Supplementary Material
VILA: On Pre-training for Visual Language Models

Ji Lin[1,2]*† Hongxu Yin[1]* Wei Ping[1] Pavlo Molchanov[1]
Mohammad Shoeybi[1] Song Han[1,2]
[1]NVIDIA [2]MIT
ViLA https://github.com/Efficient-Large-Model/VILA

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 [5] to build the FLAN-style instructions from the training set of 18 visual language datasets, as shown in Table 1. 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 [9], which has better quality and diversity (for example, it contains visual reference data like RefCOCO [8, 10]).

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 1. 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 [6] 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 [4], which we leave to future work.

C. Varying LLMs

For extra insights we here present the results with Vicuna-1.5-7B as the new LLM backbone to verify if our pretraining conclusions scale across LLMs. Firstly, we reproduce the training recipe study (originally in Table **??**) and report the average 0/4-shot accuracy in Table 2 (left). We observed the same conclusion that updating the LLM is important in the pre-training stage. Secondly, we also provide the results in the setting of Table **??** (as right of Table 2). The two backbones achieved similar accuracy on the benchmarks. Overall, our conclusions are general across LLM backbones.

PreT	SFT	0-shot	4-shot
✗	✓	62.6	59.2
✓	✓	**69.1**	**72.8**

	VQA^{v2}	GQA	VisWiz
Llama-2 7B	79.9	62.3	57.8
Vicuna 7B	79.3	62.3	58.7

Table 2. Ablation and final performance with Vicuna-1.5-7B.

D. Details on COYO Subsampling

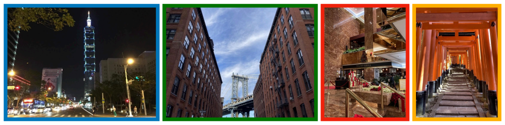
We were able to download 25M out of 30M images for the MMC4-core dataset [15]. The COYO-700M dataset [3] contains about 700M images. To maintain a similar dataset size, we subsample 25M images from the COYO-700M dataset. Specifically, we sort all the samples based on the CLIP similarity between images and captions and keep the 25M images with the highest similarities. Samples with a high CLIP similarity usually have better image-caption correspondence.

E. 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 [2, 14].

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 [14]

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 1. 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 [9]. We mark the wrong responses in red. Samples are taken from [14].

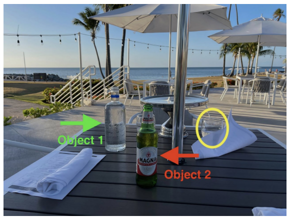
(without curation) and ask the model which city is the landmark located in (Figure 1). VILA can correctly recognize 4 out of the 4 samples, while LLaVA-1.5 only gets 2 out of the 4, with an output bias on more common cities like Tokyo and New York.

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 2 (from [14]). VILA is able to correctly figure out what is in the circled glass, while LLaVA-1.5 failed.

More logical reasoning samples. We check VILA on the most recent samples from Gemini's release [1] in Figure 3. 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 [13] to train text-to-image generative models [11, 12]. 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 4). 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 [7] (sample 2).

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

More in-context learning samples. We provide more in-context learning samples in Figure 5, 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 6. VILA is able to understand memes, reason on multiple images or video frames, and provide help on corner cases in autonomous driving.

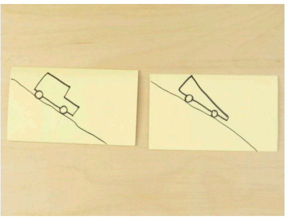
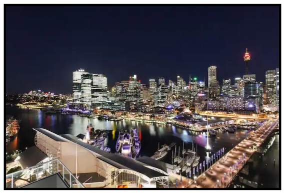
Visual input example: Visual logic reasoning	
	
User	Which of these cars is more aerodynamic? The one on the left or the right? Explain why, using specific visual details.
Gemini	The car on the right is more aerodynamic. It has a lower profile and a more streamlined shape. The car on the left has a higher 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.

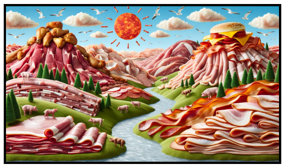
Figure 3. Our model can understand visual details on images and perform logical reasoning.

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 4. VILA can provide detailed captions. The raw captions in datasets like LAION [13] can be noisy and irrelevant. VILA can generate meaningful captions with more details compared to BLIP-2 [7]. The results are obtained by prompting the model with "Describe the image in detail.".

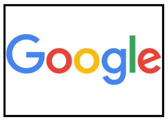
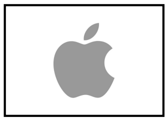
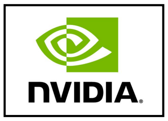
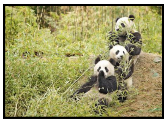
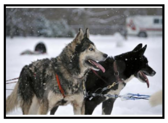
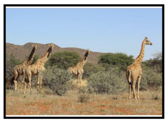
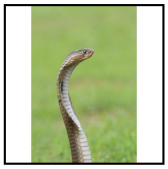
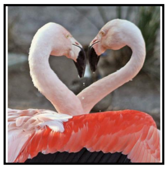
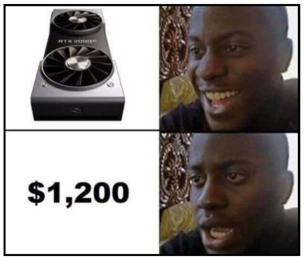
Input few shot examples + target image					**Output**
	The company is famous for its search engine.		The company is famous for iPhone and Mac.		The company is famous for its graphics processing units.
	pandas: 3		dogs: 2		giraffes: 4
	Les sanglots longs des violons de l'automne blessent mon coeur d'une 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.

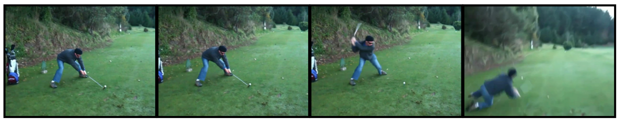
Figure 5. In-context learning samples on company knowledge, object counting, and French poem. The predictions are from VILA-13B.

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.

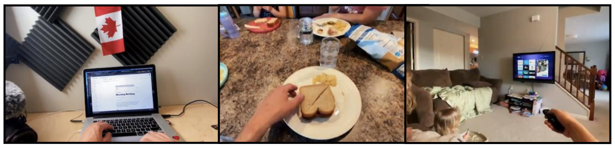


Q: What happens to the man after hitting the ball?

A: After hitting the ball, the man falls down onto the grass.

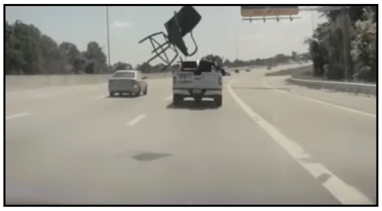
Q: Why does the man fall to the grass?

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 6. 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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.