$VILA^2$: VLM Augmented VLM with Self-Improvement

Yunhao Fang

Visual Intelligence in Large Model Era

Closed-Source

Gemini
Google

Open-Source

VILA

Qwen2-VL

Task 1: Classification

Task 2: Detection

Task n: Generation

World

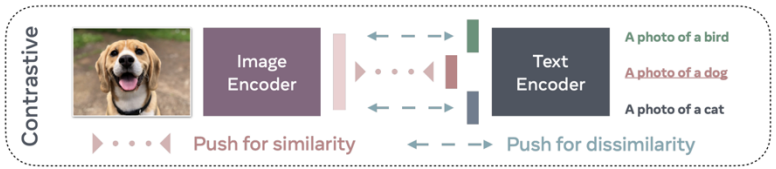
Generalist Agent

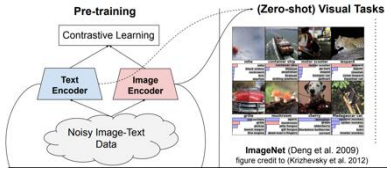
Figures from GPT-4o, Gemini, VILA, Qwen2-VL

Topics of This Talk

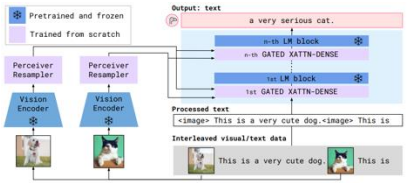
- VLM Background Introduction
- All about $VILA^2$: From Motivation to the Details
- Future Directions

Paradigm Shift: From Contrastive to Generative

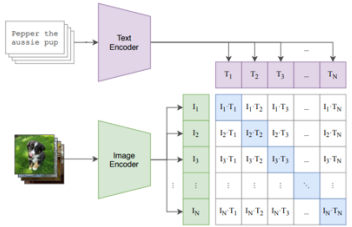




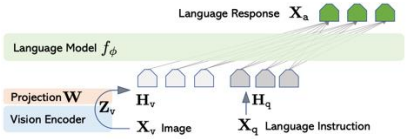
Google Research ALIGN, 2021



Flamingo, 2022



CLIP, 2021



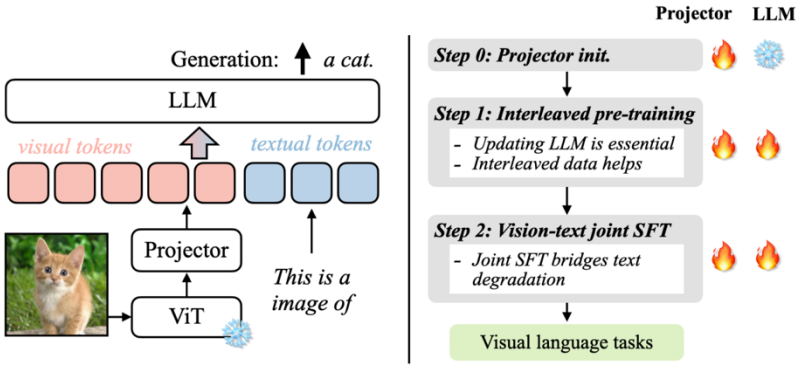
LLaVA, 2023

Figures from “An Introduction to Vision-Language Modeling”, ALIGN,CLIP, Flamingo, LLaVA

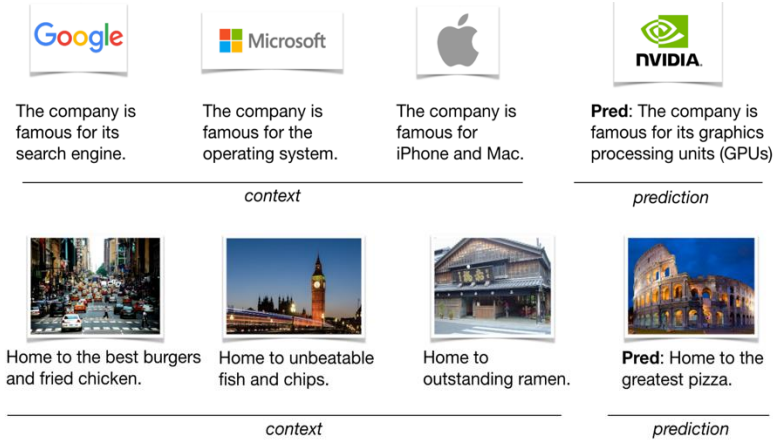
VILA (Nvidia) – Efficient AI Team

A team, a project, and a paper

Model Overview



In-Context Learning Examples



$VILA^2$: VLM Augmented VLM with Self-Improvement

Motivation:

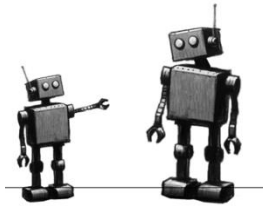
- Existing internet-scale raw images
- Distill SFT knowledge back into pre-train

In this paper, we investigated two questions:

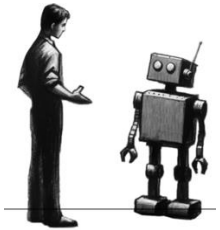
- To what extent can a VLM self-augment?
- How to surpass the limit of naive self-augmentation?

How to Improve Large Multimodal Models?

Bounded

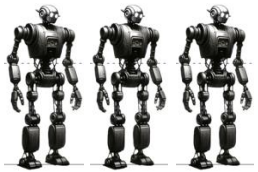


Distill from
Close-source Models



Learn from Human

Group of Teachers



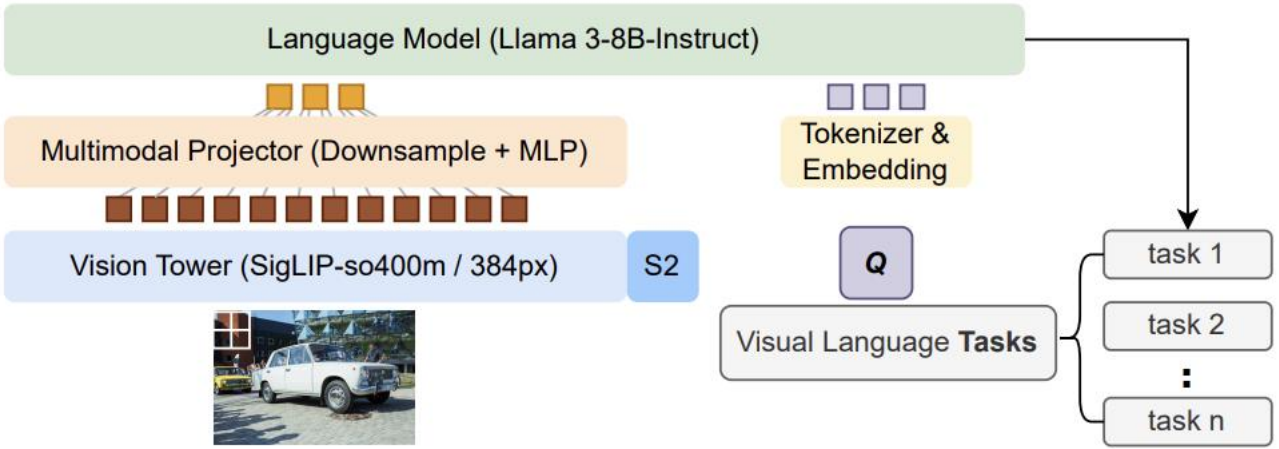
Self-evolving

- Quantitatively unscalable: Slow and high cost
- Qualitatively Insufficient: Lack of expertise / reasoning (procedure)
  No super human-level supervision

Figures from “Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision”

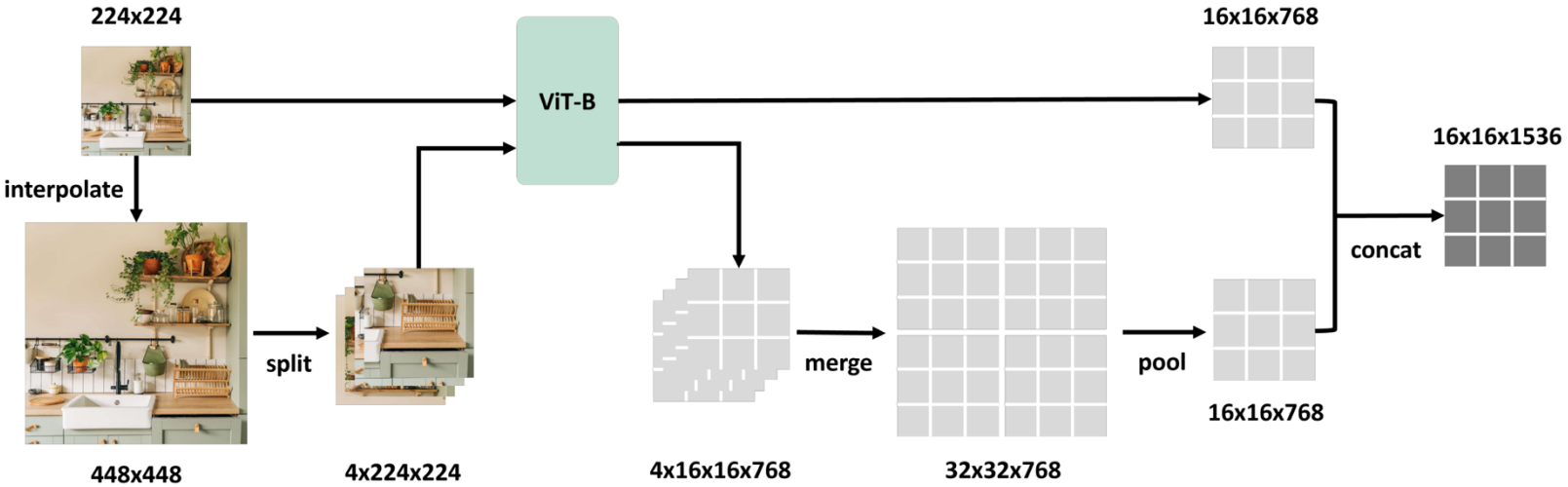
$VILA^2$: VLM Augmented VLM with Self-Improvement

Model architecture



$VILA^2$: VLM Augmented VLM with Self-Improvement

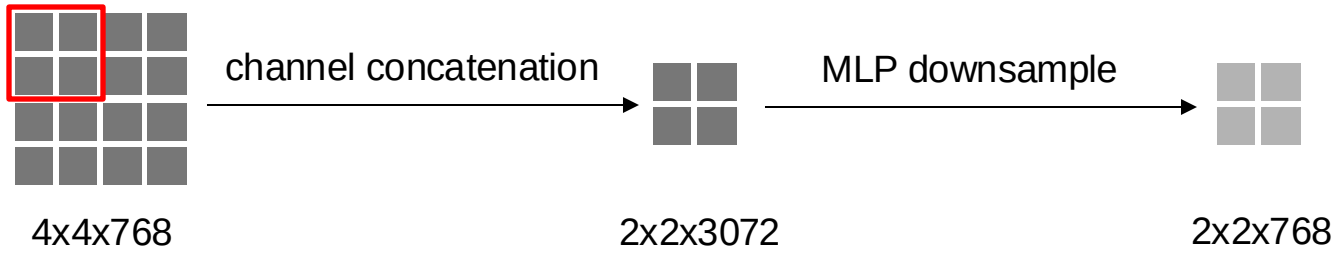
S2: Scaling up resolution instead of model size



Figures from “When Do We Not Need Larger Vision Models?”

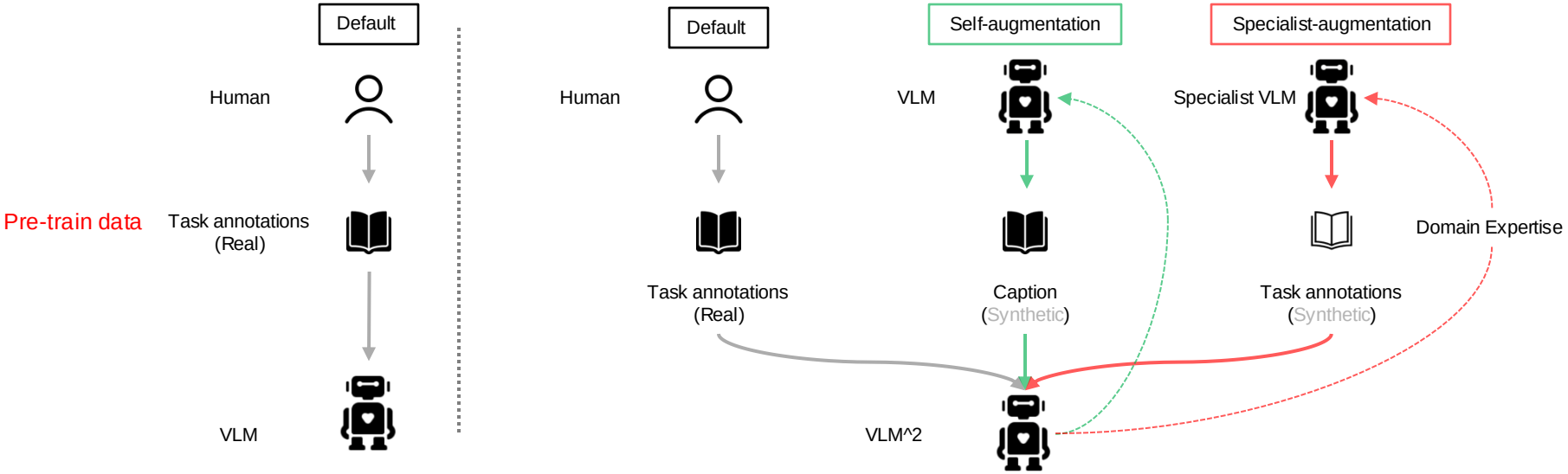
$VILA^2$: VLM Augmented VLM with Self-Improvement

Visual Token Downsample

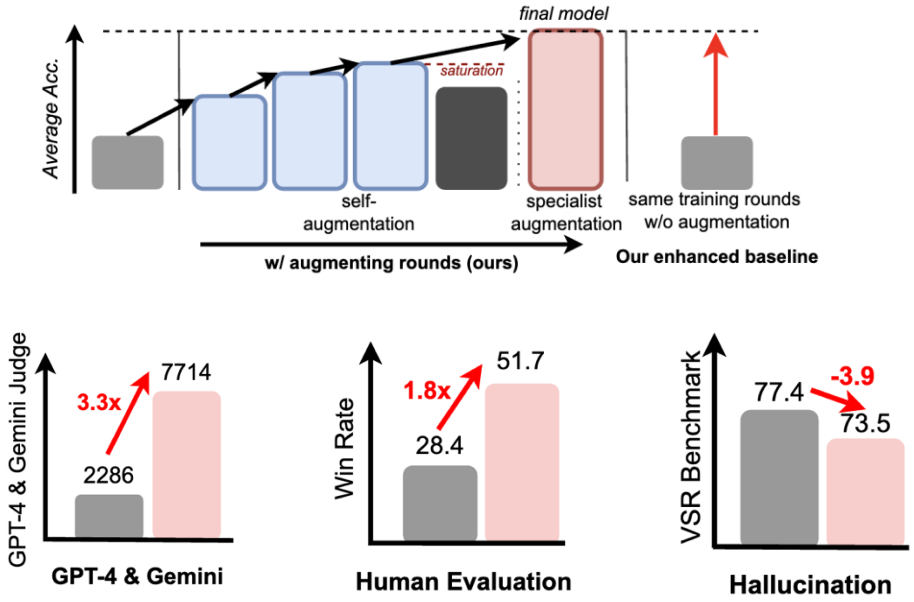


$VILA^2$: VLM Augmented VLM with Self-Improvement

Two cycles of **AI in the loop**: Self-augmentation + Specialist-augmentation

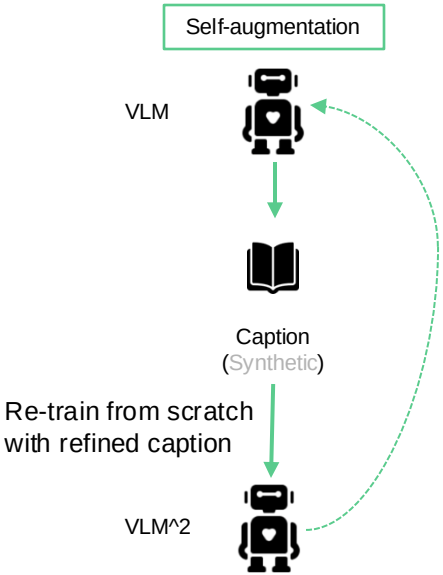


$VILA^2$: VLM Augmented VLM with Self-Improvement



$VILA^2$: VLM Augmented VLM with Self-Improvement

Self-augmentation: Iteratively refine caption quality

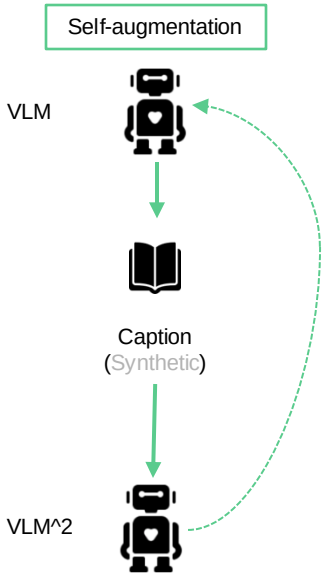


	VQA^{v2}	GQA	SQA	VQA^{T}	POPE	LLaVA^{W}	MM-Vet	MMMU
$VILA_0$ - Baseline	79.6	62.4	68.4	61.6	84.2	68.4	34.5	33.8
$VILA_1$ Pre-train with VILA0 re-cap	80.0	63.2	71.0	62.5	84.6	72.2	34.8	35.8
$VILA_2$ Pre-train with VILA1 re-cap	80.8	**63.5**	71.5	63.5	84.7	71.2	34.9	35.2
$VILA_3$ Saturated !!!	80.7	**63.5**	71.5	63.7	84.5	**72.3**	35.5	35.5
$VILA_4$	80.7	63.4	71.2	63.6	**85.0**	**72.3**	35.5	35.0

Round of Self-Augmentation

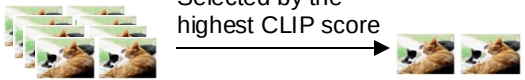
$VILA^2$: VLM Augmented VLM with Self-Improvement

Self-augmentation: Diverse prompts and human labels are helpful



- **Prompt Simple:** `<img> Describe the image briefly.`
- **Prompt Long-v1:** `<img> Describe the image in details.`
- **Prompt Long-v2:** `<img> Elaborate on the visual and narrative elements of the image in detail.`
- **Prompt Long-v3:** `<img> Instead of describing the imaginary content, only describing the content one can determine confidently from the image. Do not describe the contents by itemizing them in list form. Minimize aesthetic descriptions as much as possible.`

	Avg #words	VQA^{v2}	GQA	SQA	VQA^{T}	POPE	$LLaVA^{W}$	MM-Vet	MMMU
Baseline	17.1	79.6	62.4	68.4	61.6	84.2	68.4	34.5	33.8
Prompt Ablation for Self-Augmenting									
Self-augment Iter1 - Simple	90.4	79.4	63.0	68.7	62.4	**87.0**	68.3	34.5	33.1
Self-augment Iter1 - Long v1	94.8	80.0	62.7	**71.1**	62.2	84.0	71.7	34.5	34.4
Self-augment Iter1 - Long v2	105.4	**80.1**	63.2	70.7	62.7	84.6	71.7	34.9	34.7
Self-augment Iter1 - Long v3	102.4	**80.1**	63.4	71.0	62.9	85.0	71.4	34.4	34.7
Conversation Template Ablation for Self-Augmenting									
Mixed - re-caption text only	101.2	79.6	62.5	**71.1**	62.3	81.0	71.8	34.2	34.1
Mixed - concatenated	127.3	80.0	**63.2**	71.0	**62.5**	85.0	**72.2**	**34.8**	**35.8**



700 Million 25 Million

$VILA^2$: VLM Augmented VLM with Self-Improvement

More FLOPs on the same data?

	VQA^{v2}	GQA	SQA^{I}	VQA^{T}	POPE	$LLaVA^{W}$	MM-Vet	MMMU
$VILA_0$ (Baseline)	79.6	62.4	68.4	61.6	84.2	68.4	34.5	33.8
Train one extra epoch	-	62.5	68.7	61.9	84.0	-	34.4	33.9
Train two extra epochs	-	62.3	69.0	61.7	83.9	-	34.4	33.7

Intuition for the performance gain:

	MMC4 [26]	COYO [28]	$COYO\text{-}VILA_1$	$COYO\text{-}VILA_2$	$COYO\text{-}VILA_3$	$COYO\text{-}VILA_4$
Avg #Words	17.1 ± 25.0	11.9 ± 9.0	101.2 ± 43.0	117.1 ± 49.1	126.77 ± 50.10	125.9 ± 51.2
VQA^{v2}	N.A.	61.6	62.5	63.5	**63.7**	<u>63.6</u>

$VILA^2$: VLM Augmented VLM with Self-Improvement

A qualitative example

Raw caption: Boris Johnson riding on the street.

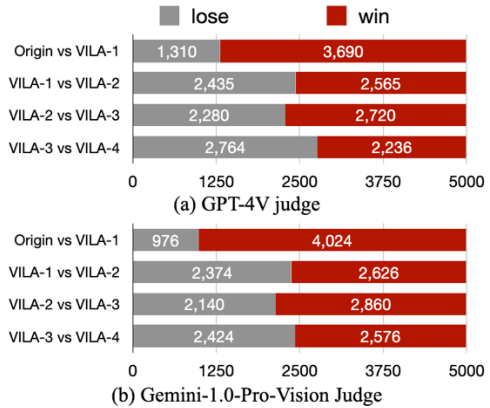
Self-Augmented Round-1: In the image, a man is seen riding a bicycle on a city street. He is dressed in a blue suit and tie, giving him a professional appearance. The bicycle he's riding has a basket attached to it, suggesting that he might be using it for transportation or leisurely rides around the city.The background of the image reveals a lively urban scene. There are buildings visible, indicating that this is likely an urban area. A group of people can also be seen walking on the sidewalk, going about their day. Adding a touch of color to the scene is a yellow flag with red and black stripes hanging from one of the buildings. This could possibly indicate the presence of a business or organization associated with the building. Overall, the image captures a moment of everyday life in a bustling city, with the man on his bicycle adding a unique element to the scene.

Self-Augmented Round-2: The image captures a moment of surprise and excitement on the face of a man, who is riding a bicycle. He's dressed in a gray jacket and blue tie, giving him an air of professionalism. His blonde hair stands out against his dark jacket, adding to the contrasting elements in the scene. The background provides context to the setting. A brick building with a yellow awning suggests it could be a commercial area or perhaps a residential neighborhood. People can be seen walking by, going about their day, oblivious to the man's unexpected ride. A notable detail is the sign for "Bacardi" visible in the background. This could indicate that the location is near a Bacardi establishment, possibly a bar or restaurant. The sign adds an element of intrigue to the scene, hinting at stories untold beyond what we see.

Self-Augmented Round-3: The image captures a lively scene on a bustling street. At the center of the frame, a man is energetically pedaling a white bicycle. He's dressed in a black jacket and blue tie, adding a touch of formality to his otherwise casual attire. His mouth is open wide as if he's shouting or laughing, perhaps expressing excitement or surprise. The background of the image provides context to the location - it appears to be a busy city street with buildings and people going about their day. The photo seems to have been taken during the day under clear skies, casting bright light that illuminates the entire scene. Despite the hustle and bustle, there's a sense of joy and freedom conveyed by the man riding the bike, making this an interesting snapshot of city life.

$VILA^2$: VLM Augmented VLM with Self-Improvement

Will rounds of self-augmentation introduces more hallucinations?



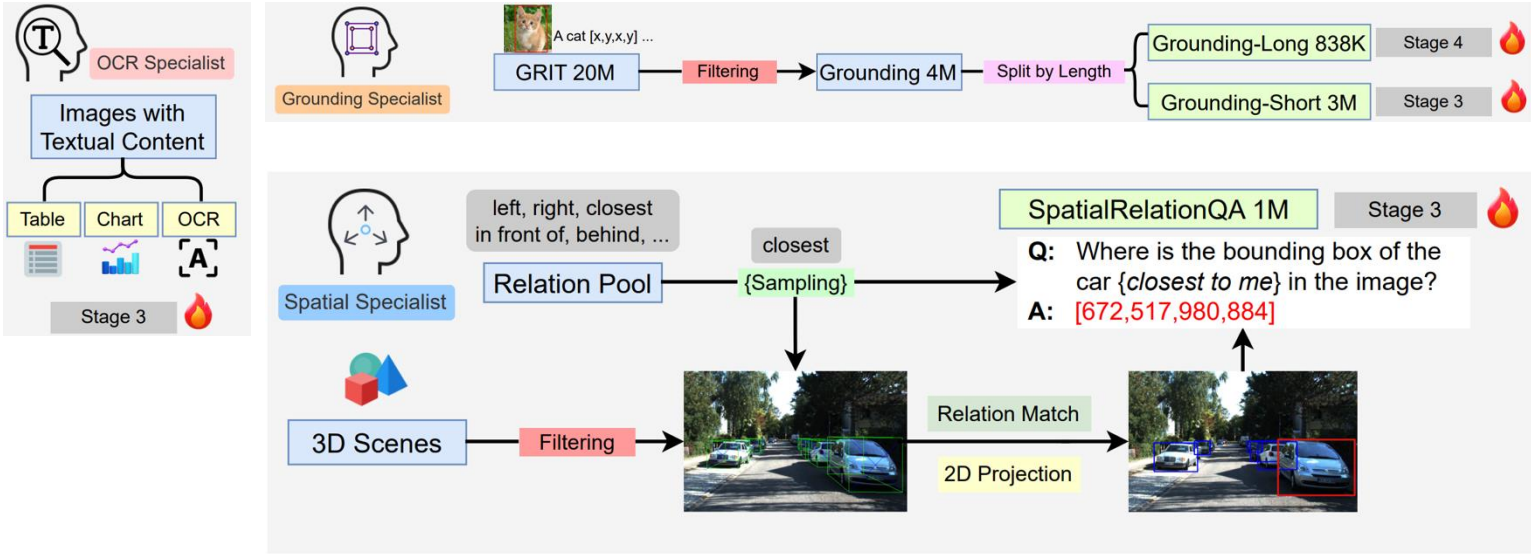
Close-source Model Judge

Metric	$VILA_0$	$VILA_1$	$VILA_2$	$VILA_3$
VSR *random* (%) ↑	73.5	75.1	76.8	**77.4**
VSR *zero-shot* (%) ↑	63.1	64.8	65.8	**66.4**
Win Rate (%) ↑	28.4	45.4	44.4	51.7

Benchmark & Human Evaluation

$VILA^2$: VLM Augmented VLM with Self-Improvement

Specialist-augmentation: Augmented with domain experts



$VILA^2$: VLM Augmented VLM with Self-Improvement

Specialist-augmentation: Prompt for task-specific annotation

Spatial Specialist
“<image> Elaborate on the visual and narrative elements of the image in detail, with a focus on spatial relations."

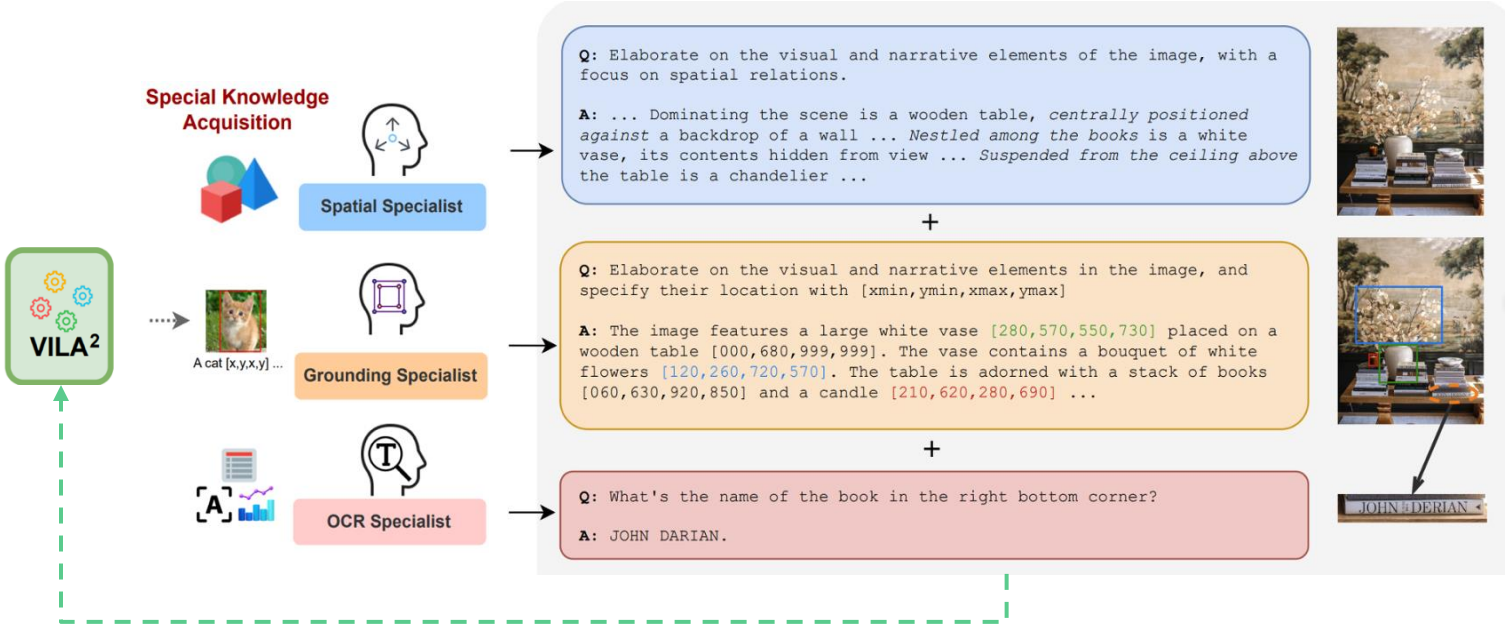
Grounding Specialist
"<image> Elaborate on the visual and narrative elements in the image, and specify their location with [xmin,ymin,xmax,ymax]."

Predict 000 – 999 special tokens w/ soft cross-entropy loss

OCR Specialist
"<image> Your task is to recognize and describe the text in the image. Please provide a brief description that focuses on the textual content."

$VILA^2$: VLM Augmented VLM with Self-Improvement

Specialist-augmentation: Overview



$VILA^2$: VLM Augmented VLM with Self-Improvement

Specialist-augmentation: Surpassing the limit of self-augmentation

Specialist-Augmented. In the image, a man is the central figure, riding a bicycle on a street. He is dressed in a gray jacket and a blue tie, giving him a professional appearance. The bicycle he's riding has a blue sign on the front, although the text on the sign is not discernible from the image. The man is not alone; he is accompanied by a group of people who are walking behind him. Their exact number is hard to determine from the image, but they appear to be a small crowd. The setting of the image is a street, with a building visible in the background. The building's details are not clear, but it adds context to the scene. The man on the bicycle is facing towards the right side of the image, suggesting that he might be moving in that direction. The people walking behind him are also facing the same direction, indicating that they might be following the man on the bicycle. Overall, the image captures a moment of everyday life, with the man on the bicycle as the main focus, and the people and the building providing the backdrop. The image does not contain any text. The relative positions of the objects suggest a dynamic scene, with the man on the bicycle leading the way for the people behind him.

	VQA^{v2}	GQA	SQA	VQA^{T}	POPE	$LLaVA^{W}$	MM-Vet	MMMU
$VILA_0$ - Baseline	79.6	62.4	68.4	61.6	84.2	68.4	34.5	33.8
$VILA_1$	80.0	63.2	71.0	62.5	84.6	72.2	34.8	35.8
$VILA_2$	80.8	**63.5**	71.5	63.5	84.7	71.2	34.9	35.2
$VILA_3$	80.7	**63.5**	71.5	63.7	84.5	**72.3**	35.5	35.5
$VILA_4$	80.7	63.4	71.2	63.6	**85.0**	**72.3**	35.5	35.0
$VILA_3$+Spatial Specialist	**81.1**	62.8	**72.9**	**65.0**	**85.0**	71.4	**37.1**	**36.8**

$VILA^2$: VLM Augmented VLM with Self-Improvement

Increase the number of experts?

	VQA^{v2}	GQA	VQA^{T}	POPE	SEED-I	MME	MM-Vet	MMMU (T)
Pretrain Data: 10% MMC4-core+10% COYO-25M+ShareGPT4V-Pretrain								
Original Caption	81.4	63.8	65.2	85.5	70.6	1472.5	34.0	31.8
+ Spatial Specialist	**81.9**$_{\uparrow 0.5}$	**64.1**$_{\uparrow 0.3}$	**66.0**$_{\uparrow 0.8}$	85.9$_{\uparrow 0.4}$	71.8$_{\uparrow 1.2}$	1476.5$_{\uparrow 4.0}$	36.7$_{\uparrow 2.7}$	32.5$_{\uparrow 0.7}$
+ OCR Specialist	81.8$_{\uparrow 0.4}$	64.0$_{\uparrow 0.2}$	65.3$_{\uparrow 0.1}$	86.4 $_{\uparrow 0.9}$	**72.1**$_{\uparrow 1.5}$	1500.2$_{\uparrow 27.7}$	34.3$_{\uparrow 0.3}$	32.1$_{\uparrow 0.3}$
+ Grounding Specialist	81.8$_{\uparrow 0.4}$	64.0$_{\uparrow 0.2}$	65.1$_{\downarrow 0.1}$	**86.7**$_{\uparrow 1.2}$	71.0$_{\uparrow 0.4}$	**1536.4**$_{\uparrow 63.9}$	**37.5**$_{\uparrow 3.5}$	**32.6**$_{\uparrow 0.8}$

Scale-up the recaption data?

Pretrain Data: MMC4-core+COYO-25M+ShareGPT4V-Pretrain								
Original Caption	82.2	63.9	66.7	**86.5**	71.2	1518.2	42.6	33.4
+ All 3 Specialist	**83.0**$_{+0.8}$	**64.7**$_{+0.8}$	**70.9**$_{+4.2}$	86.4$_{-0.1}$	**74.0**$_{+2.8}$	**1656.2** $_{+142}$	**44.7** $_{+2.1}$	**35.8**$_{+2.4}$

$VILA^2$: VLM Augmented VLM with Self-Improvement

Specialist-augmentation: Benchmark performance &

"Weak-to-strong generalization"

Method	LLM	Res.	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}	MMB	MMB^{CN}	SEED	LLaVA^{W}	MM-Vet
BLIP-2 [20]	Vicuna-13B	224	41.0	41	19.6	61	42.5	–	–	46.4	38.1	22.4
InstructBLIP [59]	Vicuna-7B	224	–	49.2	34.5	60.5	50.1	36	23.7	53.4	60.9	26.2
InstructBLIP [59]	Vicuna-13B	224	–	49.5	33.4	63.1	50.7	–	–	–	58.2	25.6
Qwen-VL [22]	Qwen-7B	448	78.8	59.3	35.2	67.1	63.8	38.2	7.4	56.3	–	–
Qwen-VL-Chat [22]	Qwen-7B	448	78.2	57.5	38.9	68.2	61.5	60.6	56.7	58.2	–	–
LLaVA-1.5 [60]	Vicuna-1.5-7B	336	78.5	62.0	50.0	66.8	58.2	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5 [60]	Vicuna-1.5-13B	336	80.0	63.3	53.6	71.6	61.3	67.7	63.6	61.6	70.7	35.4
VILA-7B [25]	Llama 2-7B	336	79.9	62.3	57.8	68.2	64.4	68.9	61.7	61.1	69.7	34.9
VILA-13B [25]	Llama 2-13B	336	80.8	63.3	60.6	73.7	66.6	70.3	64.3	62.8	73.0	38.8
LLaVA-NeXT-8B [36]	Llama 3-8B	672	–	**65.2**	–	72.8	64.6	72.1	–	–	80.1	–
Cambrian-1-8B [61]	Llama 3-8B	1024	–	64.6	–	80.4	71.7	75.9	–	–	–	–
Mini-Gemini-HD-8B [62]	Llama 3-8B	1536	–	64.5	–	75.1	70.2	72.7	–	–	–	–
VILA^2-8B (ours)	Llama 3-8B	384	**82.9**	64.1	**64.3**	**87.6**	**73.4**	**76.6**	**71.7**	**66.1**	**86.6**	**50.0**

Method	**Release Date**	**MMMU (Valid/Test)**
LLaVA-NeXT (LLM 34B)	Jan. 2024	46.7 / (missing)
MM1-30B-Chat	Mar. 2024	44.7 / 40.3
Mini-Gemini-HD (LLM 34B)	Mar. 2024	48.0 / 44.9
Cambrian-1 (LLM 34B)	Jun. 2024	49.7 / (missing)
VILA^2 (ours, LLM 34B)	May 2024	53.0 / 46.9

$VILA^2$ Extension

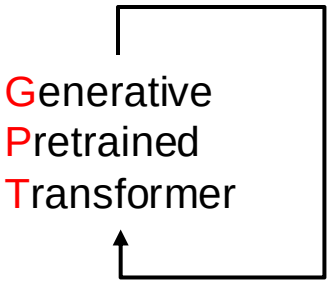
Free lunch for video understanding?

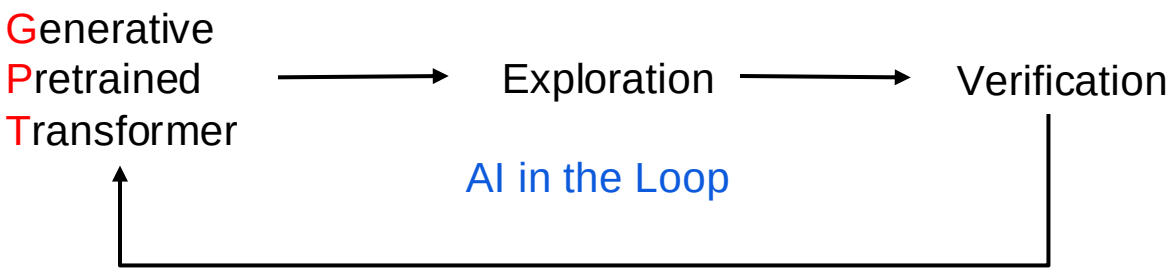
- Length generalization → Train w/ 8 frames, inference w/ 32 frames
- VILA^2 video training loop bounded by video annotation quality

Better visual-language alignment

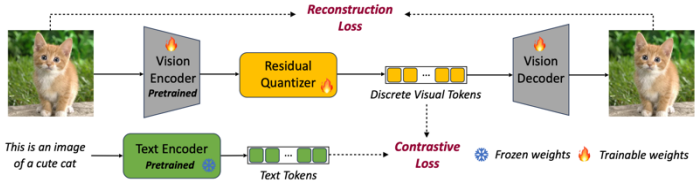
- Detailed region caption → Incorporate more details into image / video caption
- Modality synergy → Unifying generation and understanding

Close the Loop for Self-Evolving Large Models

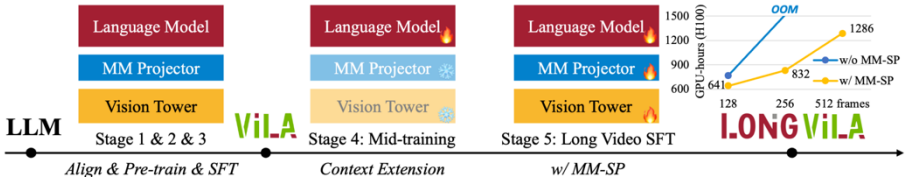




VILA Family (with More to Come):



VILA-U: Unified Understanding and Generation



LongVILA

Join VILA Community!

Github: https://github.com/NVlabs/VILA

My email: seerkfang@gmail.com

Apply for Zhijian Liu's Group: zhijian@ucsd.edu :)