MONAI Multi-modal Model (M3)

A vision-language model for medical applications that interprets medical images and text prompts to generate relevant responses.

NEXT

Holger Roth, NVIDIA

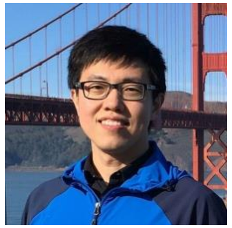
MONAI Day October 24th, 2024

Dev team.


Holger Roth


Vishwesh Nath

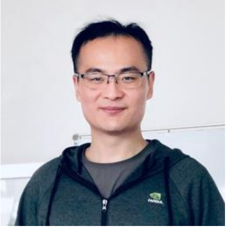

Dong Yang


Mingxin Zheng


Wenqi Li

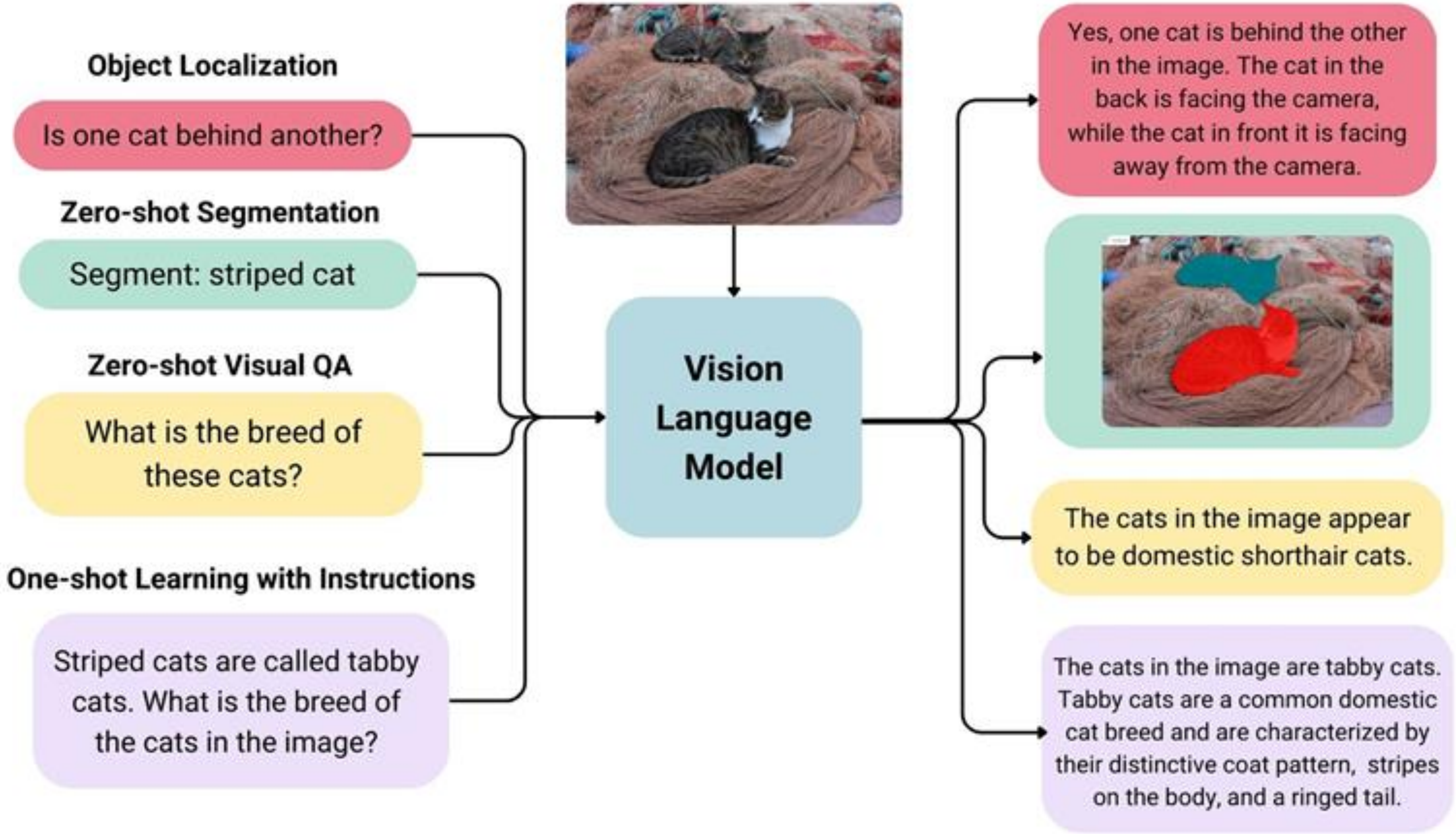

Andriy Myronenko


Daguang Xu


Nic Ma

Located: London, Maryland, Virginia, Idaho, California, Shanghai

Vision Language Models.

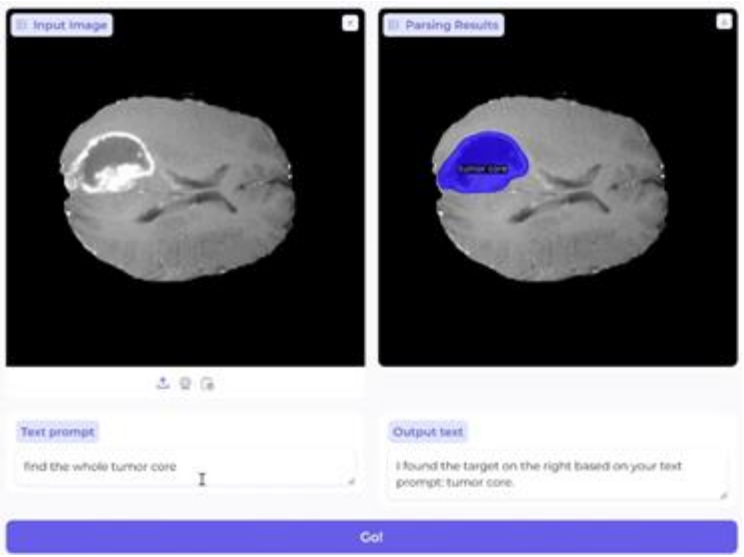


Source: Vision Language Models Explained (huggingface.co)

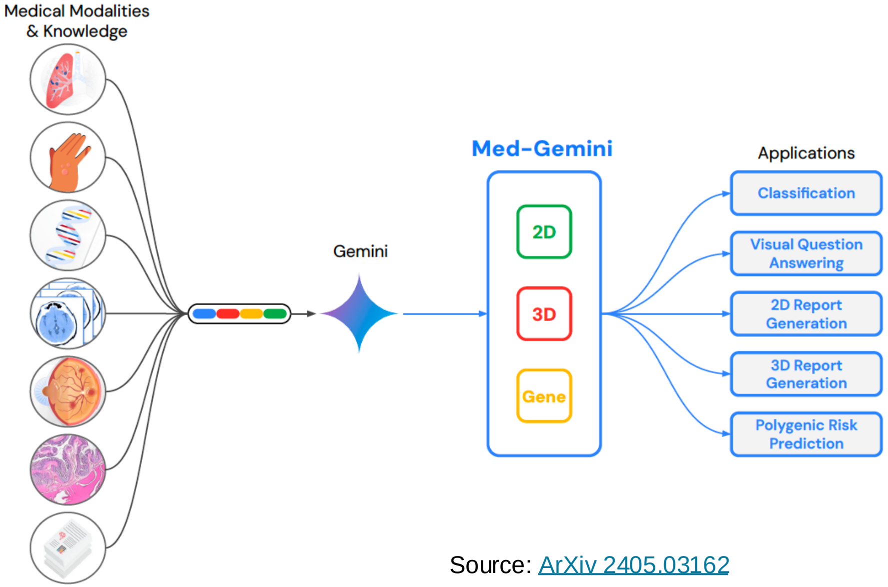
VLMs in Healthcare.

Healthcare vision language models
- Llava-Med, Llava-Rad: *task-specific*
- Med-Gemini (1.5T params): *generalist*
- BioMedparse: *prompted segmentation*
- MONAI VILA-M3 (3B-13B params): *generalist + expert model integration (segmentation, classification)*

Source: https://github.com/microsoft/LLaVA-Med

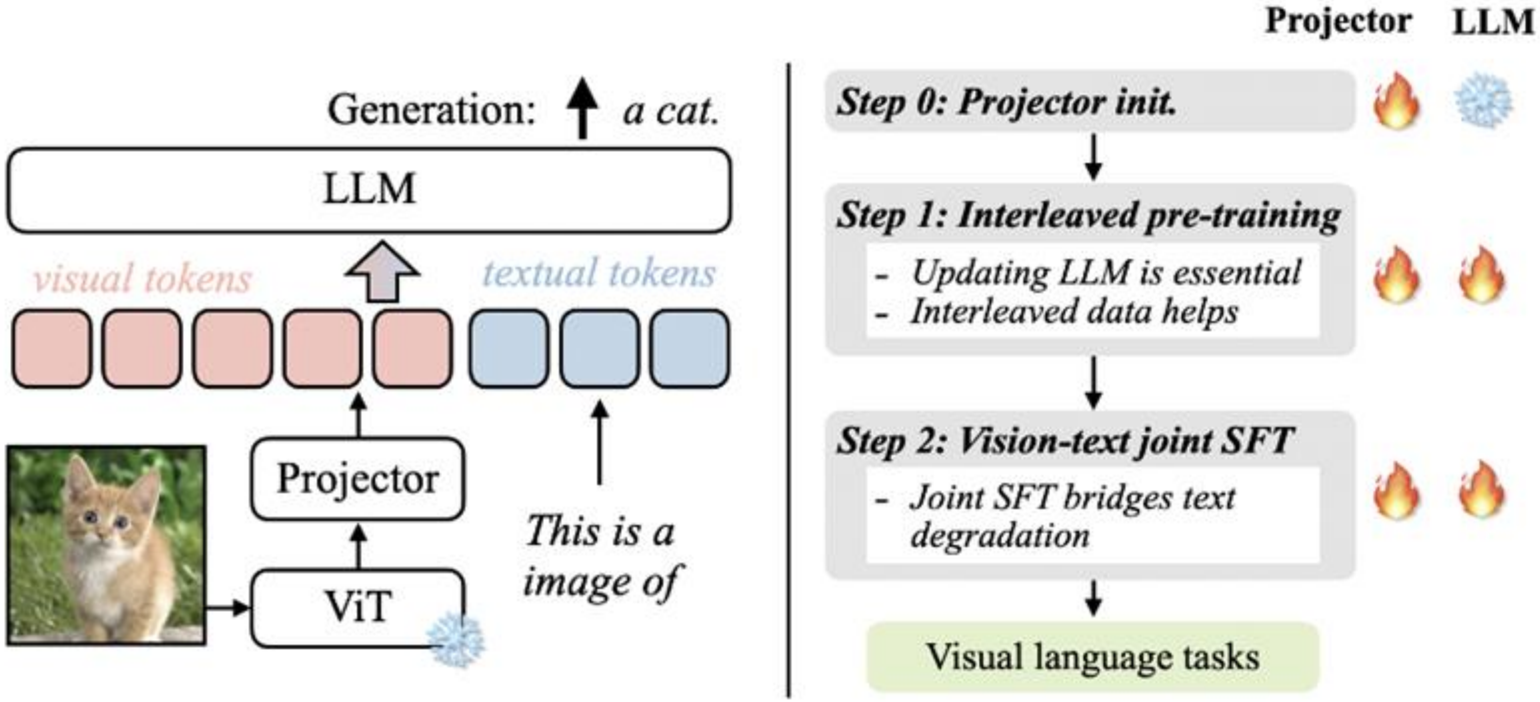


Source: https://microsoft.github.io/BiomedParse



Source: ArXiv 2405.03162

VILA: On Pre-training for Visual Language Models.

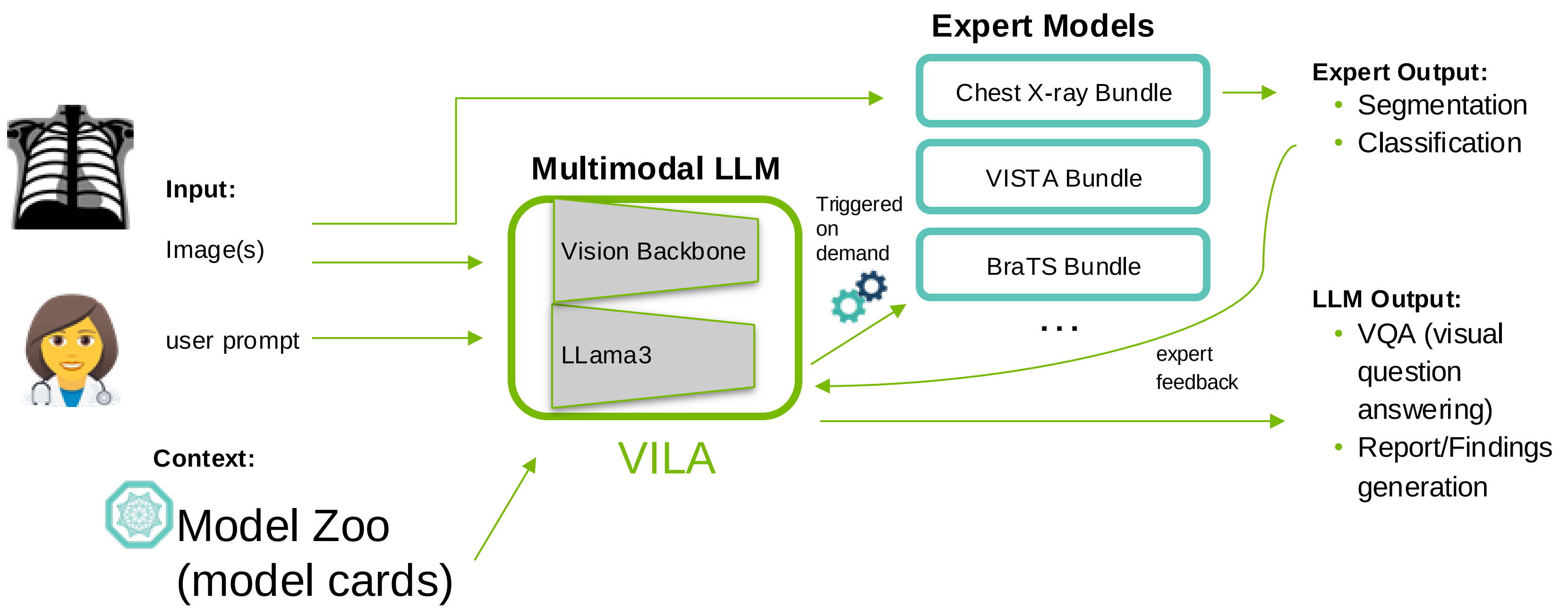


VILA 1.5

- 336x336 image resolution
- 4096 max input token length
- Up to 8 2D images supported

CVPR 2024 (Arxiv 2312.07533)

VILA-M3 Architecture.



GitHub: https://github.com/Project-MONAI/VLM

VILA-M3 Model Cards.

<VISTA3D(args)>

Modality: CT,

Task: segmentation,

Overview: domain-specialized interactive foundation model developed for segmenting and annotating human anatomies with precision,

Accuracy: 127 organs: 0.792 Dice on average,

Valid args are: 'everything', 'hepatic tumor', 'pancreatic tumor', 'lung tumor', 'bone lesion', 'organs', 'cardiovascular', 'gastrointestinal', 'skeleton', or 'muscles'

<CXR(args)>

Modality: chest x-ray (CXR),

Task: classification,

Overview: pre-trained model which are trained on large cohorts of data,

Accuracy: Good accuracy across several diverse chest x-rays datasets,

Valid args are: None

Data preparation

NEXT

How do we train VILA-M3?

VILA-M3 Train/Evaluation Data.

Dataset	QA/Text Pairs	Images	Train/Eval
PathVQA	~32,000	~4,000	Train/Eval
RadVQA	~25,000	~7,000	Train/Eval
SLAKE	~45,000	~14,000	Train/Eval
Medical-Diff-VQA	~429,000	~129,000	Train/Eval
MIMIC-CXR-JPG	~271,000	~271,000	Train/Eval
ChestXRay14	~2,000	~2,000	Eval
CheXpert	500	500	Eval
Totals	**>800,000**	**>427,000**	

Expert Selection Training Data.

Modality	Expert	Datasets
CT	*VISTA3D*	MSD (liver, spleen, pancreas), TotalSegmentatorV2
MRI	*BRATS (SegResNet)*	BRATS (2018)
Chest X-Ray	*TorchXRayVision*	MIMIC (Reports, VQA)

- Feed 2D slice selected by user to VILA-M3
- Process 2D/3D volume as supported by expert model

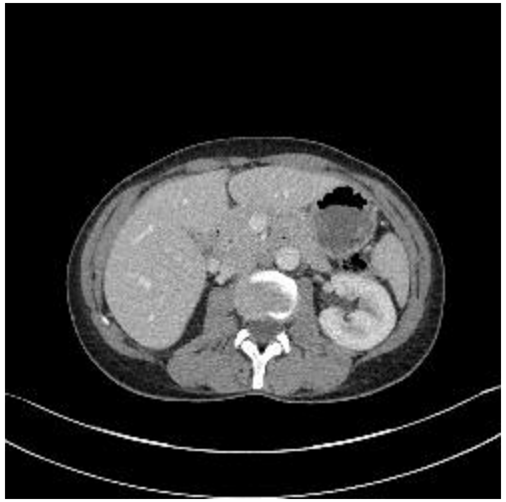
VISTA3D & BRATS models are from MONAI Model Zoo: https://monai.io/model-zoo.html
TorchXRayVision model: https://github.com/mlmed/torchxrayvision

Segmentation Task Training Data.

Llama 3.1 NIM
https://build.nvidia.com/meta/llama-3_1-70b-instruct

```
Here are some rephrased question-answer pairs with variations in the description of the segmentation
target, while maintaining a biomedical professional tone and accuracy:

**Q1:** "Identify neoplastic lesions in Liver CT"
**A1:** "Neoplastic lesions were identified using [VISTA]"

**Q2:** "Delineate tumor boundaries in Liver CT"
**A2:** "Tumor boundaries were delineated using [VISTA]"

**Q3:** "Detect cancerous growths in Liver CT"
**A3:** "Cancerous growths were detected using [VISTA]"

**Q4:** "Outline malignant masses in Liver CT"
**A4:** "Malignant masses were outlined using [VISTA]"


…

I used various synonyms for "tumors" (e.g., "neoplastic lesions", "cancerous growths", "malignant masses",
"oncological targets", "tumor tissues") and "segment" (e.g., "identify", "delineate", "detect", "outline",
"recognize", "isolate", "highlight", "define", "extract") to create diverse question-answer pairs while
maintaining accuracy and precision. Let me know if you need more!
```

Expert result feedback (VISTA3D).



'red: liver, blue: spleen, yellow: pancreas, magenta: left kidney, green: spinal cord'

Expert Model Selection with VILA.

Modality	Train	Test
CT	50k	50k
CXR	50k	50k

Epoch	3
Test CXR	99.87%
Test VISTA	99.98%

```
Wrong! Pred  vs. GT <CXR()>
Wrong! Pred  vs. GT <CXR()>
Wrong! Pred <VISTA3D(everything)> vs. GT <CXR()>
Wrong! Pred  vs. GT <CXR()>
Wrong! Pred  vs. GT <CXR()>
```

```
Wrong! Pred <VISTA3D(everything)> vs. GT <VISTA3D(hepatic tumor)>
Wrong! Pred <VISTA3D(hepatic tumor)> vs. GT <VISTA3D(pancreatic tumor)>
Wrong! Pred <VISTA3D(hepatic tumor)> vs. GT <VISTA3D(pancreatic tumor)>
Wrong! Pred <VISTA3D(everything)> vs. GT <VISTA3D(hepatic tumor)>
Wrong! Pred <VISTA3D(everything)> vs. GT <VISTA3D(hepatic tumor)>
```

Report standardization.

- Background:
  - Data curation is critical for metric calculation in captioning/report generation.
  - Manual curation is not effective/feasible.
- Automated curation using LLM ([Llama 3.1 NIM]())

  #1 – Collecting text pool

  A list of simplified sentences, focusing only on the most common findings:

  "The cardiac silhouette is normal in size."

  "The lungs are low in volume."

  "The lungs are clear."

  "No pneumothorax." ...

  #2 – Rephrasing texts in train/test set, keeping a consistent report structure.

Original report:

> **Lungs are low in volume.** Congestion of the pulmonary vasculature, small bilateral pleural effusions and presence of septal lines reflects mild pulmonary edema. Consolidations in the right mid lung and retrocardiac location could reflect a concurrent pneumonia. **Cardiac size is top normal with a normal cardiomediastinal silhouette.**

Standardized report:

> **The cardiac silhouette is at the upper limits of normal in size. The lungs are low in volume.** There is mild pulmonary vascular congestion. No pleural effusions. No focal consolidation is seen. Consolidations in the right mid lung and retrocardiac location could reflect a concurrent pneumonia.

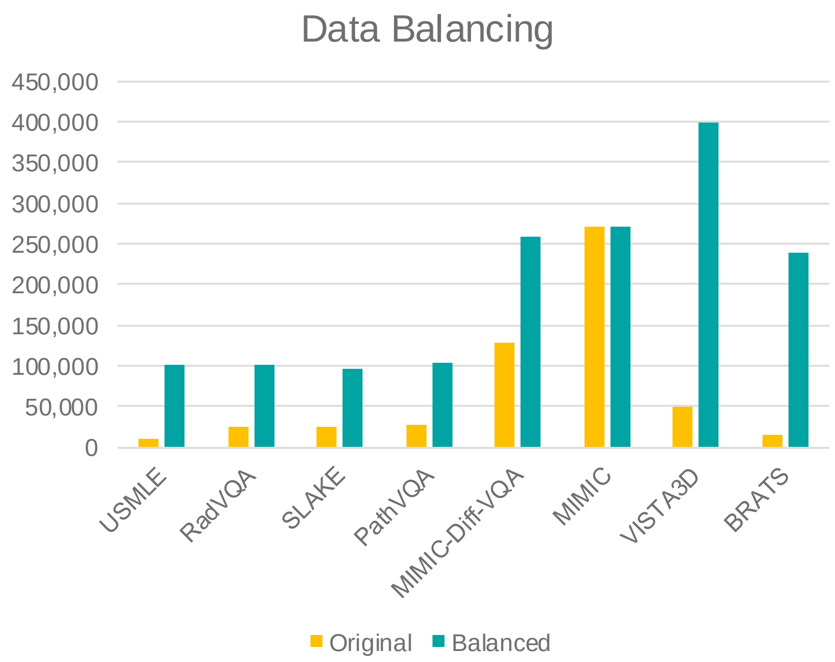
Training & Implementation

Datasets & cluster environment

NEXT

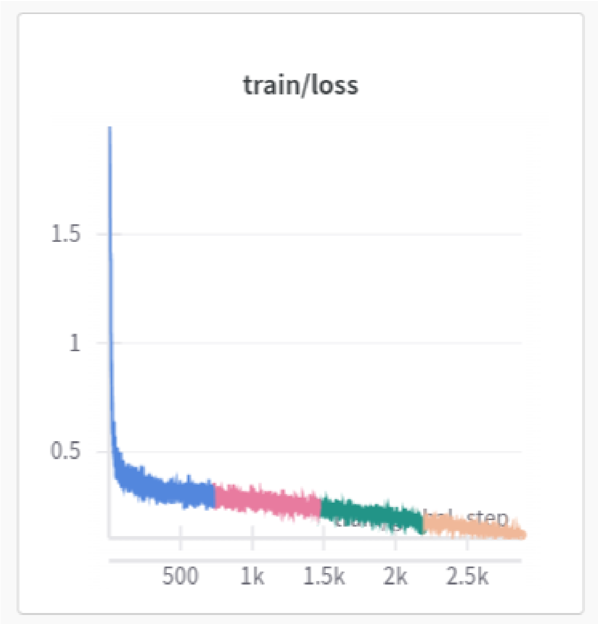
VILA-M3 Training data balancing.

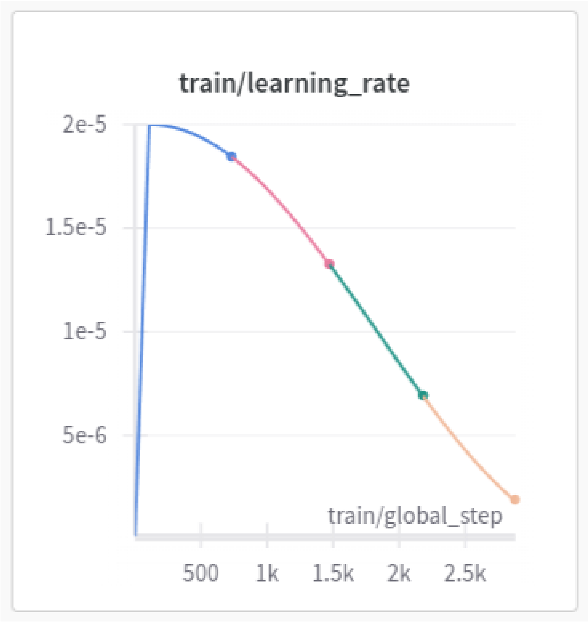
Type	Dataset	Category	Original	Frequency	Balanced
Raw	USMLE	Lang	10,178	10	101,780
Raw	RadVQA	VQA	25,124	4	100,496
Raw	SLAKE	VQA	23,888	4	95,552
Raw	PathVQA	VQA	26,034	4	104,136
Expert	MIMIC-Diff-VQA	VQA	129,232	2	258,464
Expert	MIMIC	Report	270000	1	270,000
Expert	VISTA3D	Seg	50,000	8	400,000
Expert	BRATS	Seg	15,000	16	240,000
		Total	**819,456**		**1,840,428**



Slurm cluster environment.

- VILA training code with Torch distributed
- 4 nodes with 8xA100 GPUs (80 GB each)
- Cosine learning rate decay with warmup





# Parameters	Training time
3 billion	5.5 hours
8 billion	11.0 hours
13 billion	19.5 hours

Benchmarking VILA-M3

We evaluate VILA-M3 on several different healthcare datasets & tasks

NEXT

VILA-M3 Benchmark: VQA.

Model	Type	VQA-RAD*	SLAKE-VQA	Path-VQA	Average
Llava-Med	Task-specific	*84.2*	*86.8*	*91.7*	*87.6*
Med-Gemini-1.5T	Generalist	78.8	**84.8**	83.3	82.3
Llama3-VILA-M3-3B	Generalist	78.2	79.8	87.9	82.0
Llama3-VILA-M3-8B	Generalist	**84.5**	84.5	90.0	**86.3**
Llama3-VILA-M3-13B	Generalist	80.5	83.2	**91.0**	84.9

*Comparisons to Llava-Med & Med-Gemini are not direct as data splits are not available.

VILA-M3 Benchmark: Report generation.

Model	Type	BLUE-4*	ROUGE*	GREEN*
Llava-Med	Task-specific	*1.0*	*13.3*	-
Llava-Rad	Task-specific	*15.4*	*30.6*	-
Med-Gemini-1.5T	Generalist	20.5	28.3	-
Llama3-VILA-M3-3B	Generalist	20.2	31.7	39.4
Llama3-VILA-M3-8B	Generalist	21.5	**32.3**	**40.0**
Llama3-VILA-M3-13B	Generalist	**21.6**	32.1	39.3

GREEN: Generative Radiology Report Evaluation and Error Notation (ArXiv 2405.03595)

*Comparisons to Llava-Med & Med-Gemini are not direct as data splits are not available.

Classification with expert results.

	Without Expert		With Expert	
Model	**ChestX-ray14**	**CheXpert**	**ChestX-ray14**	**CheXpert**
Med-Gemini-1.5T	46.7	48.3	-	-
TorchXRayVision	-	-	50	51.5
Llama3-VILA-M3-3B	48.4	57.4	**51.3**	60.8
Llama3-VILA-M3-8B	45.9	**61.4**	50.7	60.4
Llama3-VILA-M3-13B	**49.9**	55.8	51.2	**61.5**

In summary, we outperform Med-Gemini on 5 out of 6 tasks!

VILA Benchmark.

Model	Average
VILA base model 3B	58.3
VILA base model 8B	63.8
VILA base model 13B	63.6
Llama3-VILA-M3-3B	52.5
Llama3-VILA-M3-8B	55.1
Llama3-VILA-M3-13B	59.2

Includes image & video QA tasks from computer vision

Demo

Interactive chat with VILA-M3

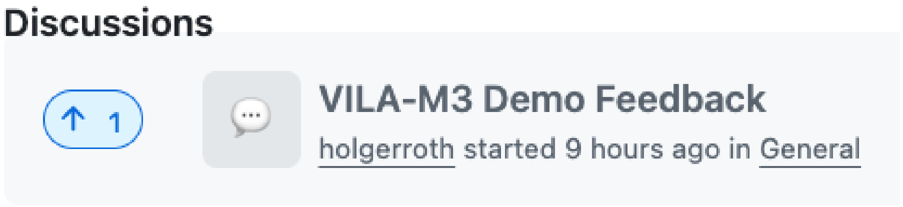
NEXT

Next steps.

- Parameter-efficient fine-tuning (LoRA, DoRA)
- Quantization for faster inference (AWQ)
- Direct 3D support

Try it out:
https://vila-m3-demo.monai.ngc.nvidia.com

Feedback:
https://github.com/Project-MONAI/VLM/discussions/35



Call for VLM Working Group

If you are interested, please contact us on MONAI Slack **#vlm**

NEXT

Thank you!

NEXT

GitHub: **https://github.com/Project-MONAI/VLM**

MONAI Hugging Face Hub: **Fine-tuned checkpoints coming soon!**

MONAI slack channel: **#vlm**