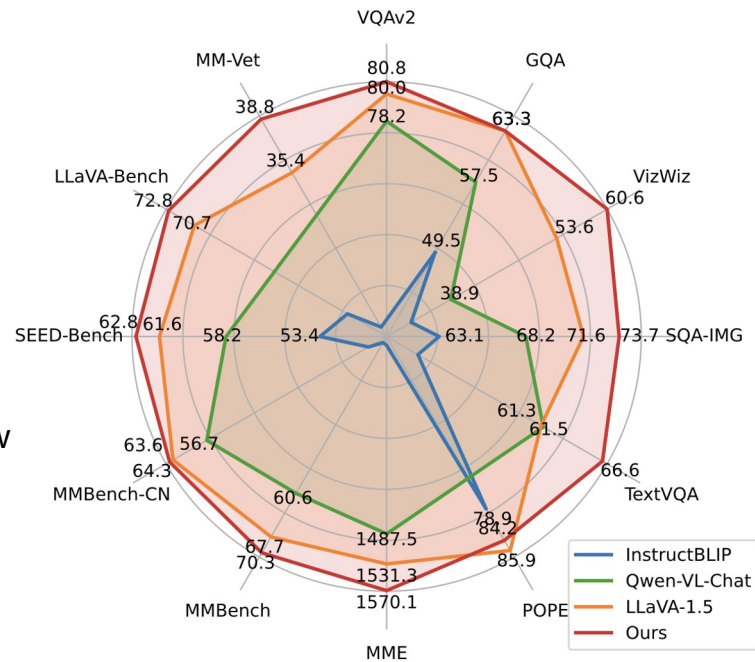


# VILA: On Pre-training for Visual Language Models

Ji Lin<sup>1,2\*</sup> Hongxu Yin<sup>1\*</sup> Wei Ping<sup>1</sup> Yao Lu<sup>1</sup> Pavlo Molchanov<sup>1</sup> Andrew Tao<sup>1</sup> Huizi Mao<sup>1</sup> Jan Kautz<sup>1</sup> Mohammad Shoeybi<sup>1</sup> Song Han<sup>1,2</sup>

<sup>1</sup>NVIDIA <sup>2</sup>MIT



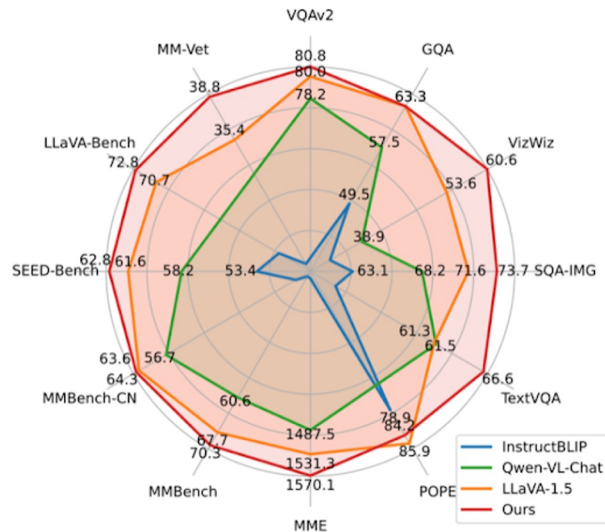
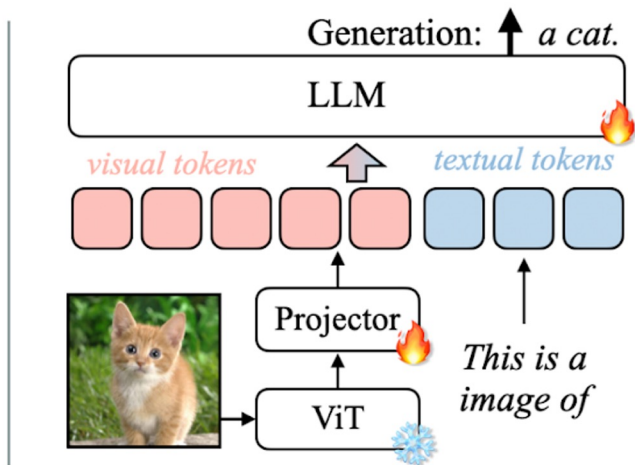
# VILA: Visual Language Model

- VILA augments the language model with image tokens (augment LLM with visual token)
- VILA unveils appealing capabilities: multi-image reasoning, in-context learning, visual chain-of-thought, and better world knowledge.
- VILA outperforms state-of-the-art models LLaVA-1.5, Fuyu-8B.
- We quantized VILA using 4bit AWQ which runs at 6ms/token for VILA-7B on a single 4090 GPU.

Step 0: projector training

Step 1: interleaved training

Step 2: text-vision joint SFT



# VILA – Our Findings

- **On LLM.** Freezing LLMs during pre-training can achieve decent zero-shot performance, but lack in-context learning capability, which requires unfreezing the LLM. Updating the LLM achieves better generalization for visual-language tasks (compared to prompt tuning).
- **On data structure.** Interleaved pre-training data is beneficial whereas image-text pairs alone are not optimal; Aligning visual and textual tokens improves in-context learning.
- **On data blending.** Joint instruction fine-tuning (adding text-only) is essential. 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.

# VILA – Our Findings

- **On LLM.** Freezing LLMs during pre-training can achieve decent zero-shot performance, but lack in-context learning capability, which requires unfreezing the LLM. Updating the LLM achieves better generalization for visual-language tasks (compared to prompt tuning).

	PreT	SFT	Projector	OKVQA		TextVQA		COCO		Flickr		Average	
	Train LLM?			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.

# VILA – Our Findings

- **On data structure.** Interleaved pre-training data is beneficial whereas image-text pairs alone are not optimal; Aligning visual and textual tokens improves in-context learning.

Pre-train Data	VLM acc (avg)		MMLU acc.
	0-shot	4-shot	
<i>Llama-2</i>	-	-	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	<b>69.0%</b>	<b>71.3%</b>	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.

# VILA – Our Findings

- **On data blending.** Joint instruction fine-tuning (adding text-only) is essential. 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.

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

# VILA: Visual Language Model

In-context learning (counting)



pandas: 3



dogs: 2



**Pred:** cats: 3

---

*context*

---

*prediction*

# VILA: Visual Language Model

In-context learning (world knowledge)



The company is famous for its search engine.



The company is famous for the operating system.



The company is famous for iPhone and Mac.



**Pred:** The company is famous for its graphics processing units (GPUs)

---

*context*

---

*prediction*



# VILA: Visual Language Model

In-context learning (world knowledge)



Home to the best burgers  
and fried chicken.



Home to unbeatable  
fish and chips.



Home to  
outstanding ramen.



**Pred:** Home to the  
greatest pizza.

---

*context*

---

*prediction*

# VILA: Visual Language Model

## Visual Chain-of-Thought (CoT)



Wine & Beers			
White			
Principessa Gavia, Clera		Rose	
Stemmari Chardonnay, Sicily	13/48	Costaripa Rosamara, Lombardia	15/59
Bucci Verdicchio Dei, Castelli di Jesi	11/43		
	15/59	Sparkling	
		Canella Prosecco, Clera	10/42
Reds			
San Felice, Chianti Calssico, Tuscany	13/50	Beers	
Col di Sasso, Blend, Tuscany	10/42	Medalla	5
Stemmari, Nero D' Avola, Sicily	11/43	Magna	6
		Heineken	6
		Presidente	6
Pane			

**Question:** Photo: [<image1>](#) Menu: [<image2>](#)  
How much should I pay for all the beer on the table according to the price on the menu? Think step-by-step.

-----

**Pred:** 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  $\$6.00 \times 2 = \mathbf{\$12.00}$ .

# VILA: Visual Language Model

## On Synthetic Images

- High-quality data generation for diffusion models
  - Image-text pairs are noisy (LAION sample)
  - Better textual descriptions => better correspondence/controllability



**Q:** Describe the image in details.

**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.

# VILA: Visual Language Model

## Explaining NVIDIA meme



**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.

# VILA for Autonomous Vehicle

## Zero-shot learning: corner case handling in self-driving



**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.

# VILA for Construction

## Zero-shot learning: crack detection



**User:** Is the building in health condition?

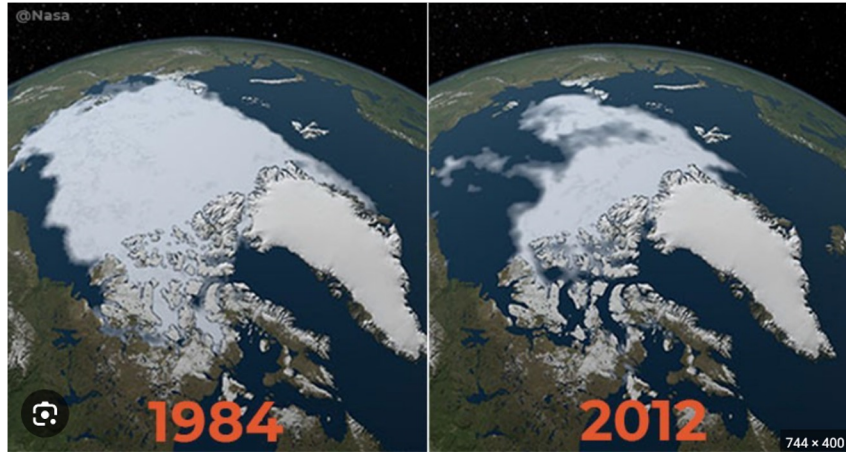
**VILA:** The building appears to be in a state of disrepair, with visible cracks and chips in the paint. This suggests that the building may be old, poorly maintained, or has experienced some form of damage or wear over time. The presence of these structural issues could potentially pose safety risks or negatively impact the building's overall appearance and value.

**User:** What are the likely service needed for this building?

**VILA:** The building in the image has a cracked wall and a damaged roof, which indicates that it requires repair and maintenance services. The cracked wall may require repairs to ensure structural integrity and prevent further damage. The damaged roof may need to be replaced or repaired to prevent water leakage and potential damage to the building's interior. Additionally, the building may need to be inspected for any other potential issues that may arise due to the damage.

# VILA for Science

## Zero-shot learning: satellite image comparison



**User:** What is the implication of temperature based on this image?

**VILA:** The image shows a comparison of the Arctic ice cover in 1984 and 2012. The implication of this image is that the Arctic ice cover has significantly decreased over the years, which is a result of climate change. The reduction in ice cover can have various implications, such as increased exposure of the Arctic environment to warmer temperatures, which can lead to the melting of sea ice, loss of habitat for polar bears and other Arctic species, and potential disruptions to the global climate system. This image serves as a visual representation of the impact of climate change on the Arctic environment and the need for action to mitigate its effects.

# VILA for HealthCare

## Zero-shot learning: training session understanding



**User:** What is the training session about?

**VILA:** The training session is about learning how to perform CPR (cardiopulmonary resuscitation) on a mannequin. The group of people, including nurses and possibly other medical professionals, are gathered around a mannequin to practice their skills and learn the proper techniques for administering CPR.

**User:** Will the practice cause pain to the patient in the image?

**VILA:** The image shows a group of people, including nurses and a mannequin, practicing CPR on a mannequin. While practicing CPR is essential for healthcare professionals, it is important to note that it is a simulation and not a real patient. The practice does not cause pain to the patient in the image, as it is a mannequin and not a living person.



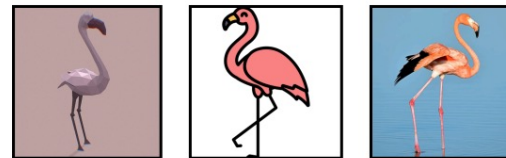
# VILA: Visual Language Model

## More on LLaVA - Visual Instruction Tuning is not Enough

**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.

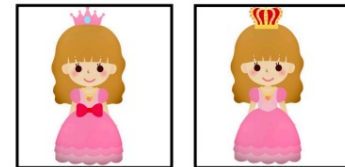


**Ours:** The common theme among these three images is the presence of a flamingo. In the first image, the flamingo is a 3D model, while in the second image, it is a pink cartoon. In the third image, it is a real flamingo standing in the water. All three images showcase the bird in different styles and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

# VILA: Visual Language Model

## Performance

Method	LLM	Res.	PT	IT	VQA <sup>v2</sup>	GQA	VisWiz	SQA <sup>I</sup>	VQA <sup>T</sup>	POPE	MME	MMB	MMB <sup>CN</sup>	SEED	LLaVA <sup>W</sup>	MM-Vet
BLIP-2 [32]	Vicuna-13B	224	129M	-	41.0	41	19.6	61	42.5	85.3	1293.8	-	-	46.4	38.1	22.4
InstructBLIP [15]	Vicuna-7B	224	129M	1.2M	-	49.2	34.5	60.5	50.1	-	-	36	23.7	53.4	60.9	26.2
InstructBLIP [15]	Vicuna-13B	224	129M	1.2M	-	49.5	33.4	63.1	50.7	78.9	1212.8	-	-	-	58.2	25.6
Shikra [10]	Vicuna-13B	224	600K	5.5M	77.4*	-	-	-	-	-	-	58.8	-	-	-	-
IDEFICS-9B [27]	LLaMA-7B	224	353M	1M	50.9	38.4	35.5	-	25.9	-	-	48.2	25.2	-	-	-
IDEFICS-80B [27]	LLaMA-65B	224	353M	1M	60.0	45.2	36.0	-	30.9	-	-	54.5	38.1	-	-	-
Qwen-VL [7]	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 [7]	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 [35]	Vicuna-1.5-7B	336	0.6M	0.7M	78.5*	62.0*	50.0	66.8	58.2	<b>85.9</b>	1510.7	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5 [35]	Vicuna-1.5-13B	336	0.6M	0.7M	<u>80.0*</u>	<b>63.3*</b>	53.6	<u>71.6</u>	61.3	<b>85.9</b>	1531.3	67.7	<u>63.6</u>	<u>61.6</u>	<u>70.7</u>	<u>35.4</u>
VILA-7B (ours)	Llama-2-7B	336	50M	1M	79.9*	<u>62.3*</u>	<u>57.8</u>	68.2	<u>64.4</u>	<u>85.5</u>	<u>1533.0</u>	<u>68.9</u>	61.7	61.1	69.7	34.9
VILA-13B (ours)	Llama-2-13B	336	50M	1M	<b>80.8*</b>	<b>63.3*</b>	<b>60.6</b>	<b>73.7</b>	<b>66.6</b>	84.2	<b>1570.1</b>	<b>70.3</b>	<b>64.3</b>	<b>62.8</b>	<b>73.0</b>	<b>38.8</b>

VILA outperforms 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.

# VILA: Visual Language Model

## Performance

Method	LLM	Res.	PT	IT	VQA <sup>v2</sup>	GQA	VisWiz	SQA <sup>I</sup>	VQA <sup>T</sup>	POPE	MME	MMB	MMB <sup>CN</sup>	SEED	LLaVA <sup>W</sup>	MM-Vet
BLIP-2 [32]	Vicuna-13B	224	129M	-	41.0	41	19.6	61	42.5	85.3	1293.8	-	-	46.4	38.1	22.4
InstructBLIP [15]	Vicuna-7B	224	129M	1.2M	-	49.2	34.5	60.5	50.1	-	-	36	23.7	53.4	60.9	26.2
InstructBLIP [15]	Vicuna-13B	224	129M	1.2M	-	49.5	33.4	63.1	50.7	78.9	1212.8	-	-	-	58.2	25.6
Shikra [10]	Vicuna-13B	224	600K	5.5M	77.4*	-	-	-	-	-	-	58.8	-	-	-	-
IDEFICS-9B [27]	LLaMA-7B	224	353M	1M	50.9	38.4	35.5	-	25.9	-	-	48.2	25.2	-	-	-
IDEFICS-80B [27]	LLaMA-65B	224	353M	1M	60.0	45.2	36.0	-	30.9	-	-	54.5	38.1	-	-	-
Qwen-VL [7]	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 [7]	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 [35]	Vicuna-1.5-7B	336	0.6M	0.7M	78.5*	62.0*	50.0	66.8	58.2	<b>85.9</b>	1510.7	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5 [35]	Vicuna-1.5-13B	336	0.6M	0.7M	80.0*	<b>63.3*</b>	53.6	71.6	61.3	<b>85.9</b>	1531.3	67.7	63.6	61.6	70.7	35.4
VILA-7B (ours)	Llama-2-7B	336	50M	1M	79.9*	<u>62.3*</u>	<u>57.8</u>	68.2	<u>64.4</u>	<u>85.5</u>	<u>1533.0</u>	<u>68.9</u>	61.7	61.1	69.7	34.9
VILA-13B (ours)	Llama-2-13B	336	50M	1M	<b>80.8*</b>	<b>63.3*</b>	<b>60.6</b>	<b>73.7</b>	<b>66.6</b>	84.2	<b>1570.1</b>	<b>70.3</b>	<b>64.3</b>	<b>62.8</b>	<b>73.0</b>	<b>38.8</b>

VILA outperforms 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.

# VILA + AWQ (4bit quantization)

## Deployable with TRT-LLM

We built the full Training-Finetuning-Quantization-Deployment pipeline for VILA, across hardware stack from the server to the edge.

# VILA

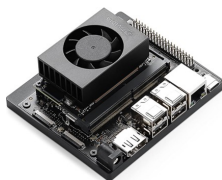
TRT-LLM



A100  
(server)



4090  
(desktop)



Jetson Orin  
(edge)

	Precision	A100	4090	Orin
VILA1.5-3B	fp16	104.6	137.6	25.4
VILA1.5-3B-AWQ	int4	182.8	215.5	42.5
VILA1.5-3B-S2	fp16	104.3	137.2	24.6
VILA1.5-3B-S2-AWQ	int4	180.2	219.3	40.1
Llama-3-VILA1.5-8B	fp16	74.9	57.4	10.2
Llama-3-VILA1.5-8B-AWQ	int4	168.9	150.2	28.7
VILA1.5-13B	fp16	50.9	OOM	6.1
VILA1.5-13B-AWQ	int4	115.9	105.7	20.6
VILA1.5-40B	fp16	OOM	OOM	--
VILA1.5-40B-AWQ	int4	57.0	OOM	--

# Benchmark on MMMU

Human Expert

Open-Source

Proprietary

Reset	Overall	Art & Design	Business	Science	Health & Medicine	Human. & Social Sci.	Tech & Eng.
GPT-4V(ision) (Playground)	55.7	65.3	64.3	48.4	63.5	76.3	41.7
SenseChat-Vision-0423-Preview*	50.3	62.7	44.1	42.3	55.7	74.7	43.5
VILA1.5*	46.9	62.1	40.6	37.7	51.7	74.0	39.5
Qwen-VL-MAX*	46.8	64.2	39.8	36.3	52.5	70.4	40.7
InternVL-Chat-V1.2*	46.2	62.5	37.6	37.9	49.7	70.1	40.8
LLaVA-1.6-34B *	44.7	58.6	39.9	36.0	51.2	70.2	36.3
Marco-VL-Plus*	44.3	57.4	34.7	38.5	48.7	72.2	36.7
Yi-VL-34B*	41.6	56.1	33.3	32.9	45.9	66.5	36.0
Qwen-VL-PLUS*	40.8	59.9	34.5	32.8	43.7	65.5	32.9
Marco-VL*	40.4	56.5	31.0	31.0	46.9	66.5	33.8
Weitu-VL-1.0-15B*	38.4	56.6	30.5	31.1	38.4	59.0	34.2
InternLM-XComposer2-VL*	38.2	56.8	32.8	30.1	39.8	60.7	31.8
Yi-VL-6B*	37.8	53.4	30.3	30.0	39.3	58.5	34.1
InfIMM-Zephyr-7B*	35.5	50.0	29.6	28.2	37.5	54.6	31.1
InternVL-Chat-V1.1*	35.3	53.7	31.7	28.2	36.5	56.4	28.0
SVIT*	34.1	48.9	28.0	26.8	35.5	50.9	30.7

# Benchmark on MMMU

Human Expert

Open-Source

Proprietary

Reset	Overall	Art & Design	Business	Science	Health & Medicine	Human. & Social Sci.	Tech & Eng.
GPT-4V(ision) (Playground)	55.7	65.3	64.3	48.4	63.5	76.3	41.7
SenseChat-Vision-0423-Preview*	50.3	62.7	44.1	42.3	55.7	74.7	43.5
VILA1.5*	46.9	62.1	40.6	37.7	51.7	74.0	39.5
Qwen-VL-MAX*	46.8	64.2	39.8	36.3	52.5	70.4	40.7
InternVL-Chat-V1.2*	46.2	62.5	37.6	37.9	49.7	70.1	40.8
LLaVA-1.6-34B *	44.7	58.6	39.9	36.0	51.2	70.2	36.3
Marco-VL-Plus*	44.3	57.4	34.7	38.5	48.7	72.2	36.7
Yi-VL-34B*	41.6	56.1	33.3	32.9	45.9	66.5	36.0
Qwen-VL-PLUS*	40.8	59.9	34.5	32.8	43.7	65.5	32.9
Marco-VL*	40.4	56.5	31.0	31.0	46.9	66.5	33.8
Weitu-VL-1.0-15B*	38.4	56.6	30.5	31.1	38.4	59.0	34.2
InternLM-XComposer2-VL*	38.2	56.8	32.8	30.1	39.8	60.7	31.8
Yi-VL-6B*	37.8	53.4	30.3	30.0	39.3	58.5	34.1
InfIMM-Zephyr-7B*	35.5	50.0	29.6	28.2	37.5	54.6	31.1
InternVL-Chat-V1.1*	35.3	53.7	31.7	28.2	36.5	56.4	28.0
SVIT*	34.1	48.9	28.0	26.8	35.5	50.9	30.7

# Resources

**Paper (CVPR'24):** <https://arxiv.org/abs/2312.07533>

**NVIDIA Developer Blog:** <https://developer.nvidia.com/blog/visual-language-models-on-nvidia-hardware-with-vila>

**NVIDIA Jetson Orin Nano Tutorial:** <https://developer.nvidia.com/blog/visual-language-intelligence-and-edge-ai-2-0>

**Repository:** <https://github.com/NVlabs/VILA>



Ji Lin



Hongxu (Danny) Yin



Wei Ping



Yao (Jason) Lu



Pavlo Molchanov



Andrew Tao



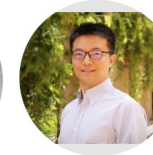
Huizi Mao



Jan Kautz



Mohammad Shoeybi



Song Han