VILA-M3: Enhancing Vision-Language Models with Medical Expert Knowledge

Supplementary Material

1. Dataset Details

1.1. Dataset Balancing

As shown by experiments in the *main* manuscript, dataset balancing is necessary to achieve the best performance when combining many datasets together (Main Manuscript, Figure 5). To balance the datasets, we categorized them into three broad categories: visual question answering (VQA), report generation, expert segmentation data, and language (see *Category*, Table. 1). For example, if the dataset counts are summed up category-wise, their ratios for the entire dataset can be estimated. The proportion of each category differs significantly between the original and balanced versions. While VQA was modestly increased (from 24.9% to 30.4%), language and expert segmentation data were substantially increased (1.2% to 5.5% and 7.9% to 34.8%, respectively), whereas report generation was decreased (33.0% to 14.7%).

Type	Dataset	Category	Original	Freq.	Balanced
Raw	USMLE	Lang	10,178	10	101,780
Raw	RadVQA	VQA	6,281	16	100,496
Raw	SLAKE	VQA	5,972	16	95,552
Raw	PathVQA	VQA	26,034	4	104,136
Expert	MIMIC-Diff-VQA	VQA	129,232	2	258,464
Expert	MIMIC	Report	270,000	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

Table 1. Balanced training dataset statistics showing original and balanced sample counts.

1.2. Report Dataset Curation

The preparation involves downloading datasets and refining report text with a Large Language Model (LLM) to create high-quality inputs for generating reliable medical reports. The primary dataset used is the **MIMIC Chest X-ray JPG Database v2.0.0** [3, 4], which contains over 377,000 images and 227,827 free-text radiology reports. These data are de-identified to comply with HIPAA requirements. The process incorporates text enhancements and cleansing to optimize the quality of report inputs.

1.2.1 Download Datasets

The process begins with downloading the MIMIC-CXR-JPG dataset, which provides chest X-ray images and corresponding radiology reports. Data splits and labels are standardized, and enhanced text versions are utilized for improved report quality. The dataset is designed for tasks involving medical image analysis and natural language processing. Enhanced text versions ensure clarity and remove noise for better performance in model training. To refine the quality of the reports and eliminate noise, we utilize an enhanced text version developed by DCL [6], and subsequently apply additional cleansing procedures to further optimize report accuracy.

1.2.2 Sentence Pool Collection

To standardize the language used in medical reports, a pool of sample sentences is created. This pool consists of commonly recurring phrases or sentence structures that appear in radiology reports. Using python scripts, the LLM (Llama-3.1-8b-instruct) analyzes the dataset and extracts these patterns, which are stored in a file (e.g., `sentence-pool.txt`) to guide the text transformation process.

Example: Sentence Pool:

- The cardiac silhouette is normal in size.
- The lungs are low in volume.
- No pleural effusions.
- No pulmonary edema.
- There is mild pulmonary vascular congestion.

Prompt Example for Collection:

"Analyze the medical report dataset and extract recurring sentence structures. Compile these into a pool of sample sentences to be used for standardizing input text."

1.2.3 Text Conversion Using LLM

The LLM (Llama-3.1-8b-instruct) utilizes the sentence pool to process the text, replacing or reformatting sentences and medical terminology into a consistent and standardized format. This step ensures uniform input data for the VLM model, improving its ability to generate reliable and accurate reports.

Example:

- **Before Conversion:**
  "Lungs are low in volume. Congestion of the pulmonary vasculature, small bilateral pleural effusions and presence of septal lines reflect mild pulmonary edema."
- **After Conversion:**
  "The lungs are low in volume. There is mild pulmonary vascular congestion. No pleural effusions are seen."

Prompt Example for Conversion:

"Using the provided sentence pool, transform the medical report text into a consistent format. Ensure the output text aligns with the predefined sentence structures without altering the content's meaning. Return only the standardized text."

1.3. Expert Dataset Curation

Segmentation: For segmentation models, we can train VILA-M3 to trigger a suitable expert model together with several different input arguments. For example, when triggering the VISTA3D model, we generate several variations (using Llama-3.1) of instructions to the model to segment a certain anatomical structure in the image and give a suitable trigger message. For instance:

- User: `Can you identify any liver masses or tumors?`
- VILA-M3: `This looks like a CT image. Let me trigger <VISTA3D(hepatic tumor)>.`
- User: `Can you assist me in segmenting the bony structures in this image?`
- VILA-M3: `I segmented the skeleton using <VISTA3D(skeleton)>.`
- User: `Segment the entire image.`
- VILA-M3: `I segmented the entire image using <VISTA3D(everything)>.`

For these types of instructions, VILA-M3 predicts the suitable arguments for the VISTA3D model in order to trigger the segmentation of the correct anatomical structure as illustrated in Fig. 1. We built these types of instruction and answer pairs into the expert training dataset in order to build segmentation capabilities into VILA-M3. Table 2 summarizes the expert models and datasets used in this work to build up the expert training dataset.

Table 2. Expert Selection Training Data.

Modality	Expert	Datasets
CT	VISTA3D	MSD (liver, spleen, pancreas), TotalSegmentatorV2
MRI	BRATS (SegResNet)	BRATS (2018)
CXR	TorchXRayVision	MIMIC (Reports, VQA)

Report Generation: Converting expert model predictions into conversation format involves transforming classification outputs into structured dialogues for AI training in medical report generation. Initially, ensemble predictions are created by combining probabilities of various medical conditions (like Atelectasis and Effusion) from multiple expert models from TorchXRayVision applied to chest X-ray images. These probabilities are interpreted to determine the presence ("yes") or absence ("no") of each condition based on a threshold.

The formatted expert predictions are then integrated into conversation prompts that include an image placeholder, a prompt for report generation, and the expert results as additional information. For example, the prompt might be:

```
<image>
Describe the image in detail.
When answering, please
incorporate the expert model
results:
Atelectasis: yes
Cardiomegaly: no
Effusion: yes
```

Each conversation is assembled into a structured format containing a unique identifier, the image reference, and the dialogue between the human and the VLM model. This approach enriches the dataset with various prompt types and simulates interactions where the VLM model provides diagnoses based on image analysis and expert insights, enhancing the model's ability to generate accurate and contextually rich medical reports.

2. Training & Compute Details

2.1. Training Details

The loss curves for all model variants when training for 2 epochs can be observed in Fig. 2. The largest change in the training loss can be observed between 3B and 8B, 13B and 40B models. There is a significant learning gap that one can observe between 3B and 8B onwards, which demonstrates that the 3B model learning capabilities are limited. It should also be noted that the 40B model has a noisier training curve as compared to all other models, this can be attributed to two major factors, the first being that the vision backbone is much larger (6B parameters) as compared to other configurations and the second factor being that it is a Yi model [7] and their learning behavior is different. The training trends overall also indicate that larger models have diminishing returns in terms of loss convergence, note that the 8B, 13B and 40B are quite close in terms of the final loss values. The 40B model loss stays quite high during training and takes a longer time to converge below the loss values of other models, indicating that it is slower to train.

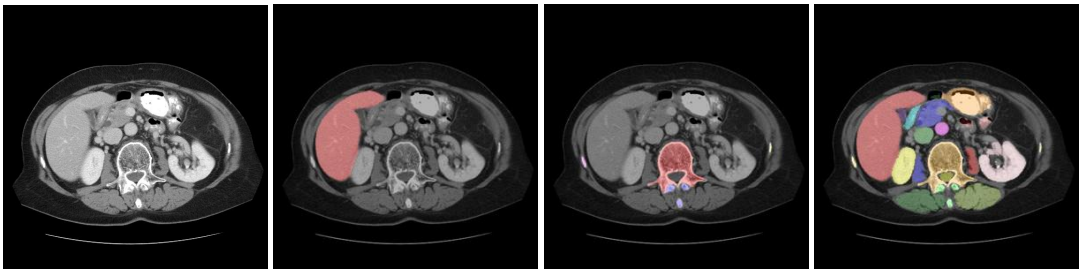
2.2. Inference Compute

We have successfully deployed an inference workflow of the proposed framework. While there are additional techniques (such as TensorRT) could be used to improve the system throughput, we list the computational costs without 'bells and whistles' to show the technical feasibility of practical usages in Table 3.

3. Additional Experiments

3.1. Scaling Law Experiment Additional Details

In the Main Manuscript, Fig. 6 shows that after an initial warm-up period, the learning curve of the fine-tuning pro-



(a) original image (b) `hepatic tumor` (c) `skeleton` (d) `everything`

Figure 1. (a) The original CT image slice. (b-d) The selected argument by VILA-M3 to the VISTA3D expert model call.

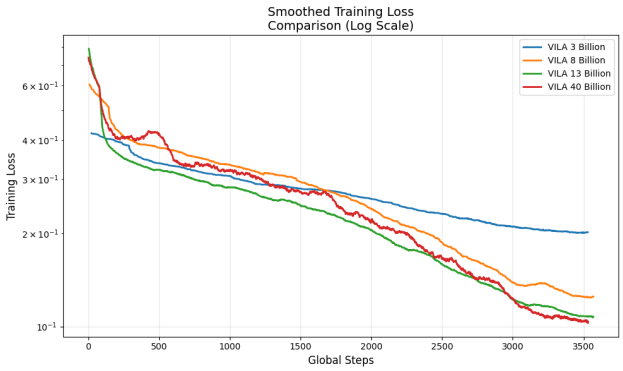


Figure 2. Comparison of training trends across all model variants by parameter size. The compared variants are 3, 8, 13, and 40 billion parameter models.

Model		VLM Properties		
Type	# Params	VRAM (GB)	tokens/sec	Max context len
VILA-M3	3B	7	41	4,096
VILA-M3	8B	18	32	4,096
VILA-M3	13B	30	26	4,096
VILA-M3	40B	77	9	4,096

Table 3. Inference computational costs for VILA-M3 variants.

Table 4. Fits to $L(N, S)$ of the loss scaling law defined in [25] Equation 5.6. These parameters are visualized in Main Manuscript Fig 6.

Parameter	α_N	α_S	N_c	S_c
Value	0.78	1.09	1.50×10^8	3.92×10^2

cesses can be approximately fit by a universal function parameterized by model size and number of steps (using 32 GPUs in this case). Table 4 summarizes the empirical fit of the power law parameters based on Equation 5.6 in [5]. It is interesting to observe that although our fine-tuning dataset size is relatively small compared with a typical LLM training scratch setup, the training process scales according to the model size and training steps following a similar pattern.

3.2. Balanced & Unbalanced Datasets Training

Since the original size of datasets varies a lot as can be observed in Table. 1. In the main manuscript we showed the comparison between balanced and unbalanced datasets for the 3B model. Here, we plot additional results for the 8B and 13B models both as shown in Fig.3.

The improvements based on the balanced training dataset as compared to the unbalanced dataset (original dataset size) can be observed in Fig. 3. Quantitatively, an average improvement of $\sim$4% of all metrics is gained by data balancing for the 3B and 13B models. The 8B model shows an improvement of $\sim$2%. Furthermore, the training trends in Fig. 4 indicate that the balanced dataset provides an earlier convergence in the loss and an overall lower loss training trend is achieved.

3.3. Comparisons With GPT-4o For Classification

Tables 5 & 6 (Main Manuscript) show a comparison with GPT-4o [1, 2] for classification experiments. The inference on GPT-4o was performed with the same prompt as being used for VILA-M3. The images were pre-appended to the prompt via API call using a python script and then the responses were collected with retry attempts set to 10.

We often found that with API usage the GPT-4o provided inconsistent responses as it took more 3-5 retries for more than quite a few images for both CheXpert and Chest X-ray datasets. We also found that if we appended the *expert model information*, the GPT-4o refused to provide responses for many images and therefore a quantitative analysis would be unfair. For more than 50% images responses could not be successfully retrieved. We believe the guardrails of GPT-4o likely come into effect when a sufficient amount of medical terminology is used within the prompt itself.

3.4. Statistical significances of the classification experiments

We conduct McNemar's Chi-square tests to show that the proposed M3 models (with expert models) significantly differ from GPT-4o (results shown in Main Manuscript Table 5

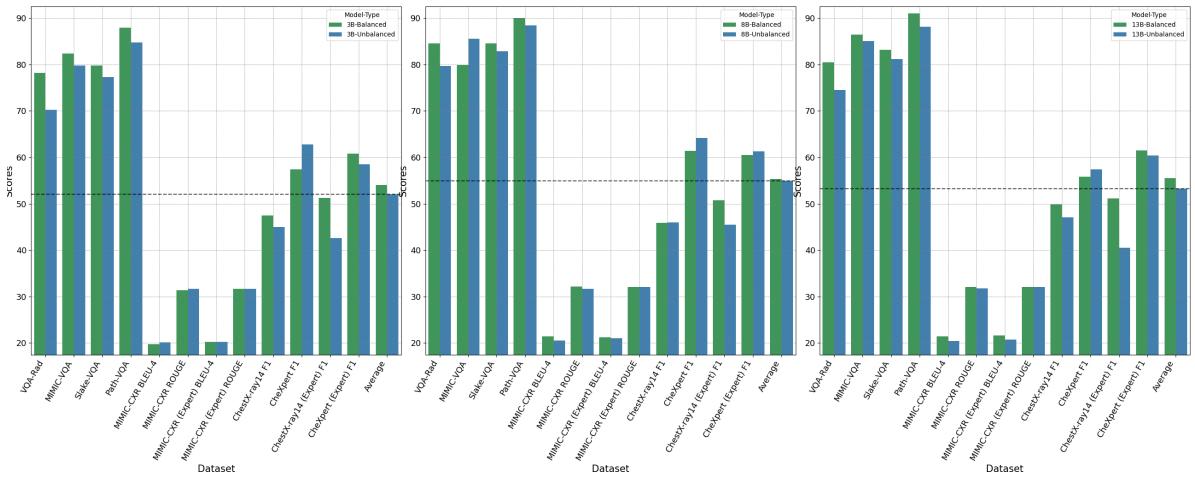


Figure 3. Comparison of VILA-M3 training with balanced and unbalanced healthcare datasets. Comparison for 3B, 8B and 13B model are shown with a training of two epochs each

Figure 4. Comparison of training trends across 3B model to show the effect of balanced and unbalanced dataset on model training.

and 6). The values that are smaller than 0.05 indicate significance. For the corresponding model and specific class denoted with *, M3 is significantly better than GPT-4o.

	Fracture	Pneumothorax	Lung opacity
M3-3B	5.5e-10*	3.2e-5*	6.1e-41*
M3-8B	1.3e-1	3.1e-1	4.5e-14*
M3-13B	5.9e-20*	5.0e-8*	1.0e-44*
M3-40B	3.1e-12*	2.5e-8*	3.2e-49*

Table 5. p-value for the null hypothesis that the classification models perform at the same level as GPT-4o for each class on ChestXray14.

3.5. Additional Results Expert-Guided IFT

From the main manuscript in Fig 4. it was observed that the 8B model overfits when the 3 epoch variant is trained. In Fig. 5 results for all datasets for all model variants (3B, 8B, 13B and 40B) can be observed. The performance deterioration for all datasets is evident for the 3 epoch model for all

	Atel.	Cardio.	Consol.	Edema	Pl. Eff.
M3-3B	2.3e-53*	2.1e-37*	6.5e-43*	7.0e-48*	2.6e-51*
M3-8B	3.7e-59*	1.0e-16*	3.4e-33*	1.3e-30*	1.3e-32*
M3-13B	2.0e-46*	2.9e-30*	4.6e-22*	1.3e-41*	1.3e-32*
M3-40B	6.6e-52*	2.5e-18*	2.8e-23*	3.2e-47*	3.8e-40*

Table 6. p-value for the null hypothesis that the classification models perform at the same level as GPT-4o for each class on CheXpert.

variants with the exception of report generation tasks. Since the report generation task is of a much more complex nature the models do not overfit to it.

3.6. Additional Results on Report Generation

We conducted a detailed analysis of the test set results for report generation, focusing on the disease and pattern distribution (multi-label ground truth) from the **MIMIC Chest X-ray JPG Database v2.0.0**. This evaluation utilized the best-performing 40B model, covering both with and without the integration of expert model predictions. Table 7 shows performance metrics with and without expert model predictions across various categories of medical findings. Key metrics include `BLEU-4`, `ROUGE`, and GREEN score (`GREEN`), evaluated for both settings. For most categories, the inclusion of expert models resulted in marginal improvements in the `GREEN` metric, which measures overall accuracy. For instance, in the *Atelectasis* category, `GREEN` improved from 35.93 to 36.56, and in *Cardiomegaly*, it increased from 38.99 to 39.79. However, some metrics, such as `BLEU-4` in categories like *Fracture* and *Lung Lesion*, showed slight decreases when expert model predictions were used.

Notably, the *No Finding* category exhibited the largest improvement in `green` (45.60 to 46.57), indicating that

expert models may be particularly effective in identifying cases with no abnormalities. These results suggest that while the integration of expert models generally enhances certain metrics, the gains are context-dependent and may vary across different medical conditions.

4. Additional Discussion

As also stated in the main manuscript, we will explore the integration of Retrieval-Augmented Generation (RAG) to further enhance VILA-M3 by retrieving and incorporating relevant information from large datasets dynamically during inference.

Further improvements would be to make an agentic framework with expert models, however it requires careful design. However, implementing an agentic framework necessitates careful design to balance autonomy with control, ensuring the system remains reliable and interpretable. Addressing these design challenges will be crucial for advancing the model's capabilities and for its successful deployment in complex real-world applications.

The results of this work indicate that we need to carefully curate data and factor in expert information from the many medical domain-specialized models that have been trained in the past.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[3] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 1

[4] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 1

[5] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 3

[6] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343, 2023. 1

[7] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 2

Table 7. Detailed performance on report generation with and without expert models.

Category	Quantity	Without Expert			With Expert		
		BLEU-4 (↑)	ROUGE	GREEN	BLEU-4	ROUGE	GREEN
Atelectasis	494	19.21	31.32	35.93	19.03	31.31	36.56
Cardiomegaly	426	18.62	32.11	38.99	18.38	31.98	39.79
Consolidation	99	18.46	31.47	37.23	18.20	30.70	34.97
Edema	451	20.69	33.39	38.97	20.54	33.68	39.66
Enlarged Cardiomediastinum	70	16.10	27.63	31.86	16.67	28.57	35.28
Fracture	53	19.12	31.07	37.51	17.24	30.20	35.88
Lung Lesion	83	20.75	32.92	38.89	18.64	31.56	36.17
Lung Opacity	759	18.60	30.35	35.92	17.96	30.18	35.30
No Finding	561	21.20	34.25	45.60	21.81	34.16	46.57
Pleural Effusion	678	18.52	30.60	34.53	18.20	30.44	35.12
Pleural Other	37	19.27	29.37	32.09	17.32	27.42	36.10
Pneumonia	219	19.39	32.30	38.41	20.14	32.91	39.43
Pneumothorax	58	18.76	29.97	30.02	17.39	28.62	27.71
Support Devices	635	19.25	30.70	35.24	19.19	30.42	35.91

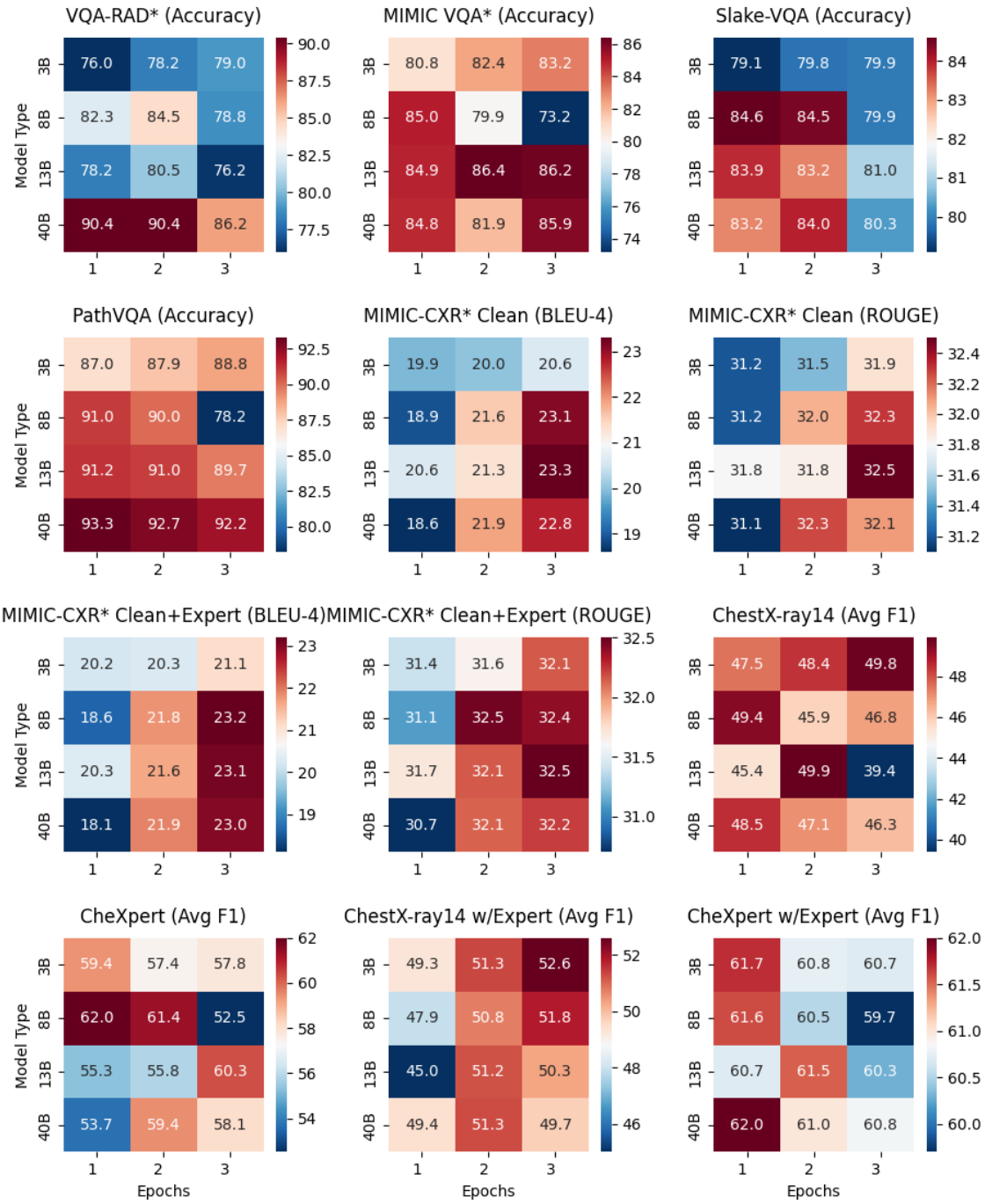


Figure 5. The heatmap shows the performance of the all model variants 3B, 8B, 13B and 40B on all datasets with trained models at 1, 2 and 3 epochs.