Adapting Vision-Language Models for Visual Question Answering Using Medical Procedure Videos

Agrim Joshi[aja121], Archita Srivastava[asa410], Saarang Anand[saa183], Van Nguyen[vtn7], and William Nguyen[wna8]

{aja121,asa410,saa183,vtn7,wna8}@sfu.ca
CMPT419 Spring 2025, Group 16, Prof. Hamarneh

Abstract. Egocentric computer vision systems show transformative potential for medical training and real-time decision support, yet remain underexplored in life-saving interventions. This work presents a comprehensive framework for visual question answering (VQA) in trauma procedures through the Trauma THOMPSON Challenge, introducing the first multimodal dataset of egocentric life-saving intervention (LSI) recordings with expert annotations. Our data set comprises of 100+ procedural videos that capture various environments, simulators, and intervention types, each annotated with 13 clinically critical multiple choice questions per frame through a scene-aware sampling strategy.

We establish a benchmark evaluation of vision-language models (VLMs) through three key contributions: (1) Systematic comparison of CLIP, BLIP, and ViLT architectures across zero-shot and fine-tuned paradigms, (2) Novel prompt engineering strategies optimizing medical domain adaptation, and (3) Quantitative demonstration that task-specific fine-tuning improves accuracy over zero-shot approaches. The framework processes raw GoPro footage through automated frame sampling, multimodal feature fusion, and tailored constrained answer prediction for clinical reasoning.

Experimental results reveal that context-sensitive prompt engineering improves zero-shot performance over basic prompts, while supervised fine-tuning yields higher accuracy for procedural questions such as catheter placement. This work establishes a foundation for robust and scalable medical VQA systems in surgical practice and emergency interventions, with future directions focused on open-ended question answering and implementation of more powerful models.

Keywords: Visual Question Answering · Action Recognition · Medical AI · Vision-Language Model · Zero-Shot Classification · Medical Computer Vision

1 Introduction

Automatic action recognition in medicine is rapidly evolving and has significant potential to improve training, evaluation, and real-time decision making. The

Trauma THOMPSON program [1] highlights the value of AI in helping medics under constrained conditions. Building on this, the Trauma THOMPSON Challenge (`https://thompson-challenge.grand-challenge.org/`) offers an egocentric video dataset of lifesaving procedures, allowing research in visual understanding from a first-person perspective. This dataset is pivotal for the development of models that can interpret medical actions effectively.

One of the key tasks in this challenge is Visual Question Answering (VQA), which aims to develop models capable of answering clinically relevant questions based on video data. An effective VQA system could assist medical professionals by providing real-time guidance, improving training simulations, and enhancing post-procedure analysis. The ability to automatically answer visual questions could reduce cognitive load during emergencies, support novice clinicians, and standardize feedback in training environments, ultimately contributing to safer and more efficient healthcare delivery. Given its potential impact, our team chose to focus on addressing the VQA task in this project.

To address this task, we utilized pre-trained models and fine-tuned them for the VQA setting using surgical video data. This strategy aligns with the successful approaches used in Task 2 of the Trauma THOMPSON Challenge, such as QUIILT3 [2], which demonstrated the effectiveness of fine-tuning Deep Modular Co-attention Networks (MCAN) [3] and incorporating GloVe pre-trained word embeddings [4] for enhanced language understanding.

The remainder of this report is structured as follows. Section 2 introduces the dataset provided to us by the Trauma THOMPSON Challenge. Section 3 describes our methodology, including data preprocessing, model selection, fine-tuning strategies, and evaluation metrics. Section 4 presents the results, assessing the effectiveness of our approach. Section 5 discusses our accomplishments, while Section 6 outlines each members' contributions. Section 7 provides the conclusion and discussion, and Section 8 explores directions for future work.

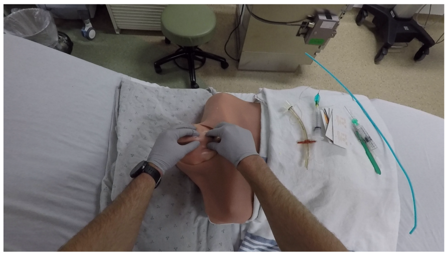
2 Materials

The Trauma THOMPSON Challenge provides both training and testing datasets, consisting of 125 videos for training and 71 for testing. For this project, we utilize the training set exclusively, as it includes ground truth labels required for supervised learning.

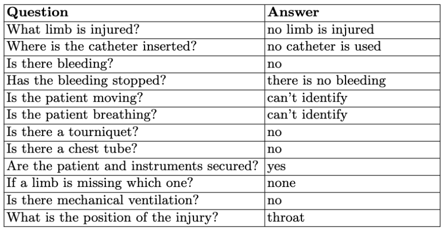
The training set for the VQA task consists of 125 videos, each paired with a corresponding .json file, resulting in 125 annotated video samples. Each video captures a life-saving procedure performed on a mannequin from various camera angles (see Fig. 1). In total, the dataset contains 380,403 frames, averaging 3,043 frames per video, with a range from 809 to 19,583 frames and a standard deviation of 2,482.55. The accompanying JSON files provide frame-level questions and their corresponding ground truth answers (refer to Appendix Fig. 5 for exact distribution).

While the dataset includes a wide variety of questions and answers, each question is limited to a predefined set of possible responses. Table 1 illustrates the types of questions which primarily focus on the patient's condition or the actions being performed.

For our project, we sample 20 frames from 100 selected videos, resulting in a total of 20,000 frames. This subset is divided into training (60%), validation (20%), and testing (20%) sets. This stratified split enables effective model training, hyperparameter tuning, and final evaluation.



Sampled Frame from a Video



Question	**Answer**
What limb is injured?	no limb is injured
Where is the catheter inserted?	no catheter is used
Is there bleeding?	no
Has the bleeding stopped?	there is no bleeding
Is the patient moving?	can't identify
Is the patient breathing?	can't identify
Is there a tourniquet?	no
Is there a chest tube?	no
Are the patient and instruments secured?	yes
If a limb is missing which one?	none
Is there mechanical ventilation?	no
What is the position of the injury?	throat

Ground Truth for the Frame

Fig. 1: Example frame from a medical procedure video (left) and its associated ground truth labels (right).

3 Methodology

3.1 Data Collection and Processing: Frame Sampling

To enable Visual Question Answering (VQA) on procedural videos, we implemented an approach to extract representative frames. Our approach supports both remote and local video processing. For remote videos, we download files into a temporary directory for seamless analysis, whereas for local videos, we directly process files from a specified folder.

In order to accomplish a representative sampling approach, first, we leveraged ffmpeg's scene detection filter (`scdet`) to automatically identify significant scene changes by recording the precise timestamps of these transitions. When the number of detected scene changes was below our desired frame count, we supplemented our selection with randomly sampled frames to ensure a complete and consistent dataset.

In addition to this, we utilized (`ffprobe`) to extract video metadata—such as frames per second and total frame count—which enabled us to accurately convert scene timestamps into corresponding frame indices. This careful mapping ensured that the sampled frames were evenly distributed throughout each video, capturing all key transitional moments.

Finally, we organized the extracted frames into subdirectories named after their source videos, with filenames incorporating both the video identifier and the frame index, thereby streamlining our downstream analysis for VQA model inference and training.

Table 1: Possible questions and answer combinations

Question	**Possible Answers**
What limb is injured?	no limb is injured, left leg, left arm, right leg, right arm
Is the patient intubated?	can't identify, yes, no
Where is the catheter inserted?	no catheter is used, lower limb
Is there bleeding?	no, yes
Has the bleeding stopped?	there is no bleeding, no, yes
Is the patient moving?	can't identify, yes, no
Is the patient breathing?	can't identify, yes, no
Is there a tourniquet?	no, yes
Is there a chest tube?	no, yes
Are the patient and instruments secured?	can't identify, yes, no
If a limb is missing which one?	none, left arm, left leg, right leg
Is there mechanical ventilation?	can't identify, yes, no
What is the position of the injury?	thorax, throat, can't identify, lower limb, abdomen, upper limb

3.2 Models

Our methodology employs a hybrid approach combining zero-shot inference and supervised fine-tuning of vision-language models (VLMs). We first evaluate three established architectures from HuggingFace - **CLIP [5], BLIP [6], and ViLT [7]**, using their base implementations without domain adaptation. These models were selected based on three criteria: 1) Architectural compatibility with medical VQA's dual-modality requirements, 2) Computational efficiency for resource-constrained deployment scenarios, and 3) Proven efficacy in generalized vision-language tasks as documented in recent literature.

3.3 Zero-shot Prompts

To assess the zero-shot capabilities of our selected vision-language models (`CLIP`, `BLIP, and ViLT`), we designed a systematic evaluation framework using two distinct prompt sets. These prompts were carefully constructed to test how different levels of contextual framing and answer space constraints affect model performance in medical VQA tasks.

The evaluation employed five prompt variations within each set, progressively increasing in contextual information and specificity. These prompt types included: *Base Question* (minimal input), *Question + Answer Options*, *Contextualized Medical Scenario*, *Instruction-based Analysis*, and *Domain-Specific Perspective*.

The two prompt sets were designed to reflect varying degrees of clinical context. **Set 1** represents a minimal clinical framing with direct question-answer formats, whereas **Set 2** incorporates enhanced medical context, leveraging simulation-aware instructions and richer visual scene descriptions. See **Table 2** under the Appendix.

3.4 Evaluation

To assess model performance on the multi-class Visual Question Answering (VQA) task, we evaluated both zero-shot and fine-tuned models using standard classification metrics for each question type:

- **Accuracy**: Proportion of correctly predicted answers.
- **Precision (Weighted)**: Class-weighted average of correctly predicted positive observations.
- **Recall (Weighted)**: Class-weighted average of correctly identified true positives.
- **F1 Score (Weighted)**: Harmonic mean of weighted precision and recall.
- **AUC (Multiclass, OvR)**: Area Under the Curve using one-vs-rest multiclass strategy.

All labels were normalized (lowercased and stripped of whitespace), then encoded using `LabelEncoder`. For AUC, we applied `label_binarize` to convert categorical outputs into binary format per class.

4 Results

This section presents a comparative evaluation of zero-shot and fine-tuned approaches for the Visual Question Answering (VQA) task using the Trauma THOMPSON dataset. The task is formulated as a multi-class classification problem, where each of the 13 medically relevant questions is associated with a constrained set of predefined answer choices.

We evaluate three state-of-the-art Vision-Language Models — `CLIP`, `BLIP` and `ViLT` — across two evaluation paradigms:

- **Zero-shot evaluation** (9 settings): Inference is performed without any model training, using image inputs and a variety of natural language prompts drawn from two structured sets — Set 1 (minimal prompts) and Set 2 (contextualized prompts).
- **Fine-tuned evaluation** (1 setting): Selected models are trained using image and question-only inputs, with cross-entropy loss applied over the discrete set of answer classes.

Model behavior is analyzed using exploratory data analysis (EDA) of the predictions across all question types. The results highlight the performance gap between general-purpose zero-shot inference and domain-adapted fine-tuning for medical simulation understanding.

4.1 Zero-shot Evaluation

In the zero-shot evaluation setting, model predictions were generated using inference over a constrained set of predefined answers, without any task-specific fine-tuning. The full results of the zero-shot evaluation are presented in Figure 2a and Figure 2b, where accuracy heatmaps illustrate model performance across all five prompt types.

Binary classification questions, such as *"Is there bleeding?"* and *"Is the patient intubated?"*, consistently achieved high accuracy and weighted F1 scores across all models. This can be attributed to the limited answer space and relatively unambiguous visual cues associated with these tasks, making them more amenable to inference in both zero-shot and supervised settings.

In contrast, multi-class or spatial reasoning questions—including *"What limb is injured?"* and *"What is the position of the injury?"*—demonstrated lower performance across all models. These tasks typically involve greater label entropy and visual complexity, requiring more nuanced spatial understanding and precise localization. Such capabilities are generally underdeveloped in models not fine-tuned on specialized medical data.

Notably, Prompt Set 2, which incorporates richer clinical context, yielded a performance boost for `CLIP`, suggesting that it benefits from enhanced prompt engineering. However, this effect was less pronounced for `BLIP` and `ViLT`, which exhibited more stable performance across both prompt sets. This observation suggests that some models may inherently encode contextual robustness, while others are more reliant on prompt phrasing.

4.2 Fine-tuning Evaluation

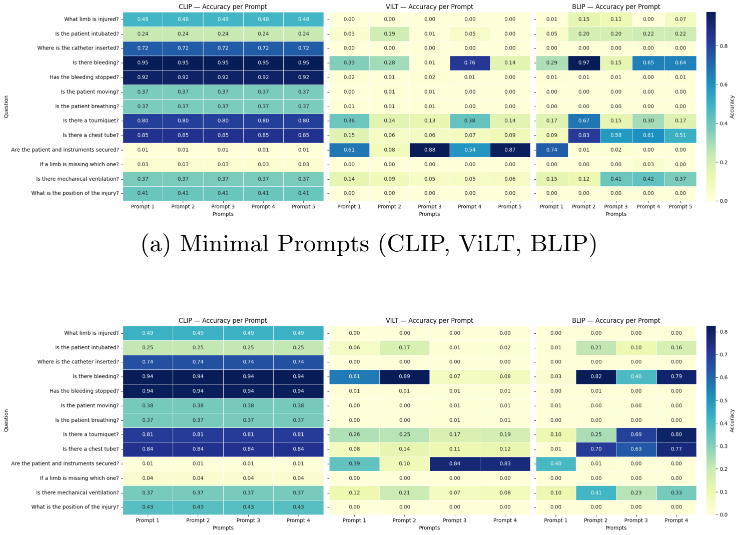
We fine-tuned `CLIP` and `ViLT` using 20,000 image-question pairs extracted from 100 videos in the training set. The data was split as follows: 60% for training, 20% for validation, and 20% for testing.

Each model was trained as a supervised multi-class classifier, predicting a label from predefined answer options for each question type. Training `BLIP` proved to be significantly more time-intensive and due to limited computational resources, we deferred this model to future work.

- `CLIP` achieved its highest validation F1 score of **0.9464** at epoch 10, see Fig. 3a.
- `ViLT` converged faster, peaking at epoch 6 with a validation F1 score of **0.9926**, see Fig. 3b.

Both models showed marked improvements over their zero-shot baselines, especially for procedurally grounded questions such as catheter placement and injury localization.

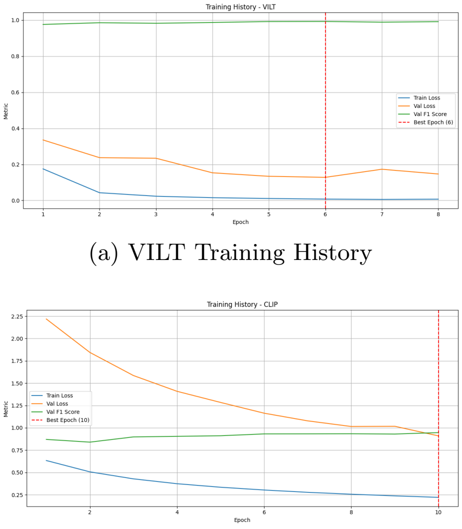
As per the heatmap visualization (Fig. 4), `ViLT` outperforms `CLIP` across all key evaluation metrics—accuracy, precision, recall, F1-score, and AUC—demonstrating fewer misclassifications and more robust overall performance. Additionally, the accuracy delta plot (Fig. 6) illustrates a consistent advantage for



(a) Minimal Prompts (CLIP, ViLT, BLIP)

(b) Contextual Prompts (CLIP, ViLT, BLIP)

Fig. 2: Accuracy heatmaps comparing Vision-Language Models across 2 prompt types.



(a) VILT Training History

(b) CLIP Training History

Fig. 3: Training loss and validation F1 trends for VILT and CLIP over epochs.

ViLT over CLIP across nearly all question types, with positive accuracy margins further validating its effectiveness. ViLT's rapid convergence and superior evaluation scores make it the preferred model for real-time decision support in our procedural VQA system. Future work will involve exploring alternative model architectures (such as BLIP) and expanding the dataset to enhance generalization and model robustness.

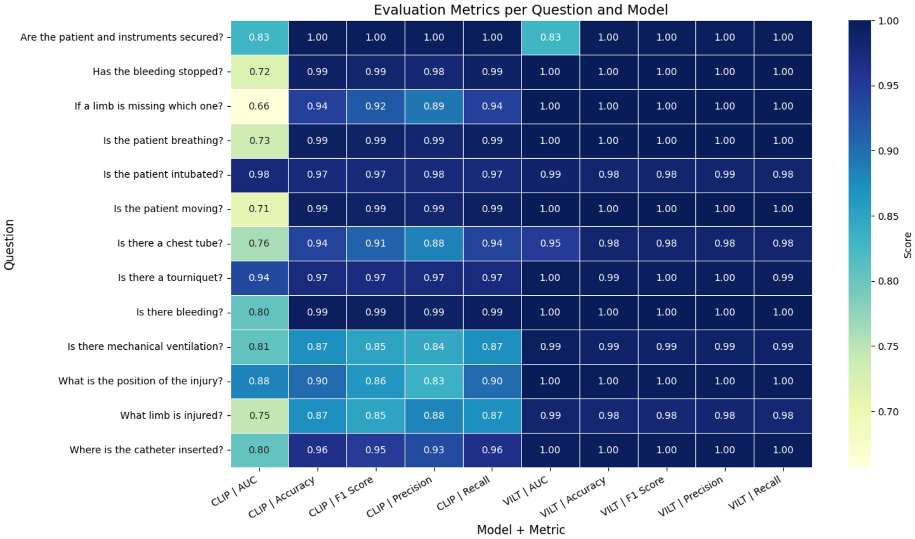


Fig. 4: Evaluation Metrics - CLIP and ViLT

5 Accomplishments

Throughout the course of this project, we gained hands-on experience in adapting large-scale Vision-Language Models (CLIP, ViLT, BLIP) to the specialized domain of medical visual question answering. We developed a full VQA pipeline encompassing custom frame sampling, prompt design, and model evaluation. Our experiments underscored how model outputs are influenced by both prompt structure and domain-specific nuances. A significant challenge encountered was the BLIP model failing during fine-tuning due to computational limitations, restricting its evaluation to zero-shot settings. Nevertheless, the fine-tuned versions of CLIP and ViLT demonstrated strong performance, highlighting the potential of general-purpose VLMs to be effectively tailored for specialized medical tasks.

6 Contributions

- **Agrim Joshi**: Contributed to the report, Created the presentation and Demo Video. Contributed to initial stages of Zero shot testing teaming up with Van. Worked out a way to mount data on Colab.
- **Archita Srivastava**: Retrieved and organized the dataset locally and on shared storage. Conducted exploratory data analysis to understand question classes and created visualizations of class distributions. Tested the inference capabilities of Video Language Models (LLaVa) and VQA Models (mini-cpm) for preliminary testing phases. Contributed to the selection of CLIP, ViLT, and BLIP as the core VQA models. Co-developed and curated prompt strategies. Implemented and executed scripts for zero-shot inference, fine-tuning, and evaluation across all models. Analyzed model outputs through Jupyter notebooks and created plots to compare performance. Contributed to the README documentation, edited the report, and assisted in preparing the project presentation video.
- **Saarang Anand**: Wrote the frame sampling script used to prepare training dataset. Designed context-based prompt templates. Wrote and formatted some sections of the report. Contributed to initial image-to-text model testing, teaming up with William. Executed zero-shot and fine-tuning of BLIP. Documented project progress.
- **Van Nguyen**: Collaborated with Agrim to test a pre-trained model on zero-shot tasks. Wrote the report.
- **William Nguyen**: Created the narration script for the video presentation. Assisted Agrim in designing the slides and editing the video. Wrote the evaluation script used to analyze model performance, and contributed to the initial stages of zero shot testing and model evaluation.

7 Conclusion and Discussions

In this project, we addressed the task of Visual Question Answering (VQA) on a specialized medical dataset by leveraging pre-trained vision-language models from Hugging Face, including BLIP, ViLT, and CLIP.

Our initial approach treated the task as an image-to-text generation problem using models like LLaVa. However, after discussions with the teaching team and further evaluation of task requirements, we transitioned to a more structured classification-based VQA framework, which better aligned with the discrete nature of the medical questions.

While we successfully established a working system and demonstrated the feasibility of adapting general-purpose models to a highly specialized domain, several challenges emerged. However, limited and imbalanced data posed generalization challenges, and fine-tuning large models was constrained by computational resources.

Critically, our findings showed that ViLT consistently outperformed CLIP across accuracy, F1, and AUC metrics, as visualized in both the evaluation heatmaps and the accuracy delta plots. However, performance still varied significantly by question type, revealing that current models may struggle with visually ambiguous or sparsely represented concepts. This suggests a need for more nuanced evaluation methods and targeted architectural improvements.

8 Future Work

We initially aimed to fine-tune the `BLIP` model for our Visual Question Answering (VQA) task; however, its significantly longer training times—compared to `ViLT` and `CLIP`—prevented us from completing this within the project timeframe. As such, our evaluation of `BLIP` was limited to zero-shot inference. Given `BLIP`'s strong performance on multimodal benchmarks, future work will revisit this model for fine-tuning using domain-specific data to assess its potential for further performance gains.

While our current framework focuses on a fixed set of predefined procedural questions, we envision extending the system to handle open-ended, natural language queries. This would enhance flexibility and align more closely with real-world clinical scenarios.

In addition, we propose incorporating domain-specific pretraining using medical imaging datasets (e.g., MIMIC-CXR or surgical video frames), refining prompt engineering to inject clinically relevant context, and exploring lightweight model variants for real-time deployment. Integrating explainability techniques (e.g., Grad-CAM, attention visualizations) would also improve transparency and trustworthiness—critical requirements for adoption in high-stakes medical environments.

Overall, this work lays the groundwork for more robust and scalable medical VQA systems, and highlights key pathways for future research in both methodology and clinical applicability.

Acknowledgements

We would like to thank the organizers of the Trauma THOMPSON Challenge for providing the videos along with corresponding question-answer annotations, which formed the foundation of this project. We are also grateful to our course instructor, Professor Ghassan Hamarneh, and the teaching assistants for their valuable guidance, feedback, and support throughout the development of this project. Additionally, we acknowledge the use of ChatGPT, which assisted us in better understanding complex concepts and helped us troubleshoot and refine our code during the development process. We also utilized a range of pretrained models available through Hugging Face (`https://huggingface.co/`), which played a central role in model experimentation and fine-tuning.

References

1. E. Birch et al. Trauma thompson: Clinical decision support for the frontline medic. *Military Medicine*, vol. 188, no. Supplement_6, pp. 208–214, 2023. November 2023,

2. T. T. L. Vuong et al. Quiil at t3 challenge: Towards automation in life-saving intervention procedures from first-person view. 2024,
3. Z. Yu et al. Deep modular co-attention networks for visual question answering. 2019,
4. J. Pennington et al. GloVe: Global vectors for word representation. In A. Moschitti et al., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532–1543, Doha, Qatar, October 2014, Association for Computational Linguistics
5. A. Radford et al. Learning transferable visual models from natural language supervision. 2021,
6. J. Li et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. 2022,
7. W. Kim et al. Vilt: Vision-and-language transformer without convolution or region supervision. 2021,

Appendix

The data can be obtained by signing up for the Trauma THOMPSON Challenge at `https://thompson-challenge.grand-challenge.org/`. The codes and output can be obtained from our group project GitHub at `https://github.com/sfu-cmpt419/2025_1_project_16`. The Jupyter Notebooks used to produce the graphics in this report can be found at: `src/analysis/analyze.ipynb`, `src/eval_train.ipynb`, `src/eval_zero_shot.ipynb`.

To reproduce the result, please follow the instructions in the README.md.

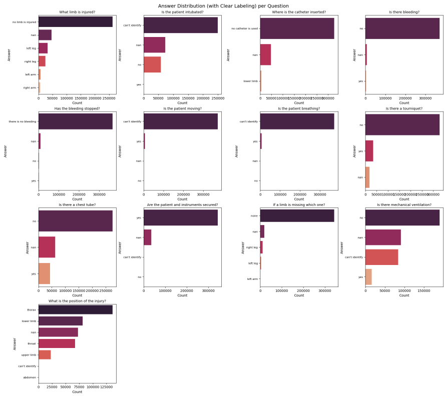


Fig. 5: Distribution of answers for each question

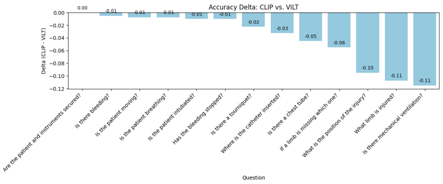


Fig. 6: Accuracy Delta

Table 2: Comparison of Zero-shot Prompt Templates: Minimal Input vs. Contextualized

Prompt Type	**Set 1: Minimal Input Prompts**	**Set 2: Contextualized Prompts**
Base Question	[Question]	Carefully examine the image and answer this medical question based solely on what is visually observable. Respond with the most likely answer based on the scene: [Question]
Question + Options	[Question]? Choose from: [Answers]. Respond with one or ’NA: Cannot be determined’.	Observe the image and answer the medical question: [Question]. Choose *only one* answer from these options: [Answers]. If no answer is visually inferable, reply ’NA: Cannot be determined’.
Contextualized Scenario	The image shows a medical simulation. [Question]? Options: [Answers]. Choose one or ’NA: Cannot be determined’.	You are analyzing an emergency trauma scene image captured from a video recorded in a high-stress environment... [Question]. Choose only one answer from the options below: [Answers]. Say ’NA: Cannot be determined’ if uncertain.
Instruction-based	Analyze this clinical simulation image. [Question]? Select from: [Answers]. or say ’NA: Cannot be determined’.	You are a medical AI assistant helping triage trauma patients... Analyze the image and answer: [Question]. Choose from: [Answers]. Say ’NA: Cannot be determined’ if needed.
Domain Perspective	[Question]? From the egocentric medical view, answer: [Answers]. or say ’NA: Cannot be determined’.	—