VILA: On Pre-training for Visual Language Models

Ji Lin[1,2 * †] Hongxu Yin[1 *] Wei Ping[1] Yao Lu[1] Pavlo Molchanov[1]
Andrew Tao [1] Huizi Mao [1] Jan Kautz[1] Mohammad Shoeybi[1] Song Han[1,2]
[1]NVIDIA [2]MIT

Abstract

Visual language models (VLMs) rapidly progressed with the recent success of large language models. There have been growing efforts on visual instruction tuning *to extend the LLM with visual inputs, but lacks an in-depth study of the visual language* pre-training *process, where the model learns to perform joint modeling on both modalities. In this work, we examine the design options for VLM pre-training by augmenting LLM towards VLM through step-by-step controllable comparisons. We introduce three main findings: (1) freezing LLMs during pre-training can achieve decent zero-shot performance, but lack in-context learning capability, which requires unfreezing the LLM; (2) interleaved pre-training data is beneficial whereas image-text pairs alone are not optimal; (3) re-blending text-only instruction data to image-text data during instruction fine-tuning not only remedies the degradation of text-only tasks, but also boosts VLM task accuracy. With an enhanced pre-training recipe we build **VILA**, a **Vi**sual **La**nguage model family that consistently outperforms the state-of-the-art models, e.g., LLaVA-1.5, across main benchmarks without bells and whistles. Multi-modal pre-training also helps unveil appealing properties of VILA, including multi-image reasoning, enhanced in-context learning, and better world knowledge.*

1. Introduction

Large language models (LLMs) have demonstrated superior capabilities for natural language tasks [4, 8, 10, 15, 16, 19, 31, 45, 50, 58–60]. Augmenting LLMs to support visual inputs allows the final model to inherit some of the appealing properties like instruction following, zero-shot generalization, and few-shot in-context learning (ICL), empowering various visual language tasks [1, 2, 6, 9, 14, 20, 35, 39, 70]. The central challenge of unifying vision and language for collaborative inference resides in connecting the LLM and the vision foundation model (*e.g.*, a CLIP encoder): both foundation models are usually pre-trained individually, before aligned via vision-language joint training. Most of the efforts in this field have been focusing on improving the visual language instruction-tuning process, *i.e.*, supervised fine-tuning (SFT) or reinforcement learning from human feedback (RLHF) [38, 39, 56]. However, there lacks a thorough study of the pre-training process, where the model is trained on image-text datasets/corpora at scale [11, 53, 71]. This process is costly but critical for the modality alignment.

In this work, we aim to explore different design options for enhanced visual language model pre-training. In particular, we aim to answer *"How do various design choices in visual language model pre-training impact the downstream performance?"* We followed the pre-training + SFT pipeline and ablated different design options for pre-training overseeing dataset properties and training protocols. We discover several findings: (1) Freezing the LLM during pre-training can achieve a decent zero-shot performance, but not in-context learning (ICL) capability, whereas updating the LLMs encourages deep embedding alignment, which we

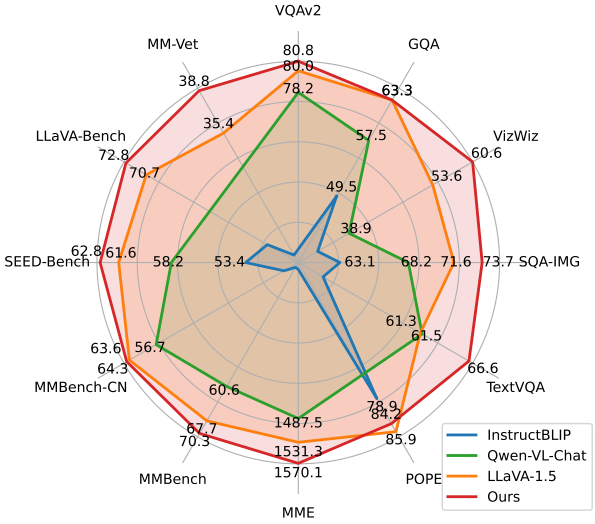


Figure 1. VILA's enhanced visual-language pre-training consistently improves the downstream task accuracy under a comparison to recent methods [8, 18, 39].

* Equal contribution. † Work done during an internship at NVIDIA.

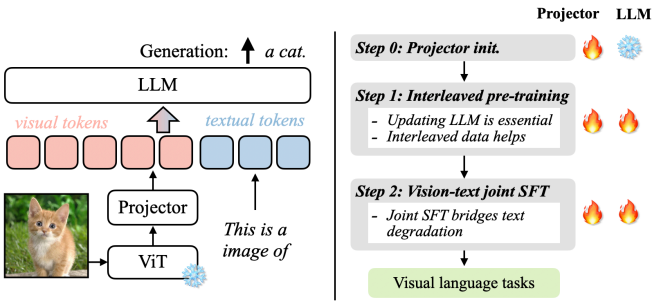


Figure 2. We study auto-regressive visual language model, where images are tokenized and fed to the input of LLMs. We find updating the LLM is essential for in-context learning capabilities, and interleaved corpus like [71] helps pre-training. Joint SFT with text-only data helps maintain the text-only capabilities.

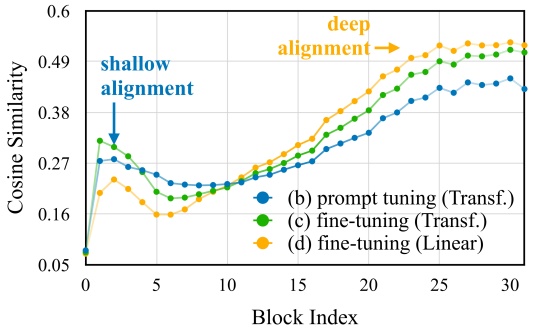


Figure 3. Prompt-tuning to support visual tokens can only enable shallow alignment, while fine-tuning the LLM leads to alignment at deeper layers. From configuration (b) to (d) (as in Table 1), the alignment improves at deeper layer, so as ICL accuracy (4-shot).

found is important for ICL; (2) Interleaved visual language data is essential for pre-training, that provides accurate gradient update and maintains text-only capability; (3) Adding in text-only instruction data during SFT can further remedy text-only degradation and boost visual language task accuracy.

We introduce practical guidance to design **Vi**sual **La**nguage models, dubbed **VILA**. Without bells and whistles, VILA outperforms the state-of-the-art model [38] by noticeable margins across a wide range of vision language tasks (Figure 1), thanks to the help of improved pre-training. Moreover, we observe that the pre-training process unlocked several interesting capabilities for the model, such as (i) multi-image reasoning (despite the model only sees single image-text pairs during SFT), (ii) stronger in-context learning capabilities, and (iii) enhanced world knowledge. We hope our findings can provide a good pre-training recipe for future visual language models.

2. Background

Model architecture. Multi-modal LLMs can be generally categorized into two settings: cross-attention-based [6, 35] and auto-regressive-based [2, 20, 39]. The latter VLM family tokenizes images into visual tokens, which are concatenated with textual tokens and fed as the input to LLMs (*i.e.*, treating visual input as a foreign language). It is a natural extension of text-only LLMs by augmenting the input with visual embeddings and can handle arbitrary interleaved image-text inputs. In this study, we focus on the pre-training of auto-regressive VLMs due to its flexibility and popularity. As shown in Figure 2, auto-regressive VLMs consists of three components: a *visual encoder*, an *LLM*, and a *projector* that bridges the embeddings from the two modalities. The projector can be a simple linear layer [39] or more capable Transformer blocks [7, 18] – we will compare their efficacy in our experiments. The model takes visual and text input and generates text outputs.

Training stages. Following common practice [7, 20, 39], we study how to augment a pre-trained text-only LLM with visual input support. The training can be categorized into three stages:

0. Projector initialization. The LLM and ViT are separately pre-trained, while the projector is usually initialized from random weights. Therefore, we first pre-train the projector while freezing both ViT and LLMs on image-caption pairs following existing literature [18, 35, 39].

1. visual language pre-training. We then pre-train the model (LLM and the projector) on visual language corpus. We consider two types of corpus: interleaved image-text corpus (*e.g.*, MMC4 [71]) and image-text pairs (*e.g.*, COYO [11] and LAION [53]). We focus the study of this work on the pre-training process, which are most costly and important for visual language alignment.

2. Visual instruction-tuning. Finally, we further perform instruction tuning of the pre-trained model on visual language instruction datasets. We convert existing visual language datasets into FLAN [63] style (*i.e.*, with dataset-specific prompts) following [18]. Please find the data blend of the visual instruction data in the supplementary.

Evaluations. During our ablation study, we evaluate the fine-tuned model on 4 visual language tasks: accuracy for OKVQA [44] and TextVQA [54], and CIDEr score for COCO [37] and Flickr [66]. We evaluate both 0-shot and 4-shot performance, which reflects the models' in-context learning capability.

3. On Pre-training for Visual Language Models

In this section, we discuss practical design choices and learned lessons for the visual language pre-training process.

3.1. Updating LLM is Essential

Fine-tuning *vs*. prompt tuning. There are two popular ways to augment a pre-trained text-only LM with visual

	PreT Train LLM?	SFT Train LLM?	Projector	OKVQA		TextVQA		COCO		Flickr		**Average**	
				0-shot	4-shot	0-shot	4-shot	0-shot	4-shot	0-shot	4-shot	0-shot	4-shot
(a)	✗	✗	Transformer	10.4	19.2	14.8	23.1	17.4	60.2	11.0	47.4	13.4	37.5
(b)	✗	✓	Transformer	47.1	47.7	37.2	36.6	109.4	88.0	73.6	58.1	66.8	57.6
(c)	✓	✓	Transformer	44.8	49.8	38.5	38.8	112.3	113.5	71.5	72.9	66.8	68.8
(d)	✓	✓	Linear	45.2	50.3	39.7	40.2	115.7	118.5	74.2	74.7	68.7	70.9

Table 1. Ablation study on whether to train LLM or freeze LLM and only perform prompt tuning during visual language pre-training (PreT). Interestingly, freezing the LLM during pre-training does not hurt the 0-shot accuracy, but leads to worse in-context learning capability (worse 4-shot). Using a simple linear projector forces the LLM to learn more and leads to better generalization. We report accuracy for VQA datasets (OKVQA, TextVQA) and CIDEr score for captioning (COCO and Flickr). *Note*: we used a different evaluation setting just for ablation study; the absolute value in this setting is lower and should not be compared against other work.

inputs: *fine-tune* LLMs on the visual input tokens [20, 39], or *freeze* the LLM and train only the visual input projector as *prompt tuning* [18, 35]. The latter is attractive since freezing the LLMs prevents the degradation of the pre-trained text-only LLM. Nonetheless, we found updating the base LLM is essential to inherit some of the appealing LLM properties like in-context learning.

To verify the idea, we compare the two training protocols in Table 1. We use a Transformer block for the projector instead of a single linear layer [39] in setting a-c, which provides enough capacity when freezing LLMs. We use MMC4-core [71]* for the comparison. We observed that:

(1) Training only the projector during SFT leads to poor performance (setting a), despite using a high-capacity design. It is rewarding to fine-tune LLM during SFT.

(2) Interestingly, freezing the LLM during pre-training does *not* affect *0-shot performance*, but *degrades in-context learning capabilities* (*i.e.*, 4-shot, comparing setting b and c). The gap is even larger for captioning datasets (COCO & Flickr) since they are out-of-distribution (the instruction tuning data is mostly VQA-alike, see supplementary), showing the worse generalization capability when freezing LLMs.

(3) When using a small-capacity projector (a linear layer instead of a Transformer block), the accuracy is slightly better (comparing c and d). We hypothesize a simpler projector forces the LLM to learn more on handling visual inputs, leading to better generalization.

The deep embedding alignment hypothesis. To understand why fine-tuning LLM is beneficial, we hypothesize that it is important to *align the distribution of visual and textual latent embeddings* (especially in the deeper layers), so that the model can seamlessly model the interaction between the two modalities. It is essential if we want to inherit some of the good properties of LLM like in-context learning for visual language applications.

To verify the idea, we calculate the Chamfer distance of visual and textual embeddings in different layers to measure how well they align in Figure 3. We calculate the pairwise cosine similarity to exclude the affect of magnitude. From configuration (b) to (d), the similarity of deeper layer goes higher, so as the 4-shot accuracy in Table 1, showing the positive relationship between deep embedding alignment and in-context learning.

Given the observations, we *fine-tune the LLM during both pre-training and instruction-tuning* in later studies, and use a *simple linear projection* layer.

*We downloaded only 25M of 30M images amid some expired URLs.

Dataset	Type	Text Src.	#img/sample	#tok./img
MMC4 [71]	Interleave	HTML	4.0	122.5
COYO [11]	Img-text pair	Alt-text	1	22.7

Table 2. Two image-text corpus considered for pre-training. The COYO captions are generally very short, which has a different distribution compared to the text-only corpus for LLM training. We sample each data source to contain 25M images by choosing samples with high CLIP similarities.

Pre-train Data	VLM acc (avg)		MMLU acc.
	0-shot	4-shot	
Llama-2	-	-	*46.0%*
COYO	51.1%	50.3%	28.8% (-17.2%)
MMC4-pairs	46.4%	44.5%	32.4% (-13.6%)
MMC4	68.7%	70.9%	40.7% (-5.3%)
MMC4+COYO	**69.0%**	**71.3%**	40.2% (-5.8%)

Table 3. Pre-training on MMC4 data provides better visual language accuracy (0-shot and few-shot) and smaller degradation on text-only accuracy compared to caption data (COYO). The benefits comes from the interleave nature but not the better text distribution (MMC4 *vs*. MMC4-pairs). Blending interleaved and caption data provides a better diversity and downstream accuracy.

..., "Easy to grow, and fruitful, these are a favorite amongst those who like to have their own at home garden.", **<im1>**, "There is a large number of different kinds of **tomatoes**, meaning that there is one for every different size garden.", "You can choose from a variety of these plants to accomplish different food missions, be them small cucumbers for pickling or larger ones for slicing.", ..., **<im2>**, "These tender **green beans** are durable and quite at home growing in the hot season air.", "To get the most out of planting them, space the seeds 4 to 6 inches apart and allow 3 feet between rows, …

Figure 4. A sample from MMC4 [71] dataset consisting of interleaved images and text segments. The images are placed *before* the corresponding text. The text are *weakly conditioned* on images: only colored text can be better inferred with the help of images.

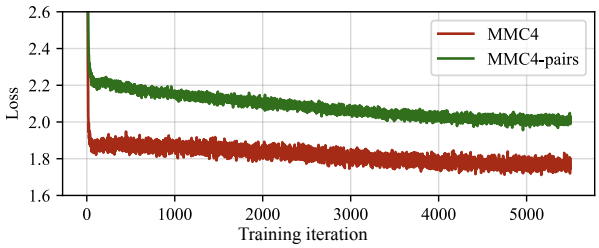


Figure 5. The training loss is lower when pre-training on MMC4 compared to MMC4-pairs (samples broken into image-text pairs), since the text segments provide more information for language modeling.

3.2. Interleaved Visual Language Corpus Helps Pre-training

Our goal is to "augment" the LLM to support visual input, instead of training a model that *only* works well on visual language inputs. Therefore, it is essential to preserve the text-only capabilities of LLMs. We found that data blending is a key factor, both for pre-training and instruction tuning.

Pre-training dataset options. Most of the VLM pre-training [35, 39, 62] relies on image-text pairs (*i.e.*, image and captions) due to the wide availability and large diversity (*e.g.*, LAION [53], COYO [11]). On the other hand, interleaved image-text datasets (MMC4 [71], M3W [6]) follow a more similar distribution compared to the text-only corpus and is found to be important in Flamingo-style model training [6]. We hypothesize that the interleaved dataset is even *more important* for VLMs when LLM backbone is updated to accommodate the visual input. For a better understanding of the two data types, we compare statistics in Table 2: COYO suffers from a short text distribution since the accompanying text is taken from alt-text. We subsample the COYO dataset by ranking CLIP similarities and keep only 25M images (a similar size as MMC4-core).

We follow the same pre-training + SFT process and ablate different pre-training corpus. We compare the 0-shot and few-shot visual language accuracy as well as text-only accuracy (MMLU [27]) in Table 3. Due to space limit, we report the average accuracy over four datasets (as in Table 1).

Interleaved data is essential. We notice using image-text pairs (*i.e.*, COYO) for pre-training can lead to catastrophic forgetting. The text-only accuracy (MMLU) degrades by 17.2%. The visual language accuracy is also much worse compared to MMC4 pre-training. Noticeably, the 4-shot accuracy is even worse than 0-shot, showing the model cannot properly do in-context learning for visual language inputs (probably because it never sees more than one image during pre-training). We hypothesize the catastrophic forgetting is due to the distribution of text-based captions, which are generally very short and concise.

On the contrary, dataset like MMC4 has a much closer distribution compared to text-only corpus (*e.g.*, C4 [50]). When using the interleaved data for pre-training, the degradation on MMLU is only ~5%. The degradation would be even smaller when using a larger base LLM [20]. With proper instruction tuning (Section 3.3), this degradation can be fully recovered. It also promotes visual in-context learning, leading to a higher 4-shot accuracy compared to 0-shot.

Interleave data structure matters, but not the text distribution. We further question whether the benefits come from the better text distribution (*e.g.*, longer) or from the interleave nature. To ablate this, we construct a new MMC4 variant by only keeping the images and their corresponding text segments, without considering the interleave nature, denoted as "MMC4-pairs". For example an MMC4 sample may look like:

<txt1>**<im1>**<txt2><txt3>**<im2>**<txt4>

It will be converted into two MMC4-pairs samples[†]:

<im1><txt2>, **<im2>**<txt4>

However, training on MMC4-pairs does not lead to a satisfactory result: it slightly reduces the degradation on MMLU due to a longer text distribution, but the VLM accuracy is even lower compared to pre-training on COYO; there is also no in-context improvement. We hypothesize the MMC4 samples do not have a very strict image-text correspondence; the image only provides marginal information for text modeling (*i.e.*, most of the information is still from pure text modeling; an example is provided in Figure 4). It is also demonstrated by the loss curves in Figure 5, where training on the interleave corpus leads to a much lower loss, indicating the full

[†]We followed [71] to match the image and text segments by CLIP scores.

PT data	SFT data	VLM acc. (avg)		MMLU acc.
		0-shot	4-shot	
Llama-2	-	-	-	*46.0%*
MMC4	Visual	68.7%	70.9%	40.7% (-5.3%)
MMC4+COYO	Visual	69.0%	71.3%	40.2% (-5.8%)
Llama-2	*Text*	-	-	*51.2%*
MMC4	Vis.+Text	71.0%	72.1%	51.4% (+0.2%)
MMC4+COYO	Vis.+Text	**72.3%**	**73.6%**	50.9% (-0.3%)

Table 4. Joint SFT (Vis. + Text) not only bridges the degradation of text-only capability (MMLU acc.), but also improves the performance on visual-language tasks (both zero-shot and few-shot).

text segments provides more information. Therefore, the interleaved data structure is critical, allowing the model to pick up the image-related information, without over-forcing it to learn unrelated text modeling.

Data blending improves pre-training. Training on image-text pairs only led to a sharp degradation on text-only accuracy (more than 17%). Luckily, blending the interleaved corpus and image-text pairs allows us to introduce more diversity in the corpus, while also preventing the severe degradation. Training on MMC4+COYO further boosts the accuracy on visual language benchmarks (the gain is larger when we perform joint SFT, as we will show later (Table 4).

3.3. Recover LLM Degradation with Joint SFT

Despite the interleave data helps maintain the text-only capability, there is still a 5% accuracy drop. A potential approach is to maintain the text-only capability would be to add in text-only corpus (the one used in the LLM pre-training). However, such text corpus are usually proprietary even for open-source models; it is also unclear how to subsample the data to match the scale of vision-language corpus.

Luckily, we found the text-only capabilities are temporarily *hidden*, but not *forgotten*. Adding in text-only data during SFT can help bridge the degradation, despite using a much smaller scale compared to the text pre-training corpora (usually trillion scale).

Joint supervised fine-tuning. The common way for instruction tuning is to fine-tune the model on some visual language datasets (VQA/Caption style [18] or GPT-generated [39]). We found blending in text-only instruction data can simultaneously (i) recover the degradation in *text-only accuracy*, and (ii) improve the *visual language accuracy*. To this end, we also blended in 1M text-only instruction tuning data sampled from FLAN [17], which we termed as *joint SFT*. We provide the comparison in Table 4.

We can see that blending in the text-only SFT data not only bridges the degradation on text-only capability (the MMLU accuracy is on par compared to the original Llama-2 model fine-tuned on the same text-only instruction data), but also improves the visual language capability. We hypothesize that the text-only instruction data improves the model's instruction-following capability, which is also important for visual language tasks. Interestingly, the benefits of blending in COYO data is more significant with joint SFT. We believe that with joint SFT, the model no longer suffers from the text-only degradation when pre-trained with short captions, thus unlocking the full benefits from the better visual diversity.

4. Experiments

4.1. Scaling up VLM pre-training

We scale up the training of VLM in the following aspects to form our final model:

Higher image resolution. Above ablation studies used the OpenAI CLIP-L [48] with 224×224 resolutions as the visual encoder. We now use 336×336 image resolutions to include more visual details for the model, which can help tasks requiring fine-grained details (*e.g.*, TextVQA [54]).

Larger LLMs. By default, we used Llama-2 [60] 7B for ablation study. We also scaled to a larger LLM backbone (*e.g.*, Llama-2 [60] 13B) to further improve the performance.

Pre-training data. We used both interleaved image-text data and image-text pairs for pre-training (we sample roughly 1:1 image proportions) to improve the data diversity. The total the pre-training corpus contains about 50M images. It is smaller than the billion-scale pre-training data [6, 14, 62], but already demonstrates impressive improvements on downstream tasks.

SFT data. We also include a better SFT data blend from LLaVA-1.5 [38], which is more diverse (*e.g.*, contains reference-based annotations) and has high-quality prompt. The new SFT data blend can significantly improve the downstream evaluation metrics. We include details the Appendix.

Limitations. Due to the limited compute budget, we have not been able to further scale up the size of the pre-training corpus to billion-scale, which we leave as future work. Nonethess, pre-training on 50M images already demonstrated significant performance improvement.

4.2. Quantitative Evaluation

visual language tasks. We perform a comprehensive comparison with state-of-the-art models on 12 visual language benchmarks in Table 5. Compared to existing models (*e.g.*, LLaVA-1.5 [38]), our model achieves consistent improvements over most datasets at different model sizes under a head-to-head setting (using the same prompts and base LLM;

Method	LLM	Res.	PT	IT	VQA^{v2}	GQA	VisWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{CN}	SEED	LLaVA^{W}	MM-Vet
BLIP-2 [35]	Vicuna-13B	224	129M	-	41.0	41	19.6	61	42.5	85.3	1293.8	–	–	46.4	38.1	22.4
InstructBLIP [18]	Vicuna-7B	224	129M	1.2M	–	49.2	34.5	60.5	50.1	–	–	36	23.7	53.4	60.9	26.2
InstructBLIP [18]	Vicuna-13B	224	129M	1.2M	–	49.5	33.4	63.1	50.7	78.9	1212.8	–	–	–	58.2	25.6
Shikra [12]	Vicuna-13B	224	600K	5.5M	77.4*	–	–	–	–	–	–	58.8	–	–	–	–
IDEFICS-9B [30]	LLaMA-7B	224	353M	1M	50.9	38.4	35.5	–	25.9	–	–	48.2	25.2	–	–	–
IDEFICS-80B [30]	LLaMA-65B	224	353M	1M	60.0	45.2	36.0	–	30.9	–	–	54.5	38.1	–	–	–
Qwen-VL [9]	Qwen-7B	448	1.4B	50M	78.8*	59.3*	35.2	67.1	63.8	–	–	38.2	7.4	56.3	–	–
Qwen-VL-Chat [9]	Qwen-7B	448	1.4B	50M	78.2*	57.5*	38.9	68.2	61.5	–	1487.5	60.6	56.7	58.2	–	–
LLaVA-1.5 [38]	Vicuna-1.5-7B	336	0.6M	0.7M	78.5*	62.0*	50.0	66.8	58.2	**85.9**	1510.7	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5 [38]	Vicuna-1.5-13B	336	0.6M	0.7M	80.0*	**63.3***	53.6	71.6	61.3	**85.9**	1531.3	67.7	63.6	61.6	70.7	35.4
VILA-7B (ours)	Llama-2-7B	336	50M	1M	79.9*	62.3*	57.8	68.2	64.4	85.5	1533.0	68.9	61.7	61.1	69.7	34.9
VILA-13B (ours)	Llama-2-13B	336	50M	1M	**80.8***	**63.3***	**60.6**	**73.7**	**66.6**	84.2	**1570.1**	**70.3**	**64.3**	**62.8**	**73.0**	**38.8**
+ShareGPT4V	Llama-2-13B	336	50M	1M	80.6*	63.2*	62.4	73.1	65.3	84.8	1556.5	70.8	65.4	61.4	78.4	45.7

Table 5. Comparison with state-of-the-art methods on 12 visual-language benchmarks. Our models consistently outperform LLaVA-1.5 under a head-to-head comparison, using the same prompts and the same base LLM (Vicuna-1.5 is based on Llama-2), showing the effectiveness of visual-language pre-training. We mark the best performance **bold** and the second-best underlined. Benchmark names are abbreviated due to space limits. VQA-v2 [25]; GQA [29]; VisWiz [26]; SQA^{I}: ScienceQA-IMG [41]; VQA^{T}: TextVQA [54]; POPE [36]; MME [24]; MMB: MMBench [40]; MMB^{CN}: MMBench-Chinese [40]; SEED: SEED-Bench [33]; LLaVA^{W}: LLaVA-Bench (In-the-Wild) [39]; MM-Vet [67]. *The training images of the datasets are observed during training. We also tried adding the ShareGPT4V [13] to the SFT blend on top of VILA-13B (last row), leading to a significant improvement on LLaVA-Bench and MM-Vet (marked in green).

Size	Model	MMLU [27]	BBH [57]	DROP [22]
7B	Llama-2	46.0%	32.0%	31.7%
	Llama-2+SFT	**51.8%**	**39.3%**	**53.1%**
	Vicuna-1.5	49.8%	36.9%	29.2%
	VILA	50.8%	38.5%	52.7%
13B	Llama-2	55.7%	37.5%	41.6%
	Llama-2+SFT	54.3%	43.2%	59.2%
	Vicuna-1.5	55.8%	38.4%	43.6%
	VILA	**56.0%**	**44.2%**	**63.6%**

Table 6. VILA maintains competitive accuracy on text-only benchmarks. There is a small gap compared to the text-only model under 7B; but the accuracy is even better under 13B.

Vicuna-1.5 is based on Llama-2). Remarkably, we 7B model is able to outperform LLaVA-1.5 13B on VisWiz [26] and TextVQA [54] by a large margin thanks to the pre-training. Our 7B model even outperforms the 13B LLaVA model on these datasets. Our model also has multi-lingual capability despite the vision-language instruction data is in English, outperforming LLaVA-1.5 on MMBench-Chinese benchmark. Our results demonstrates the benefits of vision-language pre-training on downstream tasks, even when using a high-quality instruction tuning dataset [38].

Text-only performance. Our goal is to augment an LLM to support visual inputs. It is essential that the model can retain the text-only capability. Therefore, we further evaluate the text-only performance of the models under three benchmarks: MMLU [27], BBH [57], and DROP [22] in Table 6. We did not choose benchmarks like MT-Bench [69] since text instrution tuning is not the focus of the work. Overall, our model achieves performance as Llama-2 fine-tuned with the same text SFT data: the accuracy of our 7B model is a bit lower, while the 13B is higher. We suspect the smaller model may suffer from a larger text performance degradation during the pre-training, as observed in [20].

4.3. Qualitative Evaluation

Here we study how visual language pre-training enables new capabilities for the model. Part of the image samples are taken from [6, 14, 64].

Multi-image reasoning. Thanks to the pre-training, our model has the ability to reason over multiple images, despite the SFT data is composed of single-image samples. We provide two examples in Figure 6. In the first example, our model is able to figure out the common object (*i.e.*, a flamingo) across the three images and the different art styles of each one, while the LLaVA model failed. The LLaVA model hallucinates and cannot distinguish the information from different input images. In the second example, our model is able to find one of the differences (*i.e.*, the headwear) out of the two.

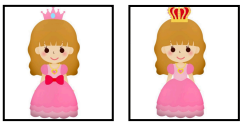
In-context learning. In-context learning is an important characteristic of LLMs, allowing people to prompt the LLM with few-shot samples to enable new tasks. We provide in-context learning samples in Figure 7. Interestingly LLaVA-1.5 [38] can also perform in-context learning to some extend, despite only being trained on single-image-text-paired samples. We believe the capability is inherited from text-only

Prompt: Image 1: <img1>\nImage 2: <img2>\nImage 3: <img3>\n
What is the common thing about these three images?

LLaVA-1.5: The common theme among these three images is the presence of a pink flamingo. In each image, the flamingo is standing in the water, with one image showing the bird in the water, another image showing the bird's legs, and the third image showing the bird's head. The flamingo is the central subject in all three images, and its distinctive pink color and unique appearance make it the focal point of each photo.

Ours: The common theme among these three images is the presence of a flamingo. In the first image, the flamingo is a 3D model, while in the second image, it is a pink cartoon. In the third image, it is a real flamingo standing in the water. All three images showcase the bird in different styles and formats, but they all share the same subject.

Prompt: Image 1: <img1>\nImage 2: <img2>\n
Tell me about the differences between image 1 and image 2.



LLaVA-1.5: In image 1, there is a small girl wearing a pink dress and a crown. She is smiling and appears to be happy. In image 2, the girl is wearing a pink dress and a crown, but she is not smiling. The main difference between the two images is the girl's facial expression, with one showing a smile and the other not showing any expression.

Ours: The two images feature a cartoon girl wearing a pink dress. In the first image, the girl is wearing a crown, while in the second image, she is wearing a tiara. Both images showcase the girl in a princess-like attire, but the difference lies in the type of headwear she is wearing. The crown is a more traditional and regal accessory, while the tiara is a more playful and common accessory associated with princesses in popular culture.

Figure 6. Our model VILA can reason over multiple images thanks to the pre-training process. The samples are taken from [6, 64].

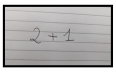
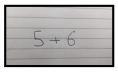
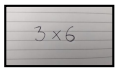
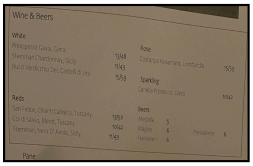
Input few shot examples + target image			Output
 Underground.	 Congress.	 ?	**LLaVA-1.5**: Soulemes. **Ours**: Soulomes.
 2+1=3	 5+6=11	 ?	**LLaVA-1.5**: 3x6=18 **Ours**: 3x6=18
 Romanticism	 Surrealism	 ?	**LLaVA-1.5**: Surrealism **Ours**: Impressionism

Figure 7. VILA has better in-context learning capability thanks to interleaved image text pretraining rather than single image-text pairs.We feed two image+text pairs and a third image as the context to prompt the VLM. LLaVA failed the first sample due to limited OCR capability, and failed the third examples by repeating the second sample semantic.

pre-training of the base LLM. Nonetheless, our model outperforms LLaVA-1.5 for in-context learning: LLaVA-1.5 failed the first sample due to limited OCR capability, and failed the third example by repeating the semantics.

Visual Chain-of-Thoughts (CoT). Our model is able to perform chain-of-thought reasoning given visual inputs. As shown in Figure 8, VILA is able to perform complex CoT reasoning over the input images (multi-image or single-image) when adding "Think step-by-step" to the end of the prompt. We believe the CoT capability is inherited from text-only SFT, despite there is no such samples from the visual language instruction data.

Example: Visual Chain-of-Though (CoT)

Prompt: Photo: <img1>\nMenu: <img2>\nHow much should I pay for all the beer on the table according to the price on the menu?
Answer: According to the menu, the price for the beer on the table is $6.

Prompt: … + ***Think step-by-step***.
Answer: The beer on the table is Magna, which is priced at $6.00. There are two bottles of Magna on the table. Therefore, the total cost of the beer should be $6.00 x 2 = $12.00.

Prompt: <img>\nI am getting just two custom pizza for me and my friend. How much should I pay in total according to the image?
Answer: According to the image, you should pay $11.99 for two custom pizzas.

Prompt: … + ***Let's think step-by-step***.
Answer: The image shows a chalkboard menu with a special for today, which is to create your own pizza. The price for this special is $9.99. Since you are getting two custom pizzas, you should pay $9.99 x 2 = $19.98 for the two pizzas. So, the answer is $19.98.

Figure 8. Our model is able to perform chain-of-thought reasoning given visual inputs. It is able to generate the correct answer when adding "Think step-by-step" to the prompt. *Zoom in* for a better view of the image details. Samples from [20, 64].

Better world knowledge. Since our model is pre-trained on a large-scale corpus, it has better understanding of world knowledge. We perform a case study by prompting the model to recognize the locations of some famous landmarks (please see supplementary due to space limits). VILA can correctly recognize 4 out of the 4 samples, while LLaVA-1.5

Resolution	Projector	#Tokens	Evaluation ↑		
			OKVQA	TextVQA	COCO
224	linear	256	49.9%	41.6%	116.0
336	linear	576	49.7%	49.8%	117.7
336	downsample	144	49.3%	45.6%	115.7

Table 7. Improving the image resolution from 224 to 336 can significantly improve TextVQA accuracy. The raw resolution matters more than #tokens; high-resolution with token downsampling works better than low-resolution. We report accuracy for OKVQA and TextVQA, and CIDEr for COCO. Note: the evaluation protocol is different from Table 5 and can only be compared within the table.

only gets 2 out of the 4, demonstrating the effectiveness of the pre-training. Samples are taken from [64].

4.4. Other Learnings.

Image resolution matters, not #tokens. We chose an image resolution of 336^2 since it provides more fine-grained details compared to 224^2, leading to improved accuracy on tasks like TextVQA [54]. As shown in Table 7, increasing the resolution from 224 to 336 can improve the TextVQA accuracy from 41.6% to 49.8%. However, a higher resolution leads to more tokens per image (336^2 corresponds to 576 tokens/image) and a higher computational cost. It also limits the number of demonstrations for in-context learning.

Luckily, we find that the raw resolution matters more than the #visual tokens/image. We can use different projector designs to compress the visual tokens. Here we try a "downsample" projector, which simply concatenates every 2×2 tokens into a single one and use a linear layer to fuse the information. It reduces the #tokens to 144 under the 336 resolution, that is even smaller than the 224+linear setup. Nonetheless, the TextVQA accuracy is higher ($\sim 46\%$ *vs*. 41.6%), despite still 3% worse compared to 336+linear setup, showing a large redundancy in the image tokens. The gap on other datasets such as OKVQA and COCO is smaller since they usually require higher-level semantics.

In our main results, we did not use any token compression methods to provide the best accuracy despite this encouraging observation, and leave it to future work.

Comparison to frozen LLMs with visual experts. Another interesting method for retaining the text capabilities of LLMs during the pre-training is to freeze the base LLM and add an extra visual expert to process the visual tokens [62]. The definition of expert is similar to MoE frameworks, but with a manual routing mechnism according to token types. Since the base LLM is frozen, the model fully retains the original functionality for text-only inputs during pre-training. However, we find that directly fine-tuning the LLM during visual language pre-training still leads to a better VLM accuracy and in-context learning capability (Table 8). Adding an extra visual expert also leads to near $2\times$ model size increase, which is not friendly for edge deployment. Therefore, we chose to directly fine-tune the base LLM.

	#Param	VLM acc. (avg)	
		0-shot	4-shot
Visual Expert [62]	$1.9\times$	67.0%	64.8%
Fine-tune	$\mathbf{1}\times$	**71.0%**	**72.1%**

Table 8. Directly fine-tuning the LLM during pre-training leads to better VLM accuracy and in-context learning capabilities. It also enjoys a smaller model size. Both settings are pre-trained on the MMC4-core dataset [71].

5. Related Work

Large language models (LLMs). LLMs based on Transformers [61] have fundamentally changed the language processing field. They are achieving increasing capabilities by *scaling up* the model size and the pre-training corpus [1, 10, 16, 19, 21, 23, 28, 49, 55]. It is believed that most the capability of the LLM is obtained from the large-scale *pre-training* process, which are later unlocked through instruction tuning [17, 45, 46]. There is a growing effort from the open-source community to build a strong base LLM [59, 60, 68] and the conversational variants [15, 58]. In this work, we start with the base Llama-2 model [60].

Visual language models (VLMs). VLMs are LLMs augmented with visual inputs to provide a unified interface for visual language tasks. There are two main designs for VLMs: 1. cross-attention based, where the LLM is frozen while the visual information is fused into intermediate embeddings with a cross-attention mechanism [6, 7]; 2. auto-regressive based, where the visual input is tokenized and fed to the LLM alongside text tokens [2, 5, 8, 14, 20, 35, 39, 65, 70]. The latter is a natural extension by treating visual inputs as a foreign language. VLMs are also instruction-tuned so that they can better follow human instructions or perform conversations [18, 39, 56]. In this work, we study the pre-training process of the auto-regressive VLMs due to their flexibility when handling multi-modal inputs.

Following text-only LLMs, people also study different training recipes for VLMs. Some work freezes the LLM and train auxiliary components [6, 34, 35, 62], others fine-tune the LLM to enable visual capabilities [14, 20]. There is also usage of different data corpora, including image-text pairs [14, 20, 34, 39], interleaved datasets [7], video-text pairs [42], visual-grounded annotations [38, 47], *etc*. In this work, we provide a holistic ablation of different design choices for the pre-training stage.

6. Conclusion

This paper has explored effective pretraining design options to augment LLMs towards vision tasks. Leveraging full strength of LLM learning, interleaved-nature of image-text data, and careful text data re-blending, VILA has surpassed state-of-the-art methods for vision tasks while preserving text-only capabilities. VILA has also depicted strong reasoning capability for multi-image analysis, in-context learning and zero/few-shot tasks. We hope our paper can help spur further research on VLM pretraining and collection of cross-modality datasets.

Acknowledgements

We would like to thank Bryan Catanzaro for fruitful discussions. We also appreciate the help from Zhuolin Yang, Guilin Liu, Lukas Voegtle, Philipp Fischer, Karan Sapra and Timo Roman on dataset preparation and feedback.

References

[1] GPT-4 technical report. Technical report, OpenAI, 2023. https://arxiv.org/abs/2303.08774. 1, 8

[2] Fuyu-8B: A multimodal architecture for AI agents. https://www.adept.ai/blog/fuyu-8b, 2023. 1, 2, 8

[3] Gemini: A family of highly capable multimodal models. Technical report, Gemini Team, Google, 2023. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf. 14

[4] Yi-34B large language model. https://huggingface.co/01-ai/Yi-34B, 2023. 1

[5] Emanuele Aiello, Lili Yu, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. Jointly training large autoregressive multimodal models. *arXiv preprint arXiv:2309.15564*, 2023. 8

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 2, 4, 5, 6, 7, 8, 13

[7] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, 2023. 2, 8

[8] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. Technical report, Alibaba Group, 2023. https://arxiv.org/abs/2303.08774. 1, 8

[9] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 6

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 1, 8

[11] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 1, 2, 3, 4, 13

[12] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 6

[13] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. 6

[14] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 1, 5, 6, 8, 13

[15] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1, 8

[16] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1, 8

[17] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 5, 8

[18] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. 1, 2, 3, 5, 6, 8, 13

[19] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. 1, 8

[20] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1, 2, 3, 4, 6, 7, 8

[21] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi

Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022. 8

[22] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*, 2019. 6

[23] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022. 8

[24] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 6

[25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 6

[26] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 6

[27] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300, 2020. 4, 6

[28] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 8

[29] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 6

[30] IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics, 2023. 6

[31] Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, et al. Mistral–a journey towards reproducible language model training, 2021. 1

[32] Matej Kosec, Sheng Fu, and Mario Michael Krell. Packing: Towards 2x nlp bert acceleration. 2021. 13

[33] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 6

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 8

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 3, 4, 6, 8, 14

[36] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 13

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 2, 5, 6, 8, 13

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023. 1, 2, 3, 4, 5, 6, 8

[40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 6

[41] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 6

[42] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 8

[43] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 13

[44] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2

[45] OpenAI. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt, 2023. Accessed: 2023. 1, 8

[46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 8

[47] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 8

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[49] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 8

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1, 4

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 14

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 14

[53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1, 2, 4, 14

[54] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2, 5, 6, 8

[55] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022. 8

[56] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 1, 8

[57] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022. 6

[58] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 1, 8

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 8

[60] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 5, 8

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 8

[62] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 4, 5, 8

[63] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 2

[64] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 6, 7, 8, 13

[65] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 8

[66] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2

[67] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6

[68] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. 8

[69] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. 6

[70] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 8

[71] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig

Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023. 1, 2, 3, 4, 8, 13

A. SFT Blend for Ablation Study

We used an in-house data blend for supervised fine-tuning/instruction tuning during the ablation study. We followed [18] to build the FLAN-style instructions from the training set of 18 visual language datasets, as shown in Table 9. We may see that most of the datasets are in a VQA format. For the final model, we also blend in the LLaVA-1.5 SFT dataset [38], which has better quality and diversity (for example, it contains visual reference data like RefCOCO [37, 43]).

Categories	Datasets
Captioning	Image Paragraph Captioning, MSR-VTT, TextCaps
Reasoning	CLEVR, NLVR, VisualMRC
Translation	Multi30k
VQA	ActivityNet-QA, DocVQA, GQA, iVQA, MSRVTT-QA, MSVD-QA, OCR-VQA, ST-VQA, ViQuAE, VQAv2, Visual Dialog

Table 9. The SFT blend we used during the ablation study.

B. Training Cost

We perform training on 16 A100 GPU nodes, each node has 8 GPUs. The training hours for each stage of the 7B model are: projector initialization: 4 hours; visual language pre-training: 30 hours; visual instruction-tuning: 6 hours. The training corresponds to a total of 5.1k GPU hours. Most of the computation is spent on the pre-training stage.

We have not performed training throughput optimizations like sample packing [32] or sample length clustering. We believe we can reduce at least 30% of the training time with proper optimization. We also notice that the training time is much longer as we used a high image resolution of 336×336 (corresponding to 576 tokens/image). We should be able to reduce the training time by more than 50% by using lower-resolution images for pre-training (*e.g.*, 224×224) and scale up the resolution at the later stage of the training [14], which we leave to future work.

C. Details on COYO Subsampling

We were able to download 25M out of 30M images for the MMC4-core dataset [71]. The COYO-700M dataset [11] contains about 700M images. To maintain a similar dataset size, we subsample 25M images from the COYO-700M dataset. Specifically, we sort all the samples based on the CLIP similarity between images and captions and keep the 25M images with the highest similarities. Samples with a high CLIP similarity usually have better image-caption correspondence.

D. More Qualitative Samples

Here we provide more qualitative samples that we were not able to include in the main paper due to space limits. Many of the image samples are taken from [6, 64].

Prompt: Which city is this landmark in?

<img1>
LLaVA-1.5: The landmark is in Tokyo, Japan.
Ours: The landmark in the image is located in Taipei, Taiwan.

<img2>
LLaVA-1.5: This landmark is in New York City.
Ours: This landmark is located in New York City.

<img3>
LLaVA-1.5: This landmark is in New York City.
Ours: The landmark in the image is located in New York City.

<img4>
LLaVA-1.5: This landmark is located in Tokyo, Japan.
Ours: This landmark is located in Kyoto, Japan.

Figure 9. Landmark city recognition. Visual-language pre-training gives the model better world knowledge. It reduces the bias towards answering "Tokyo" compared to LLaVA-1.5 [38]. We mark the wrong responses in red. Samples are taken from [64].

Better world knowledge. Pre-training on a large-scale corpus allows the model to have better visual-related world knowledge. Here we take four landmark images from [64] (without curation) and ask the model which city is the landmark located in (Figure 9). VILA can correctly recognize 4 out of the 4 samples, while LLaVA-1.5 only gets 2 out of the 4, with an output bias on more common cities like Tokyo and New York.

Visual reference understanding. Our model can understand visual reference overlaid on images and perform reasoning. We provide a sample of visual reference reasoning in Figure 10 (from [64]). VILA is able to correctly figure out what is in the circled glass, while LLaVA-1.5 failed.

More in-context learning samples. We provide more in-context learning samples in Figure 13, including company knowledge, object counting, and French poems. VILA demonstrates strong in-context learning capabilities under various demonstrations.

More VQA samples. We provide more VQA samples in Figure 14. VILA is able to understand memes, reason on

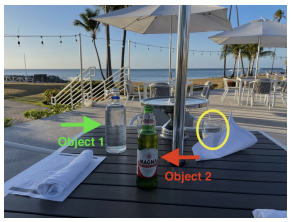
Visual input example: Understanding visual reference	
	
User	What is in the circled glass? Object 1 or 2? Describe what is Object 1 and object 2. Then check what is in the circled glass.
GPT-4V	Object 1 is a glass bottle of water. It is clear and has a blue label on it. Object 2 is a glass bottle of Magna beer. It is green and has a red and white label on it. In the circled glass, it appears to be a clear liquid, likely water. So, it is likely that the contents of the circled glass are from Object 1, the glass bottle of water.
LLaVA-1.5	Object 1 is a bottle of water, and Object 2 is a bottle of beer. The circled glass contains the beer.
Ours	Object 1 is a bottle of water, and Object 2 is a bottle of beer. The circled glass contains Object 1, which is a bottle of water.

Figure 10. Our model can understand visual reference overlaid on images and perform reasoning.

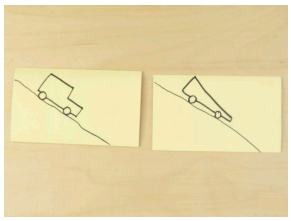
Visual input example: Visual logic reasoning	
	
User	Which of these cars is more aerodynamic? The one on the left or the right? Explain why, using specific visual details.
Gemini	The car on the right is more aerodynamic. It has a lower profile and a more streamlined shape. The car on the left has a higher profile and a more boxy shape, which makes it less aerodynamic.
LLaVA-1.5	The car on the right is more aerodynamic than the one on the left. This is evident from the fact that the car on the right is shown going down a hill, while the car on the left is shown going up a hill. When a car is going downhill, it experiences less air resistance, which allows it to maintain higher speeds with less energy expenditure. In contrast, when a car is going uphill, it encounters more air resistance, making it more challenging to maintain speed and requiring more energy to overcome the resistance. Therefore, the car on the right is more aerodynamic due to the downhill slope, which allows it to travel more efficiently.
Ours	The car on the right is more aerodynamic than the one on the left. This is because the car on the right has a more streamlined shape, which reduces air resistance and drag. The car on the left has a more traditional, boxy shape, which increases air resistance and drag. The aerodynamic car on the right would be more efficient and faster, while the car on the left would be slower and less efficient.

Figure 11. Our model can understand visual details on images and perform logical reasoning.

multiple images or video frames, and provide help on corner cases in autonomous driving.

More logical reasoning samples. We check VILA on the most recent samples from Gemini's release [3] in Figure 11. VILA is able to understand the logic utilizing visual features in detail, whereas the LLAVA-1.5 cannot yield reasonable responses.

Using VILA for detailed captioning. People have been using datasets like LAION [53] to train text-to-image generative models [51, 52]. The quality of the image-text pairs can significantly affect the performance of the trained model. Some captions in the training datasets are quite noisy: they are either not quite related to the images or are too abbreviated and contain limited details. We show that we can use VLM models to generate high-quality and detailed captioning (Figure 12). We use a simple prompt "Describe the image in detail." to generate the captions. VILA can generate more related descriptions compared to the original caption (sample 1) and provide more details compared to previous models like BLIP-2 [35] (sample 2).

Example: Detailed Captioning

Original caption (LAION): Best Location in Sydney!!!
Ours: The image captures a beautiful nighttime scene of a city skyline with a large body of water in front of it. The city is illuminated by the lights of various buildings, creating a vibrant atmosphere. The water is filled with numerous boats of different sizes.

BLIP-2: There is a mountain of meat and a river in the background
Ours: The image is a surreal and colorful scene featuring a landscape made entirely of meat. The landscape is filled with various types of meat, including bacon, ham, and sausage, creating a unique and eye-catching scene. The meat is arranged in the form of mountains, hills, and valleys, giving the impression of a fantasy world. In addition to the meat, there are several birds scattered throughout the scene.

Figure 12. VILA can provide detailed captions. The raw captions in datasets like LAION [53] can be noisy and irrelevant. VILA can generate meaningful captions with more details compared to BLIP-2 [35]. The results are obtained by prompting the model with "Describe the image in detail.".

Input few shot examples + target image						Output
	The company is famous for its search engine.		The company is famous for iPhone and Mac.			The company is famous for its graphics processing units.
	pandas: 3		dogs: 2			giraffes: 4
	Les sanglots longs des violons de l'automne blessent mon coeur d'une langueur monotone.		Pour qui sont ces serpents qui sifflent sur vos têtes?			Les flamands roses s'embrassent avec passion, leurs cœurs se touchant, leur amour se partageant.

Figure 13. In-context learning samples on company knowledge, object counting, and French poem. The predictions are from VILA-13B.

Example: VQA Samples

Q: Can you explain the meme?
A: The meme depicts a man's reaction to the price of a computer graphics card. In the first image, the man is smiling and appears excited about the product. In the second image, he is shocked and disappointed by the high price of the graphics card, which is $1,200. The meme is a playful representation of the contrast between the man's initial enthusiasm and his subsequent disappointment upon learning the cost of the product.

Q: What happens to the man after hitting the ball?
A: After hitting the ball, the man falls down onto the grass.
Q: Why does the man fall to the grass?
A: The man falls to the grass because he is swinging his golf club too hard, causing him to lose his balance and fall.

Q: Photo 1, at 10:30 am: <img1>
Photo 2, at 12:45 pm: <img2>
Photo 3, at 3:45 pm: <img3>
What did I have for lunch, and what time was it?
A: I had a sandwich for lunch, and it was at 12:45 pm.

Q: What is unusual about this image?
A: The unusual aspect of this image is that a chair is flying through the air on a highway, seemingly coming out of the back of a truck.
Q: What should you do if you encounter this?
A: If you encounter this situation, you should immediately stop your vehicle and move to a safe distance from the truck and the flying chair. It is essential to avoid any potential hazards and contact the authorities to report the incident and ensure the safety of everyone involved.

Figure 14. VQA samples. VILA is able to understand memes, reason on multiple images or video frames, and provide help on corner cases in autonomous driving. The answers are from VILA-13B.