Chen, Wen-Kuang (David Chen)

VLM & Training Optimization

Motivation & Goal

Model answers human questions based on the content of an image.

Requires both visual perception and natural language understanding.

VQA combines image understanding and text reasoning into one model

Challenge: handle diverse real-world scenes and ambiguous questions.

Example: “What is on the table?” → “A coffee cup and a book.”

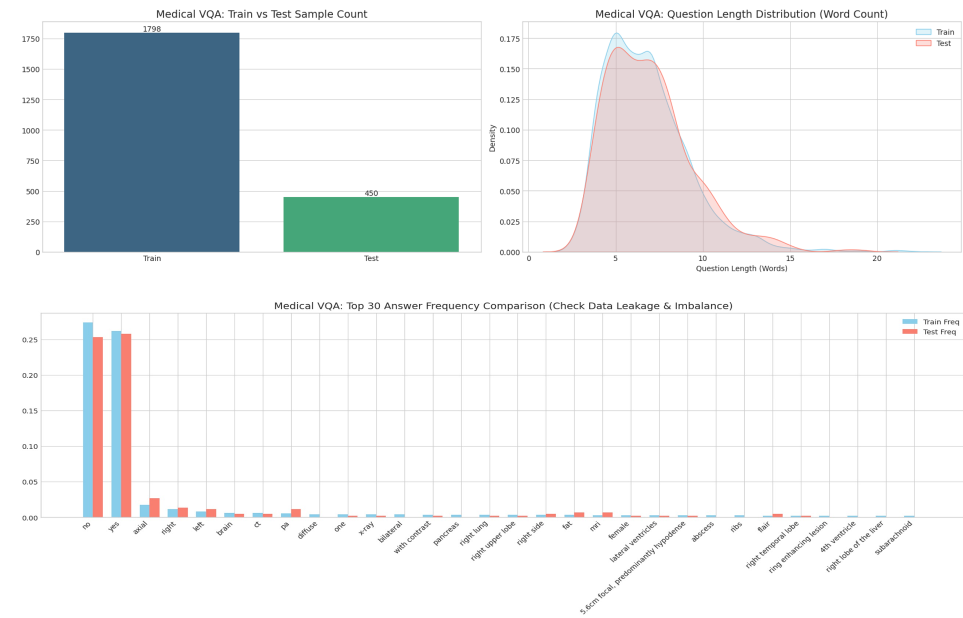
Full fine-tuning on large models is expensive and memory intensive.

Developing VLM model training and finetuning process.

Maintaining accuracy while cutting compute cost and training time.

Data

- Medical Visual Question Answering
  - Includes radiology and clinical images paired with medical questions.
  - Short medical phrases (yes/no, pneumonia, left lung opacity)
  - ~3,000–4,000 QA pairs
  - Long-tail labels
- DAQUAR
  - Indoor scenes
  - Mostly single words (chair, red, 2)
  - ~12,000 QA pairs
  - Image diversity very high




```
📊 Yes/No Question Ratio:
  - Train: 53.56%
  - Test:  51.11%
📊 Vocabulary Size:
  - Train Unique Answers: 458
  - Test Unique Answers:  167
  - Unseen Answers in Test: 59 (Model cannot predict these in classification)
```

Training Flow

Data → Training → Evaluation

Data
- Medical Visual Question Answering
- DAQUAR

Training
- Finetuning?
- Lora?

- ResNet50 + BERT
- PubMedCLIP
- Qwen2-VL
- ViLT

Evaluation

Training Env:
- Google Colab T4 + 15G GPU

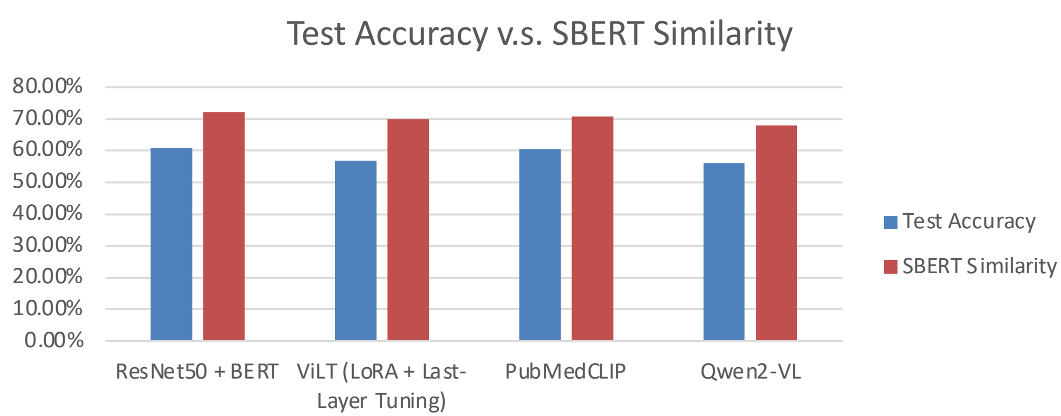
Training Model

	ResNet50 + BERT	PubMedCLIP	ViLT	Qwen2-VL
Model Type	Classification	Classification	Classification (VQA Pretrained)	Generative VQA
Image Encoder	ResNet50 (CNN)	PubMedCLIP ViT-base	Vision Transformer inside ViLT (no CNN)	Vision Encoder inside Qwen2-VL
Text Encoder	BERT-base	BERT-base	Transformer text encoder	LLM (Qwen2-2B)
Size	Small (~135M)	Medium (~150M)	~87M (ViLT-base)	Large (~2.2B)
Image Resolution (Resize)	224 x 224	224×224	224×224	448×448
Training Strategy	Fine-tuning	Fine-tuning	Parameter-Efficient Fine-Tuning (LoRA + last layer)	Fine-tuning + Lora
Strength	Stable, best on small datasets	Strong VQA inductive bias, efficient fine-tuning	Domain-specific vision encoder	Best for reasoning & open-ended responses
Weak	Not domain-specific	Overfits small datasets	Less strong on medical images	Slow, harder to adapt to classification

Training Result – Medical Visual Data

	ResNet50 + BERT	PubMedCLIP	ViLT	Qwen2-VL
Test Accuracy	60.89%	60.44%	56.89%	56%
SBERT Similarity (paraphrase-MiniLM-L6-v2)	72.22%	70.78%	69.87%	67.94%
Training Time	1 hour	2 hours	2 hours	>5 hours

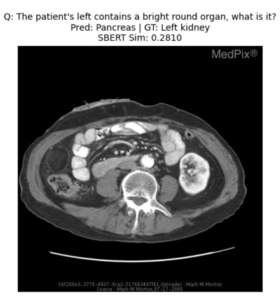
- **Higher SBERT similarity**, meaning its answers are often semantically correct even if not textually identical.
- **Qwen2-VL: Much longer training/inference time** due to image encoder + LLM decoding.



Insight

- Domain-specific pretraining matters more than model size
  - ResNet50 + BERT, the smallest and simplest model, **performed best**.
  - Old and smaller model is better than new or large model.

- Classical supervised CNNs still excel in structured tasks
  - Medical images have **strong low-level features (edges, patterns, textures)**, which CNNs extract effectively.
  - Transformer vision models require **much larger training sets**

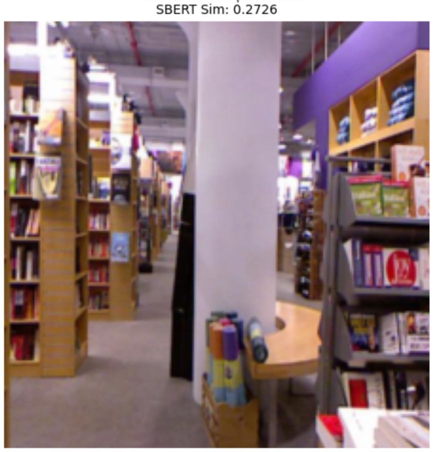


Training Result – DAQUAR

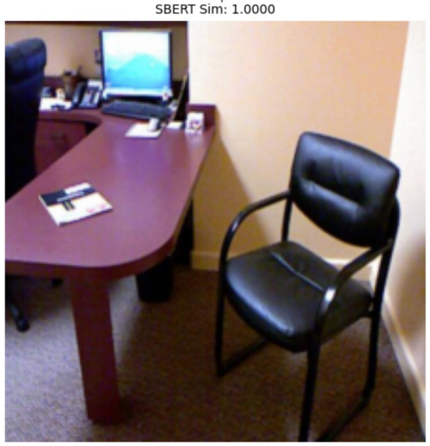
- Questions are more diverse, and answers vary widely.
- The models are not pretrained on natural indoor images, so domain mismatch is large.

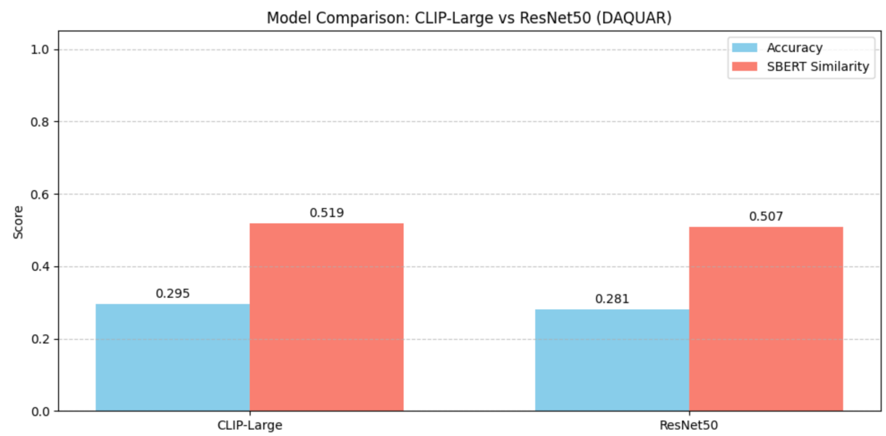
	ResNet50 + BERT	ViLT
Test Accuracy	28.1%	29.5%
SBERT Similarity	50.7%	51.9%
Training Time	1 hour	2 hours

Q: what is on the right side of the pillar
Pred: bookshelf | GT: table
SBERT Sim: 0.2726



Q: what is on the right side of the table
Pred: chair | GT: chair
SBERT Sim: 1.0000





Future Work

- Improve Image Resolution
- Explore Other Parameters or Techniques for all models
- Introduce Other Vision–Language Modules or medical-pretrained multimodal models
- User more powerful hardware to train

Thanks You

Insight

- Generative VQA models (Qwen2-VL) are flexible but less accurate
  - **Lower accuracy:** generative answers are harder to align with exact-word labels.
  - **Higher SBERT similarity**, meaning its answers are often semantically correct even if not textually identical.
  - **Much longer training/inference time** due to image encoder + LLM decoding.

- DAQUAR is a far more difficult dataset
  - DAQUAR images (indoor scenes) require high-level reasoning and object interactions.
  - Questions are more diverse, and answers vary widely.
  - The models are not pretrained on natural indoor images, so domain mismatch is large.

- LoRA improves efficiency but cannot fully overcome architectural limitations.
  - But still did not surpass ResNet50 or PubMedCLIP
  - This shows that visual language model architecture is inherently weak for medical images, where local patterns matter more than global attention.