 7B with 32 experts. This leads to a family of performant models, that outperforms most of the relevant works to the best of our knowledge. In particular, the pre-trained model MM1 is SOTA, performing better than Emu2 [96], Flamingo [3], and IDEFICS [42] on captioning and visual question answering (VQA) tasks in few-shot settings. The final models, after SFT, achieve competitive performance across 12 established multimodal benchmarks.

Thanks to large-scale multimodal pre-training, as shown in Figures 1 and 2, MM1 enjoys appealing properties such as in-context predictions, multi-image and chain-of-thought reasoning. MM1 also enables strong few-shot learning capability after instruction tuning. These strong results demonstrate that the presented recipe for building MLLMs translates the design principles to a competitive model at scale. We hope that these presented insights will remain relevant, even as specific modeling components and data sources evolve.

2 Related Work

The type of MLLMs concerned in this work build upon a strong pre-trained autoregressive LLM that consumes both text and visual tokens, the latter obtained via an image encoder [5, 14, 24, 40, 58, 70, 83]. Our approach is based on a decoder-only architecture, akin to Kosmos-1 [40].

Recent research has increasingly focused on visual instruction tuning on top of the pre-trained LLM [57]. Prominent examples include LLaVA(-1.5/NeXT) [68–70], MiniGPT-4 [119], mPLUG-Owl(-2/Doc) [109–111], Otter [54, 55], InstructBLIP [20], Honeybee [9], SPHINX(-X) [31,67], to name a few. There is also a rich body of literature on constructing instruction-tuning data [12, 32, 60, 103, 117], enabling MLLMs for referring and grounding [11, 50, 83, 105, 112, 115], image generation and editing [30, 48, 96].

The body of work that focuses on thorough ablations, in particular also on the pre-training side, is relatively sparse. VILA [65] focuses on studying various components of multimodal pre-training, but falls short of providing optimization details or detailed pre-training evaluations. Emu2 [96], on the other side, provides details regarding pre-training optimization parameters and base model results. However, they do not provide ablations that justify the various component decisions. IDEFICS [52] is another work that provides details regarding large-scale multimodal pre-training. However, their focus is primarily on closely replicating the closed-source Flamingo [3] model.

In contrast to these previous works, we aim to provide details regarding all components of our pre-training strategy, from hyperparameters to data to architecture. We also provide results for our base pre-trained models to help differentiate the impact of multimodal pre-training *vs.* instruction tuning. Furthermore, we provide extensive ablations on the precise impacts of decisions regarding visual encoders, vision-language connectors, and pre-training data mixture.

3 Recipe for Building MM1

Building performant MLLMs is a highly empirical endeavor. In this work, we present details of the ablations we have performed to arrive at a performant model. We explore three major axes of design decisions:

- **Architecture**: We investigate different pre-trained image encoders and explore varying ways of connecting LLMs with these encoders.
- **Data**: We consider different types of data and their relative mixture weights.
- **Training Procedure**: We explore how to train the MLLM including the hyperparameters and what parts of the model to train at what stage.

3.1 Empirical Setup for Ablations

In order to identify what are good choices along each of the above axes, we need an efficient way to assess model performance. As training a large MLLM can take substantial resources, we utilize a simplified setup for ablations.

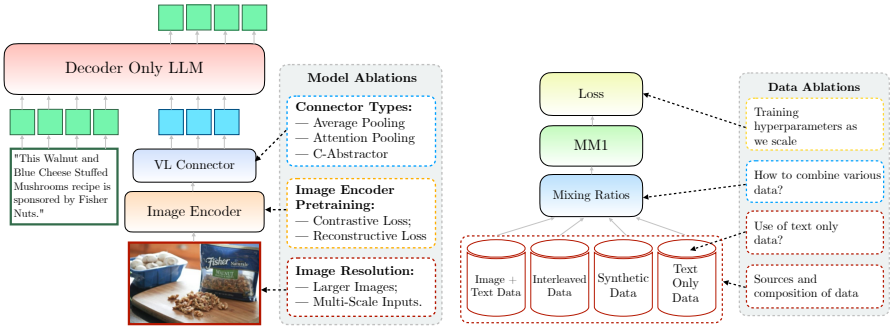


Fig. 3: *Left:* Model ablations: what visual encoder to use, how to feed rich visual data, and how to connect the visual representation to the LLM. *Right:* Data ablations: type of data, and their mixture.

More concretely, we use a smaller base configuration of our model that we ablate from. We modify one component at a time, either an architectural module or a data source, and assess the impact of the design choice for each of these components. This allows us to arrive to the final model-data configuration that we scale up, both in terms of model parameters as well as training time. The base configuration for ablations is as follows:

- **Image Encoder**: A ViT-L/14 [23] model trained with a CLIP loss [84] on DFN-5B [27] and VeCap-300M [51]; images of size 336×336.
- **Vision-Language Connector**: C-Abstractor [9] 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 [15], NoCaps [2], TextCaps [94], VQAv2 [33], TextVQA [95], VizWiz [34], GQA [41], and OK-VQA [76].

3.2 Model Architecture Ablations

In this work, we analyze components that enable an LLM to process visual data. Specifically, we investigate (1) how to best pre-train a visual encoder, and (2) how to bridge the visual features to the space of the LLM (see Figure 3, left).

Image Encoder Pre-training. Most MLLMs use a CLIP pre-trained image encoder [20, 68, 70, 110], while recent works also started to explore vision-only self-supervised models, such as DINOv2 [67, 99], as the image encoder. Here, we primarily ablate the importance of image resolution and image encoder pre-training objective. Note that unlike the rest of our ablations, here we use a 2.9B LLM (instead of 1.2B) to ensure there is sufficient capacity to utilize some of the larger image encoders.

	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
	AIM_{1B}	ViT/1B			37.9	59.5	63.3
	AIM_{3B}	ViT/3B			38.9	60.9	64.9
Contrastive	$CLIP_{DFN+VeCap}$	ViT-L	224	DFN-5B+VeCap	36.9	58.7	62.2
	$CLIP_{DFN}$	ViT-H		DFN-5B	37.5	57.0	61.4
	$CLIP_{DFN+VeCap}$	ViT-H		DFN-5B+VeCap	37.5	60.0	63.6
	$CLIP_{DFN+VeCap}$	ViT-L	336	DFN-5B+VeCap	39.9	62.4	66.0
	$CLIP_{DFN+VeCap}$	ViT-H			40.5	**62.6**	66.3
	$CLIP_{OpenAI}$	ViT-L		ImageText-400M	39.3	62.2	66.1
	$CLIP_{DFN}$	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: [26]; DFN-2/5B: [27]; VeCap: VeCap-300M [51]; OpenAI [84].

Contrastive losses. When trained on large-scale image-text datasets, the resulting models possess strong semantic understanding of the image data [84].

Reconstructive Losses. When it comes to dense prediction, CLIP-style models struggle to attain the same strong performance [87, 88, 102]. Hence, we also consider image encoders learned using reconstructive losses, as such losses explicitly capture all parts of an image. In particular, we utilize AIM [26], which has shown that an autoregressive reconstructive loss on image data alone scales well.

Encoder Lesson: Image resolution has the highest impact, followed by model size and training data composition. As we can see in Table 1, increasing image resolution from 224 to 336 results in approx. 3% boost in all metrics across all architectures. Increasing the model size from ViT-L to ViT-H, a doubling in parameters, results in a modest performance increase of usually less than 1%. Finally, adding VeCap-300M [51], a dataset of synthetic captions, yields more than 1% boost in few-shot scenarios.

When it comes to model type, the results are less conclusive. Contrastive methods tend to result in higher performance than reconstructive. In particular, encoders based on ViT-L of 300M parameters result in 0.3% to 1.5% performance gain compared to AIM_{600M} of comparable size (only 20 of the 24 AIM model layers are used at inference). This lesson is, nevertheless, inconclusive for the potential of AIM as it has been trained on less than half the data. Similarly, the widely used open sourced OpenAI model [84] perform on-par with our model of comparable capacity but trained on DFN+VeCap data mixture.

Vision-Language Connector and Image Resolution. The goal of this component is to translate the visual representation to the space of the LLM. As image encoders are ViTs, their output is either a single embedding, or a set of grid-arranged embeddings corresponding to the input image patches. Therefore, the

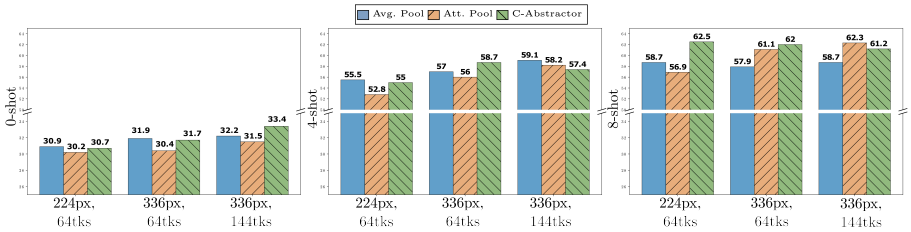


Fig. 4: 0-shot, 4-shot, and 8-shot ablations across different visual-language connectors for two image resolutions, and two image token sizes.

spatial arrangement of the image tokens needs to be converted to the sequential one of the LLM. At the same time, the actual image token representations are to be mapped to the word embedding space.

We consider using 64 or 144 tokens to represent the image, as well as two different image resolutions, 224 and 336. Further, we consider the following architectural options:

Average Pooling. Following [96], we apply $n \times n$ average pooling on the output of the ViT image encoder, followed by a linear projection ($n \in \{8, 12\}$).

Attention Pooling. Motivated by the fact that image token representations are in a different space than the LLM input embeddings, attention pooling using k learnable queries, is a natural approach. By varying k one can vary the number of inputs from a single image that are fed into the LLM (we use $k \in \{64, 144\}$).

Convolutional Mapping. More recently, Honeybee [9] has studied the above questions and proposed the C-Abstractor module. It is implemented as a ResNet [36] block that preserves local information while through adaptive pooling can change the number of image tokens.

VL Connector Lesson: Number of visual tokens and image resolution matters most, while the type of VL connector has little effect. The results shown in Figure 4 demonstrate that both zero- and few-shot performance increases as we increase the number of visual tokens or/and image resolution. However, contrary to what has been reported in the literature [9], different architectural designs do not appear to conclusively produce stronger models. After instruction tuning, all three architectures achieve very similar results at the 336px and 144 token setting. (See Appendix for fine-tuning results.)

3.3 Pre-training Data Ablation

In the following, we elaborate our pre-training data choices (see Figure 3, right). Two types of data are commonly used to train MLLMs: captioning data consisting of images with paired text descriptions; and interleaved image-text documents from the web. We also include text-only data to help preserve the language understanding capabilities of the underlying pre-trained LLM. The full list of datasets is summarized in Table 2. We use the same model setup for ablations described in Section 3.1, with the only exception that we train 200k steps to fully

Data Type	Sources	Size
Captioned Images	CC3M [92], CC12M [10], HQIPT-204M [87], COYO [8], Web Image-Text-1B (Internal)	2B image-text pairs
Captioned Images (Synthetic)	VeCap [51]	300M image-text pairs
Interleaved Image-Text	OBELICS [52], Web Interleaved (Internal)	600M documents
Text-only	Webpages, Code, Social media, Books, Encyclopedic, Math	2T tokens

Table 2: List of datasets for pre-training multimodal large language models.

leverage large-scale training. We also incorporate a set of commonly employed text tasks, referred to as TextCore, as part of the evaluation to better assess the effects of data mixture.

Data Lesson 1: Interleaved data is instrumental for few-shot and text-only performance, while captioning data lifts zero-shot performance. In Figure 5a, we present results across different mixes of interleaved and captioned data. Zero-shot performance increases consistently, from 25.8% to 39.3%, as we increase the amount of captioned data. At the same time, however, for 4- and 8-shot performance, having at least 50% of the data being interleaved is crucial to maintain over 61% for 8-shot or 58% for 4-shot. Without it, performance drops drastically to 45% and 43.7%, respectively. Since interleaved data naturally contains multiple images and accompanying text which are often interrelated, such data is inherently similar to few-shot test inputs, which aligns well with empirical results. However, due to the nature of common evaluation being heavily tailored to captioning problems (3 out of the 8 benchmarks are captioning), captioning data notably lifts zero-shot performance. Interestingly, the use of interleaved data further boosts performance on these very same captioning benchmarks in few-shot settings. Similarly, text-only performance benefits from interleaved data, likely as interleaved data contains long-form text as well.

Data Lesson 2: Text-only data helps with few-shot and text-only performance. As seen in Figure 5b, combining text-only and captioned data boost few-shot performance. In other words, long text does allow the model to utilize multiple image and text examples as context to perform better VQA and captioning. On the other side, combining text-only with interleaved data leads to a drop in performance, albeit a minor one. In both cases, text-only performance is increased as shown in the boost of TextCore numbers.

Data Lesson 3: Careful mixture of image and text data can yield optimal multimodal performance and retain strong text performance. In Figure 5c, we experiment with several mixing ratios between image (caption and interleaved) and text-only data. We see that with caption/interleaved/text ratio 5:5:1, we achieve a good balance of strong multimodal performance while still keeping comparable text-only understanding performance.

Data Lesson 4: Synthetic data helps with few-shot learning. At last, we study the importance of the synthetic caption data, VeCap [51]. It is of higher quality, but relatively small, being only 7% compared to all caption data. As

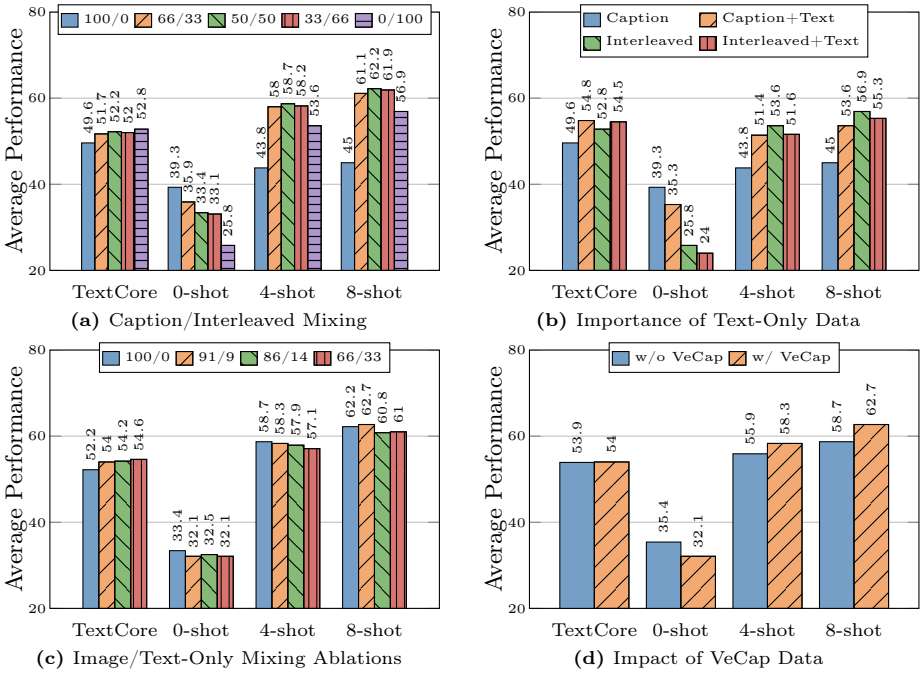


Fig. 5: Data Ablations. For each ablation, we present four metrics: TextCore, 0-shot, 4-shot, and 8-shot. **(a)** Results with image data where we present five different mixing ratios between interleaved and captioned data. **(b)** Results with and without text-only data. We mix the text-only data separately with captioned and interleaved data. **(c)** Results with different mixing ratios between image data (caption and interleaved) and text-only data. **(d)** Results with and without VeCap as part of caption data.

shown in Figure 5d, it does give a non-trivial boost in few-shot performance, of 2.4% and 4% absolute.

4 Final Model and Training Recipe

We collect the results from the previous ablations to determine the final recipe for MM1 multimodal pre-training:

- **Image Encoder**: Motivated by the importance of image resolution, we use a ViT-H [23] model with 378×378 resolution, pre-trained with a CLIP objective on DFN-5B [27].
- **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 [9].
- **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.

In order to improve the model performance, we scale up the LLM size to 3B, 7B, and 30B parameters. We initialize both the image encoder and the underlying LLM decoder weights for MM1 from in-house pre-trained models. We then perform multimodal pre-training on the above data mix for 200k steps (approx. 400B tokens). All models are pre-trained entirely unfrozen with sequence length 4096, up to 16 images per sequence at 378×378 resolution, with a batch size of 512 sequences. All models are trained using the AXLearn framework.[1]

Model Scaling. Using established scaling characteristics of LLMs [38, 39, 107, 108], we perform a grid search of learning rate at small scale, 9M, 85M, 302M, and 1.2B, while using the components identified in Sec. 3.2 to identify optimal learning rate and extrapolate it to larger scale. We use a linear regression in log space to extrapolate from smaller to larger models, resulting in the following prediction of optimal peak learning rate η given the number of (non-embedding) parameters N:

$$\eta = \exp(-0.4214 \ln(N) - 0.5535) \tag{1}$$

For $N = 3e^{10}$, this fit predicts $\eta = 2.2e^{-5}$, which is what we use for the final MM1-30B. Similar to [43], we found in preliminary experiments that validation loss wasn't strongly correlated with downstream task performance. Therefore, we directly use downstream 8-shot average performance for curve fitting.

Scaling via Mixture-of-Experts (MoE). MoE scales the total number of model parameters while keeping the activated parameters constant. It enjoys a larger model capacity without sacrificing inference speed significantly. Recently, MoE has shown promising results in language [19,25,28,44,121], multimodal [64, 81] and computer vision [13, 21, 49, 89] tasks.

In experiments, we further explore scaling the dense model by adding more experts in the FFN layers of the language model. Our MoE implementation generally follows GShard [53] and ST-MoE [121]. Specifically, we design two MoE models, a 3B-MoE using 64 experts that replaces a dense layer with a sparse layer in every-2 layers and a 7B-MoE using 32 experts that replaces a dense layer with a sparse layer in every-4 layers. The 3B-MoE contains 64B parameters in total and the 7B-MoE contains 47B parameters in total. We adopt top-2 gating with a load balance loss term with a 0.01 coefficient to encourage a better expert load balance and adopt a router z-loss term with a 0.001 coefficient to stabilize training. To convert a dense model to MoE, we only replace the dense language decoder with an MoE language decoder. The image encoder and the vision-language connector are kept the same. To train an MoE, we adopt the same training hyperparameters that are discovered for the dense backbone and identical training settings including training data and training tokens.

Multimodal Pre-training Results. We evaluate pre-trained models on captioning and VQA tasks via appropriate prompting. We evaluate zero- and few-shot, as shown in Table 3, and compare against the few approaches that report few-shot pre-training performance. When it comes to few-shot performance,

[1] `https://github.com/apple/axlearn`

Model	Shot	Captioning			Visual Question Answering			
		COCO	NoCaps	TextCaps	VQAv2	TextVQA	VizWiz	OKVQA
MM1-3B Model Comparisons								
Flamingo-3B [3]	0†	73.0	–	–	49.2	30.1	28.9	41.2
	8	90.6	–	–	55.4	32.4	38.4	44.6
MM1-3B	0	73.5	55.6	63.3	46.2	29.4	15.6	26.1
	8	**114.6**	**104.7**	**88.8**	**63.6**	**44.6**	**46.4**	**48.4**
MM1-7B Model Comparisons								
IDEFICS-9B [52]	0†	46.0*	36.8	25.4	50.9	25.9	35.5	38.4
	8	97.0*	86.8	63.2	56.4	27.5	40.4	47.7
Flamingo-9B [3]	0†	79.4	–	–	51.8	31.8	28.8	44.7
	8	99.0	–	–	58.0	33.6	39.4	50.0
Emu2-14B [96]	0†	–	–	–	52.9	–	34.4	42.8
	8	–	–	–	59.0	–	43.9	–
MM1-7B	0	76.3	61.0	64.2	47.8	28.8	15.6	22.6
	8	**116.3**	**106.6**	**88.2**	**63.6**	**46.3**	**45.3**	**51.4**
MM1-30B Model Comparisons								
IDEFICS-80B [52]	0†	91.8*	65.0	56.8	60.0	30.9	36.0	45.2
	8	114.3*	105.7	77.6	64.8	35.7	46.1	55.1
	16	116.6*	107.0	81.4	65.4	36.3	48.3	56.8
Flamingo-80B [3]	0†	84.3	–	–	56.3	35.0	31.6	50.6
	8	108.8	–	–	65.6	37.3	44.8	57.5
	16	110.5	–	–	66.8	37.6	48.4	57.8
Emu2-37B [96]	0	–	–	–	33.3	26.2	40.4	26.7
	8	–	–	–	67.8	49.3	54.7	54.1
	16	–	–	–	68.8	50.3	57.0	57.1
MM1-30B	0	70.3	54.6	64.9	48.9	28.2	14.5	24.1
	8	123.1	111.6	92.9	70.9	49.4	49.9	58.3
	16	**125.3**	**116.0**	**97.6**	**71.9**	**50.6**	**57.9**	**59.3**

Table 3: Multimodal pre-training evaluations. (*) IDEFICS includes PMD in its training data (includes COCO). (†) These models include two text-only demonstrations in their "0" prompt, whereas MM1 does not. For the full table, please refer to Appendix.

MM1 outperforms all published prior work for pre-trained MLLMs. We see superior performance at 30B across captioning benchmarks and the VizWiz-QA benchmark. On VQAv2, TextVQA, OKVQA, at that scale we are comparable to Emu2 [96]. For zero-shot performance, even without instruction fine-tuning, our models perform favorably on TextCaps across all model sizes, and comparable to Flamingo-3B at small scales for most benchmarks.

5 Supervised Fine-Tuning

In this section, we describe the supervised fine-tuning (SFT) experiments trained on top of the pre-trained models described in the previous sections.

SFT Data Mixture. We follow LLaVA-1.5 [68] and LLaVA-NeXT [69], and collect roughly 1.45M SFT examples from a diverse set of datasets, including

- Instruction-response pairs generated by GPT-4 and GPT-4V, including LLaVA-Conv and LLaVA-Complex [70] for conversations and complex reasoning, and ShareGPT-4V [12] for detailed image descriptions;
- Academic task oriented vision-language (VL) datasets, including (*i*) VQAv2 [33], GQA [41], OKVQA [76], A-OKVQA [90], and COCO Captions [15] for natural images; (*ii*) OCRVQA [80], and TextCaps [94] for text-rich images; and (*iii*) DVQA [45], ChartQA [77], AI2D [46], DocVQA [79], InfoVQA [78], and Synthdog-En [47] for document and chart understanding.
- Text-only SFT data: We include an internal text-only dataset to ensure the model is capable of text-only instruction following.

During SFT, we keep both the image encoder and the LLM backbone *unfrozen*; other SFT training details are provided in Appendix.

Scaling to Higher Resolutions. Intuitively, higher image resolution leads to better performance. To support high-resolution SFT, we use two approaches:

Positional embedding interpolation, *e.g.*, as explored in Qwen-VL [5] and BLIP2 [59]. After positional embedding interpolation, the vision transformer backbone is adapted to the new resolution during fine-tuning. Through this method, we have fine-tuned our model to support image resolutions ranging from 448×448, 560×560, to 672×672. Note that, for a resolution of 672×672, with a patch size of 14×14, an image is represented with $2,304$ tokens.

Sub-image decomposition, recently introduced by SPHINX [67], Monkey [63], and LLaVA-NeXT [69]. Computing self-attention among more than $2,000$ image tokens is computationally challenging, limiting further scaling to even higher image resolutions. Following SPHINX [67], as shown in Figure 6a, for a high-resolution input image, *e.g.*, 1344×1344, we construct five images of 672×672, and feed them as independent images into our visual encoder. Specifically, we first downsample the input image to 672×672 as a high-level representation, and also resize the input image to 1344×1344 and divide the resized image into 4 sub-images of 672×672, which preserve more detailed visual information. Using positional embedding interpolation for each sub-image, we can support image resolution as high as 1792×1792 in experiments.

5.1 SFT Results

Comparison with SOTA. Results are summarized in Table 4. We use "-Chat" to denote our MM1 models after SFT. First, on average, MM1-3B-Chat and MM1-7B-Chat outperforms all listed models of the same size, setting a new state of the art for these model sizes. MM1-3B-Chat and MM1-7B-Chat show particularly strong performance on VQAv2, TextVQA, ScienceQA, and also the more recent benchmarks (MMMU and MathVista).

Second, we explore two MoE models: (*i*) 3B-MoE with 64 experts, and (*ii*) 7B-MoE with 32 experts. Our MoE models achieve uniformly better performance

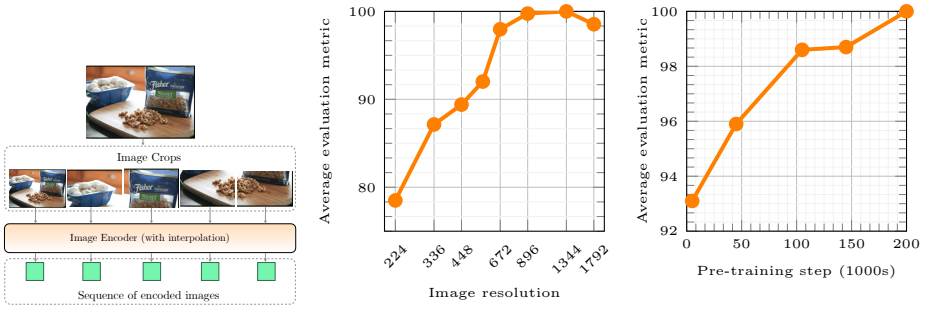
Model	VQA^{v2}	VQA^{T}	SQA^{I}	MMMU	MathV	MME^{P}	MME^{C}	MMB	SEED	POPE	$LLaVA^{W}$	MM-Vet
3B Model Comparison												
MobileVLM [17]	–	47.5	61.0	–/–	–	1288.9	–	59.6	–/–	84.9	–	–
LLaVA-Phi [120]	71.4	48.6	68.4	–/–	–	1335.1	–	59.8	–/–	85.0	–	28.9
Imp-v1 [91]	79.45	59.38	69.96	–/–	–	1434.0	–	66.49	–	88.02	–	33.1
TinyLLaVA [118]	79.9	59.1	69.1	–/–	–	1464.9	–	66.9	–/–	86.4	75.8	32.0
Bunny [37]	79.8	–	70.9	38.2/33.0	–	1488.8	289.3	68.6	62.5/–	86.8	–	–
Gemini Nano-2 [97]	67.5	65.9	–	32.6/–	30.6	–	–	–	–	–	–	–
MM1-3B-Chat	82.0	71.9	69.4	33.9/33.7	32.0	1482.5	279.3	67.8	63.0/68.8	87.4	72.1	43.7
MM1-3B-MoE-Chat	82.5	72.9	76.1	38.6/35.7	32.6	1469.4	303.1	70.8	63.9/69.4	87.6	76.8	42.2
7B Model Comparison												
InstructBLIP-7B [20]	–	50.1	60.5	–/–	25.3	–	–	36.0	53.4/–	–	60.9	26.2
Qwen-VL-Chat-7B [5]	78.2	61.5	68.2	35.9/32.9	–	1487.5	360.7	60.6	58.2/65.4	–	–	–
LLaVA-1.5-7B [68]	78.5	58.2	66.8	–/–	–	1510.7	316.1	64.3	58.6/66.1	85.9	63.4	31.1
ShareGPT4V-7B [12]	80.6	60.4	68.4	–/–	–	1567.4	376.4	68.8	–/–	–	72.6	–
LVIS-Ins4V-7B [103]	79.6	58.7	68.3	–/–	–	1528.2	–	66.2	60.6/–	86.0	67.0	31.5
VILA-7B [65]	79.9	64.4	68.2	–/–	–	1531.3	–	68.9	61.1/–	85.5	69.7	34.9
SPHINX-Intern2 [31]	75.5	–	70.4	–/–	35.5	1260.4	294.6	57.9	68.8/–	86.9	57.6	36.5
LLaVA-NeXT-7B [69]	81.8	64.9	70.1	35.8/–	34.6	1519	332	67.4	–/70.2	86.53	81.6	43.9
MM1-7B-Chat	82.8	72.8	72.6	37.0/35.6	35.9	1529.3	328.9	72.3	64.0/69.9	86.6	81.5	42.1
MM1-7B-MoE-Chat	83.4	73.8	74.4	40.9/37.9	40.9	1597.4	394.6	72.7	65.5/70.9	87.8	84.7	45.2
30B Model Comparison												
Emu2-Chat-37B [96]	84.9	66.6	–	36.3/34.1	–	–	–	–	62.8/–	–	–	48.5
CogVLM-30B [104]	83.4	68.1	–	32.1/30.1	–	–	–	–	–	–	–	56.8
LLaVA-NeXT-34B [69]	83.7	69.5	81.8	51.1/44.7	46.5	1631	397	79.3	–/75.9	87.73	89.6	57.4
MM1-30B-Chat	83.7	73.5	81.0	44.7/40.3	39.4†	1637.6	431.4	75.1	65.9/72.1	87.6	89.3	48.7
Gemini Pro [97]	71.2	74.6	–	47.9/–	45.2	–	436.79	73.6	–/70.7	–	–	64.3
Gemini Ultra [97]	77.8	82.3	–	59.4/–	53.0	–	–	–	–	–	–	–
GPT4V [1]	77.2	78.0	–	56.8/55.7	49.9	–	517.14	75.8	67.3/69.1	–	–	67.6

Table 4: Comparison with SOTA models on MLLM benchmarks. VQA^{v2} [33]; VQA^{T}: TextVQA [95]; SQA^{I}: ScienceQA-IMG [75]; MMMU [114]; MathV: MathVista [74]; $MME^{P/C}$: the Perception/Cognition split of MME [29]; MMB: MMBench [72]; SEED: SEED-Bench [56]; POPE [62]; $LLaVA^{W}$: LLaVA-Bench (In-the-Wild) [70]; MM-Vet [113]. The two numbers reported in MMMU denote the performance on the val and test split, respectively. The two numbers reported in SEED denote the performance on the whole SEED-Bench and the image part, respectively. (†) 8-shot prompting: 44.4.

than the dense counterpart on almost every benchmark. This shows the great potential of MoE for further scaling, which is left as future work.

Third, for the 30B model size, MM1-30B-Chat outperforms Emu2-Chat-37B [96] and CogVLM-30B [104] on TextVQA, SEED, and MMMU. Compared with LLaVA-NeXT [69], we also achieve competitive performance. However, LLaVA-NeXT does not support multi-image reasoning, nor few-shot prompting, as each image is represented as 2,880 tokens, while ours is only 720 in total.

Impact of Image Resolution. Figure 6b shows the impact of input image resolution on SFT performance. Compared to a baseline model with an image resolution of 336 pixels, we can achieve a 15% relative increase by supporting an image resolution of 1344×1344. Note that for the largest image resolution of 1792×1792, average performance decreases slightly. This is likely because many of the evaluation images are smaller than this resolution, and resizing artifacts may affect the model performance. By default, the results in Table 4 correspond to image resolutions of 1344×1344.



(a) High resolution image input processing.

(b) Impact of image resolution on SFT performance.

(c) Impact of pre-training on SFT performance.

Fig. 6: We study the impact of image resolution and pre-training for SFT performance.

Impact of Pre-training. In contrast to most recent MLLMs, we perform large-scale pre-training for our models. To assess the impact of pre-training on the final model performance, we perform SFT on the same pre-training run, but at different checkpoint steps. For an earlier checkpoint step, the model has seen less unique data samples than a later checkpoint step, so this is a measure of the importance of the quantity of pre-training data. In Figure 6c, we show that the model consistently improves as it has seen more pre-training data. Furthermore, large-scale multimodal pre-training enables strong in-context few-shot learning and multi-image reasoning capabilities, while most MLLM benchmarks shown in Table 4 focus on zero-shot metrics and single-image reasoning.

Few-shot Chain-of-Thought Reasoning after SFT. As seen in Section 3.3, MM1 gains few-shot capabilities thanks to interleaved data. Even though our fine-tuning data includes only single-image examples, we find that MM1-30B-Chat still exhibits multi-image reasoning. This is shown qualitatively in Figure 2, and quantitatively on MathVista [74], where we evaluate few-shot performance with chain-of-thought prompting: 4-shot performance is **41.9**, which is 2.5 points higher than zero-shot (**39.4**). To allow for more examples, we explore a *mixed resolution in-context examples* formulation, where we feed some of the examples at a lower resolution (see Appendix for details). Using this formulation with 8 in-context examples increases the performance on MathVista to **44.4**.

6 Conclusion

We study how to build performant MLLMs. Through carefully ablating modeling and data choices, we identify important lessons that yield a pre-trained model achieving SOTA results on a range of few-shot evaluations. After SFT, this model family produces competitive performance on a wide range of benchmarks, while enabling multi-image reasoning and few-shot prompting. We hope that the identified lessons will help the community in building strong models beyond any single specific model architecture or data strategy.

Acknowledgements

The authors would like to thank Floris Weers for his assistance with multimodal evaluation infrastructure; Vaishaal Shankar, Alaa El-Nouby, Yang Zhao, Shuangfei Zhai, Russ Webb, Hadi Pouransari, Hong-You Chen, Yanghao Li, and David Mizrahi for valuable guidance, suggestions, and feedback; Chen Chen and Qibin Chen for help on instruction tuning; Maitreyi Kunnavakkam Vinjimur, Megan Maher Welsh, Bhavika Devnani, and David Koski for their assistance with input pipelines and data processing; Guoli Yin, Tom Nickson and Michael Tu for assistance with the AXLearn infrastructure and early LLM work; Ankur Jain and Varsha Mohan Paidi for assistance with dataset creation and filtering; Esteban Gonzalez, Ian Clark, Jack Bailin, David Koski, and in particular Venkata Yerneni for assistance with the internal Weights & Biases instance for tracking experiments and model evaluations.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: ICCV (2019)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning (2022)
4. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
6. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS (2020)
8. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset` (2022)
9. Cha, J., Kang, W., Mun, J., Roh, B.: Honeybee: Locality-enhanced projector for multimodal llm. arXiv preprint arXiv:2312.06742 (2023)

10. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021)
11. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
12. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023)
13. Chen, T., Chen, X., Du, X., Rashwan, A., Yang, F., Chen, H., Wang, Z., Li, Y.: Adamv-moe: Adaptive multi-task vision mixture-of-experts. In: ICCV (2023)
14. Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C.R., Goodman, S., Wang, X., Tay, Y., et al.: Pali-x: On scaling up a multilingual vision and language model. arXiv preprint arXiv:2305.18565 (2023)
15. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
16. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. JMLR (2023)
17. Chu, X., Qiao, L., Lin, X., Xu, S., Yang, Y., Hu, Y., Wei, F., Zhang, X., Zhang, B., Wei, X., et al.: Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886 (2023)
18. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
19. Dai, D., Deng, C., Zhao, C., Xu, R.X., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., Xie, Z., Li, Y.K., Huang, P., Luo, F., Ruan, C., Sui, Z., Liang, W.: Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066 (2024)
20. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
21. Daxberger, E., Weers, F., Zhang, B., Gunter, T., Pang, R., Eichner, M., Emmersberger, M., Yang, Y., Toshev, A., Du, X.: Mobile v-moes: Scaling down vision transformers via sparse mixture-of-experts (2023)
22. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
24. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: PaLM-E: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
25. Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., Zoph, B., Fedus, L., Bosma, M.P., Zhou, Z., Wang, T., Wang, E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q., Wu, Y., Chen, Z., Cui, C.: GLaM: Efficient scaling of language models with mixture-of-experts. In: ICML (2022)

26. El-Nouby, A., Klein, M., Zhai, S., Bautista, M.A., Shankar, V., Toshev, A., Susskind, J., Joulin, A.: Scalable pre-training of large autoregressive image models. arXiv preprint arXiv:2401.08541 (2024)
27. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
28. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity (2022)
29. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
30. Fu, T.J., Hu, W., Du, X., Wang, W.Y., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models. arXiv preprint arXiv:2309.17102 (2023)
31. Gao, P., Zhang, R., Liu, C., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., et al.: Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. arXiv preprint arXiv:2402.05935 (2024)
32. Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., Chen, K.: Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:2305.04790 (2023)
33. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
34. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: CVPR (2018)
35. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
36. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
37. He, M., Liu, Y., Wu, B., Yuan, J., Wang, Y., Huang, T., Zhao, B.: Efficient multimodal learning from data-centric perspective. arXiv preprint arXiv:2402.11530 (2024)
38. Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T.B., Dhariwal, P., Gray, S., et al.: Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701 (2020)
39. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L.: Training compute-optimal large language models (2022)
40. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., Wei, F.: Language is not all you need: Aligning perception with language models (2023)
41. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
42. IDEFICS: Introducing idefics: An open reproduction of state-of-the-art visual language model. `https://huggingface.co/blog/idefics` (2023)
43. Isik, B., Ponomareva, N., Hazimeh, H., Paparas, D., Vassilvitskii, S., Koyejo, S.: Scaling laws for downstream task performance of large language models (2024)

44. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mixtral of experts (2024)
45. Kafle, K., Price, B., Cohen, S., Kanan, C.: Dvqa: Understanding data visualizations via question answering. In: CVPR (2018)
46. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: ECCV (2016)
47. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: ECCV (2022)
48. Koh, J.Y., Fried, D., Salakhutdinov, R.: Generating images with multimodal language models. arXiv preprint arXiv:2305.17216 (2023)
49. Komatsuzaki, A., Puigcerver, J., Lee-Thorp, J., Ruiz, C.R., Mustafa, B., Ainslie, J., Tay, Y., Dehghani, M., Houlsby, N.: Sparse upcycling: Training mixture-of-experts from dense checkpoints. In: ICLR (2023)
50. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023)
51. Lai, Z., Zhang, H., Wu, W., Bai, H., Timofeev, A., Du, X., Gan, Z., Shan, J., Chuah, C.N., Yang, Y., et al.: From scarcity to efficiency: Improving clip training via visual-enriched captions. arXiv preprint arXiv:2310.07699 (2023)
52. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: Obelics: An open web-scale filtered dataset of interleaved image-text documents (2023)
53. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: {GS}hard: Scaling giant models with conditional computation and automatic sharding. In: ICLR (2021)
54. Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., Liu, Z.: Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425 (2023)
55. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)
56. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
57. Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J.: Multimodal foundation models: From specialists to general-purpose assistants. arXiv preprint arXiv:2309.10020 (2023)
58. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
59. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
60. Li, L., Yin, Y., Li, S., Chen, L., Wang, P., Ren, S., Li, M., Yang, Y., Xu, J., Sun, X., et al.: M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. arXiv preprint arXiv:2306.04387 (2023)
61. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)

62. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
63. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. arXiv preprint arXiv:2311.06607 (2023)
64. Lin, B., Tang, Z., Ye, Y., Cui, J., Zhu, B., Jin, P., Huang, J., Zhang, J., Ning, M., Yuan, L.: Moe-llava: Mixture of experts for large vision-language models (2024)
65. Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. arXiv preprint arXiv:2312.07533 (2023)
66. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., Zitnick, C.L.: Microsoft COCO: common objects in context. arXiv preprint arXiv:1405.0312 (2014)
67. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023)
68. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
69. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
70. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
71. Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al.: Llava-plus: Learning to use tools for creating multimodal agents. arXiv preprint arXiv:2311.05437 (2023)
72. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
73. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS (2019)
74. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
75. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. NeurIPS (2022)
76. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019)
77. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022)
78. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: WACV (2022)
79. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: WACV (2021)
80. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: ICDAR (2019)
81. Mustafa, B., Ruiz, C.R., Puigcerver, J., Jenatton, R., Houlsby, N.: Multimodal contrastive learning with LIMoe: the language-image mixture of experts. In: NeurIPS (2022)

82. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
83. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
84. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
85. Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al.: Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446 (2021)
86. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020)
87. Ranasinghe, K., McKinzie, B., Ravi, S., Yang, Y., Toshev, A., Shlens, J.: Perceptual grouping in contrastive vision-language models. In: ICCV (2023)
88. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: CVPR (2022)
89. Ruiz, C.R., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A.S., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) NeurIPS (2021)
90. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: ECCV (2022)
91. Shao, Z., Ouyang, X., Yu, Z., Yu, J.: Imp: An emprical study of multimodal small language models (2024), `https://huggingface.co/MILVLG/imp-v1-3b`
92. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
93. Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., Catanzaro, B.: Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053 (2019)
94. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: ECCV (2020)
95. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: CVPR (2019)
96. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., et al.: Generative multimodal models are in-context learners. arXiv preprint arXiv:2312.13286 (2023)
97. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
98. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., et al.: Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
99. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. arXiv preprint arXiv:2401.06209 (2024)

100. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
101. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. NeurIPS (2021)
102. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. arXiv preprint arXiv:2312.01597 (2023)
103. Wang, J., Meng, L., Weng, Z., He, B., Wu, Z., Jiang, Y.G.: To see is to believe: Prompting gpt-4v for better visual instruction tuning. arXiv preprint arXiv:2311.07574 (2023)
104. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023)
105. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. arXiv preprint arXiv:2305.11175 (2023)
106. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
107. Yang, G., Hu, E.J.: Feature learning in infinite-width neural networks. arXiv preprint arXiv:2011.14522 (2020)
108. Yang, G., Hu, E.J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., Gao, J.: Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer (2022)
109. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., et al.: mplug-docowl: Modularized multimodal large language model for document understanding. arXiv preprint arXiv:2307.02499 (2023)
110. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
111. Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv preprint arXiv:2311.04257 (2023)
112. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. In: ICLR (2024)
113. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
114. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502 (2023)
115. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Zhang, L., Li, C., et al.: Llava-grounding: Grounded visual chat with large multimodal models. arXiv preprint arXiv:2312.02949 (2023)
116. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
117. Zhao, B., Wu, B., Huang, T.: Svit: Scaling up visual instruction tuning. arXiv preprint arXiv:2307.04087 (2023)

118. Zhou, B., Hu, Y., Weng, X., Jia, J., Luo, J., Liu, X., Wu, J., Huang, L.: Tinyllava: A framework of small-scale large multimodal models. arXiv preprint arXiv:2402.14289 (2024)
119. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
120. Zhu, Y., Zhu, M., Liu, N., Ou, Z., Mou, X., Tang, J.: Llava-phi: Efficient multi-modal assistant with small language model. arXiv preprint arXiv:2401.02330 (2024)
121. Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., Shazeer, N., Fedus, W.: St-moe: Designing stable and transferable sparse expert models (2022)