.

IReL, IIT(BHU) at MEDIQA-MAGIC 2025: Tackling Multimodal Dermatology with CLIPSeg-Based Segmentation and BERT-Swin Question Answering

Notebook for ImageCLEFmedical at CLEF 2025

Krishna Tewari[1,*,†], Abhyudaya Verma[1,†] and Sukomal Pal[1,†]

[1]*Indian Institute of Technology(BHU) Varanasi, India*

Abstract

Advances in multimodal learning have the potential to significantly improve automated analysis of dermatological images by integrating visual and textual clinical information. In this work, we present IReL, IIT(BHU)'s system developed for the MEDIQA-MAGIC 2025 challenge, addressing two tasks: lesion segmentation and CVQA. For segmentation, we propose a CLIPSeg-based framework that combines clinical images with contextual prompts formed by consumer questions and clinician responses. Using frozen CLIP encoders and a fine-tuned transformer decoder, our system produces detailed lesion masks being among top performing team by achieving a Dice score of 0.741 and a Jaccard score of 0.588. These results demonstrate the effectiveness of prompt-guided vision-language models in generating clinically meaningful segmentation outputs. In the VQA task, we integrate Bio_ClinicalBERT and a Swin Transformer to encode textual and visual inputs, respectively. While the model underperformed (accuracy 0.1731), likely due to suboptimal input alignment, it establishes a foundation for future enhancements. Our findings underscore the strength of vision-language fusion for dermatological segmentation and indicate that targeted improvements in multimodal alignment and input formatting could substantially improve VQA performance. Overall, this work highlights the promise of multimodal architectures in advancing intelligent clinical decision support.

Keywords

Segmentation, Multimodal Dermatology, Visual Question Answering, MEDIQA-MAGIC 2025, ClipSeg

1. Introduction

Dermatological disorders constitute a substantial portion of global disease burden, with skin conditions affecting nearly one in three individuals at some point in their lifetime [1]. Accurate diagnosis of these conditions often depends on a combination of visual inspection and clinical context, posing a multimodal challenge in healthcare. With the proliferation of patient-generated dermatology images and natural language interactions via telehealth platforms, there is an urgent need for intelligent systems capable of jointly analyzing textual and visual data.

To address this gap, the MEDIQA-MAGIC 2025 challenge [1] introduced a two-pronged benchmark for multimodal dermatology. The first subtask focuses on **lesion segmentation**, where the objective is to generate dense pixel-wise masks of dermatological anomalies using both images and contextual prompts. The second subtask addresses **closed-domain visual question answering (CVQA)**, requiring models to select the correct answer from multiple choices given a medical image and a related clinical question.

For Subtask 1, we utilize CLIPSeg [2], a vision-language segmentation model that integrates textual prompts with visual features through a frozen CLIP backbone and a transformer decoder. By

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

krishnatewari.rs.cse24@itbhu.ac.in (K. Tewari); abhyudaya.student.cse21@itbhu.ac.in (A. Verma); spal.cse@itbhu.ac.in (S. Pal)

0009-0005-6599-9956 (K. Tewari); 0000-0001-8743-9830 (S. Pal)

[1]https://www.who.int/news-room/fact-sheets/detail/skin-conditions

incorporating contextual cues formed by concatenating patient questions and clinician responses, our system is able to produce clinically meaningful lesion masks. For Subtask 2, we employ a dual-encoder architecture using Bio_ClinicalBERT [3] for textual encoding and Swin Transformer [4] for image representation. These encoders project their respective modalities into a shared embedding space to support answer classification.

The rest of the paper is structured as follows. Section 2 provides a review of related work. Section 3 introduces the datasets. Section 4 elaborates on our implementation strategies. Section 5 discusses the results obtained. Finally, Section 6 offers concluding insights and suggests directions for future work.

2. Related Work

Multimodal learning has emerged as a promising direction in dermatological AI by enabling the integration of visual and textual data for context-aware diagnosis and segmentation. While early efforts in medical visual question answering (VQA) focused on radiology and pathology [5, 6], recent work has adapted these frameworks to dermatology, addressing its unique visual and semantic challenges.

In lesion segmentation, the ISIC (International Skin Imaging Collaboration) challenges [7] have been instrumental in shaping benchmarks for melanoma detection. These competitions drove the adoption of convolutional models like U-Net [8], which remains a cornerstone in medical image segmentation due to its skip connections and encoder-decoder design that retain spatial resolution. However, such purely visual models often fail to integrate clinical context, limiting their diagnostic interpretability.

To overcome this limitation, multimodal transformers such as LXMERT [9] and ViLT [10] have been applied to align visual and textual inputs. These models support joint reasoning tasks like captioning and VQA, but their general-domain pretraining constrains their effectiveness in clinical settings. Domain-specific adaptations, such as BioViL-T [11], have improved upon this by fine-tuning with biomedical corpora, enhancing performance in tasks like image-text retrieval and clinical reporting.

Simultaneously, synthetic data generation has become vital in dermatology to address limitations in dataset availability and privacy. Generative models like StyleGAN2 [12] can synthesize realistic lesion images, though their fidelity depends heavily on careful tuning. More recent approaches [13] condition image generation on clinical text prompts, ensuring better semantic relevance and diagnostic utility.

Despite recent progress, most approaches treat lesion segmentation and VQA as separate tasks, optimizing each in isolation. While unified multimodal systems are a promising goal, addressing these tasks independently allows for specialized architectures and task-specific tuning. In this work, we adopt separate transformer-based models for segmentation and VQA, enabling targeted advancements in each area while contributing to the development of robust and interpretable dermatological AI solutions.

3. Dataset Details

The MEDIQA-MAGIC 2025 dataset [14] includes thousands of high-resolution clinical dermatology images collected in real-world settings. Each encounter contains between 1 to 5 RGB images depicting various skin lesions under natural lighting conditions. These images are captured from different angles and are often consumer-generated, mimicking real clinical workflows. For each image, expert dermatologists provided pixel-wise lesion segmentation masks. Up to three masks may be available per image, created by independent annotators from a pool of four. The masks follow a standardized file naming convention and are stored in TIFF format. Each clinical encounter is associated with one or more templated multiple-choice questions.

4. Methodology

In this section, we present the methodology and model architecture employed for both subtasks.

4.1. Segmentation (Subtask 1)

This section details our vision-language pipeline for identifying dermatological lesion regions in clinical images. As summarised in Figure 1, the workflow combines textual prompts with image features through a pre-trained CLIP backbone and a fine-tuned CLIPSeg decoder to yield pixel-accurate lesion masks.

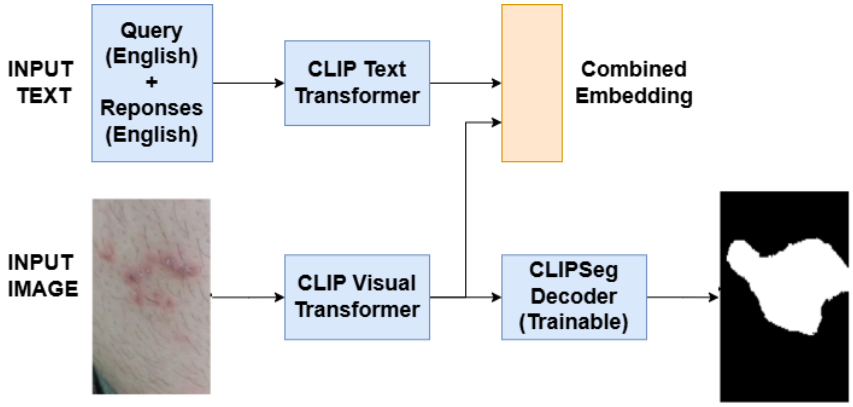

Figure 1: Step-by-step overview of the segmentation pipeline utilizing a vision-language model (CLIPSeg) to detect skin lesion regions in dermatological images.

4.1.1. Data Pre-processing and Prompt Construction

We used the *DermaVQA* corpus [14], where every clinical image is accompanied by up to three segmentation masks, each created by a dermatologist chosen from four annotators {ann0, ann1, ann2, ann3}. Masks follow the pattern:

`IMG_{ENCOUNTERID}_{IMAGEID}_mask_{ann#}.tiff`.

If a given annotator did not label an image, the corresponding file is absent. To obtain a single ground-truth mask, we perform a pixel-wise logical OR across all available annotator masks (typically three per image). Images are converted to RGB, resized to 352×352 (CLIPSeg default) and normalised with CLIP's ImageNet statistics. For the language stream, we concatenate the consumer question and clinician answer:

$$Prompt = Question \,\|\, Answer,$$

then strip HTML tags and excessive whitespace.

4.1.2. Tokenisation and CLIPSegProcessor

Text prompts are tokenized using CLIP's tokenizer, producing token IDs and an attention mask to distinguish real tokens from padding. Simultaneously, input images are resized and normalized by the vision processor to fit CLIP's vision transformer requirements. The Hugging Face CLIPSegProcessor combines these steps, outputting a dictionary with `input_ids`, `attention_mask`, and `pixel_values`, all properly padded, truncated, and normalized for seamless model input.

4.1.3. Embeddings Extraction and Model Architecture

We adopt the CLIPSeg framework [2], which builds on CLIP's dual-encoder architecture. The frozen text transformer encodes the input prompt into a fixed-length global text embedding that captures its semantic meaning. Simultaneously, the frozen vision transformer divides the image into patches, encoding each into feature vectors that retain spatial information.

To enable multimodal reasoning, both embeddings are projected into a shared latent space via learned linear layers. The global text embedding is broadcast and concatenated with each image patch embedding, forming a multimodal token sequence. This fused representation, combining semantic and spatial cues, is passed to a lightweight transformer decoder with about 6.5 million trainable parameters.

The decoder produces a dense per-pixel logits map, indicating each pixel's probability of belonging to the prompt-specified target region. During training and inference, the CLIP backbone remains frozen to retain pretrained knowledge, while only the decoder is fine-tuned, ensuring computational efficiency and reducing overfitting risks.

4.1.4. Decoder Fine-tuning and Training

The output of the CLIPSeg decoder is a single-channel logits map of the same spatial dimensions as the input image (i.e., 352×352). To convert these raw logits into interpretable probabilities, we apply a sigmoid activation function at each pixel location:

$$p_i = \sigma(z_i) = \frac{1}{1 + e^{-z_i}},$$

where z_i is the raw logit value at pixel i, and $p_i \in (0, 1)$ represents the model's confidence that pixel i belongs to the lesion region.

For supervision, we use the Binary Cross-Entropy (BCE) loss, a standard choice for binary segmentation tasks. BCE compares the predicted probability map against the binary ground-truth mask on a pixel-by-pixel basis. The loss is computed as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(x_i) + (1 - y_i) \log(1 - x_i) \right],$$

where $y_i \in \{0, 1\}$ is the ground-truth label for pixel i, x_i is the predicted probability, and N is the total number of pixels. This formulation penalises incorrect predictions more heavily when the model is confident, helping to stabilise learning and accelerate convergence.

Optimization is performed using the AdamW optimizer and we use a fixed learning rate of $9e^{-4}$, chosen based on initial grid search experiments. To ensure stable gradients and avoid numerical instabilities, we apply gradient clipping with a maximum norm of 1.0. This is particularly useful when training with mixed precision, where dynamic range limitations in float16 can cause gradients to explode in rare cases. Training is conducted for 20 to 30 epochs, with early stopping based on the validation Dice score.

4.1.5. Implementation Details

The segmentation pipeline is implemented in PyTorch [15], with preprocessing and model components integrated via Hugging Face's transformers and datasets APIs. Training and evaluation are conducted on a single NVIDIA GPU with CUDA acceleration. Batches of 32 examples are sampled using a PyTorch DataLoader. Each contains a clinical dermatology image, a textual prompt, and a binary segmentation mask formed by merging annotator masks. To maximize GPU throughput, 4–8 data loading workers are used for parallelized I/O, aiding especially with larger batches.

We employ the `CIDAS/clipseg-rd64-refined` model from the Hugging Face Hub, featuring a pretrained CLIP backbone and a transformer-based segmentation decoder. The CLIP encoders are frozen while only the decoder is fine-tuned, reducing trainable parameters and improving generalization with limited dermatological data. Mixed-precision training is enabled using PyTorch AMP, storing most activations in float16 while preserving stability through selective float32 usage. This significantly improves training efficiency. For reproducibility, random seeds are fixed across NumPy, PyTorch, and CUDA, and deterministic operations are enforced.

4.2. CVQA (Subtask 2)

This section details our multimodal pipeline for the CVQA task as shown in Figure 2, where the objective is to select the correct answer from a predefined set of options given a clinical image and a corresponding question.

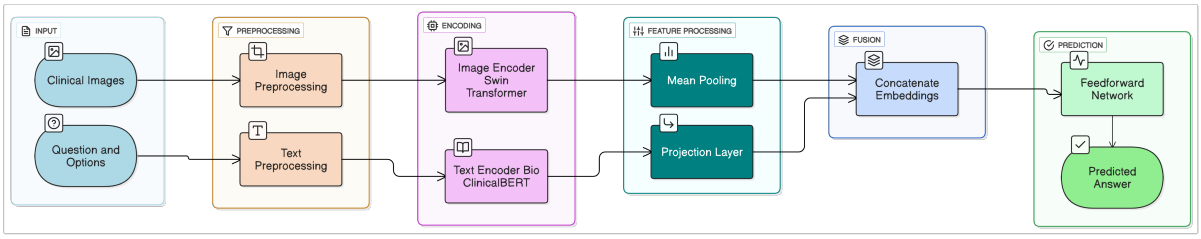

Figure 2: Step-by-step overview of the CVQA pipeline.

4.2.1. Dataset Composition and Instance Formatting

We use the *CLEF-MAGIC 2025* dataset, where each encounter consists of 1-5 clinical images and a set of templated multiple-choice questions. Each question q_i has k candidate answers $\{o_{i1}, o_{i2}, ..., o_{ik}\}$, with only one correct label.

To formulate the inputs, we first identify all images associated with an encounter, and construct paired sequences by concatenating the question with each answer option:

$$\texttt{Input}_j = q_i \,\|\, o_{ij}, \quad \forall j \in \{1, ..., k\}.$$

Each such tuple is linked to all images from the encounter. When multiple images are present, we average their embeddings (after encoding) to form a unified visual context. This strategy preserves the collective diagnostic content of the encounter while simplifying input dimensionality.

4.2.2. Preprocessing Pipeline

All image-question-option pairs are processed via Hugging Face's unified `AutoProcessor` interface, which wraps both the tokenizer and image feature extractor for compatibility with the model architecture.

Each clinical image is converted to RGB format and resized to 224×224 pixels. Normalization is applied using ImageNet mean and standard deviation statistics to match the expected input distribution of the Swin Transformer backbone. Question-option strings are tokenized using the BERT tokenizer with truncation and padding to a maximum sequence length of 128 tokens. The tokenizer outputs `input_ids` and `attention_mask`, which are used to mask padding tokens during attention computation.

4.2.3. Model Architecture

Our architecture consists of a dual-stream vision-language encoder, followed by a scoring module that computes relevance scores over candidate answers. We use Bio_ClinicalBERT [3] to encode medical question-option strings into dense representations. The final [CLS] token embedding is extracted as the global textual representation. Images are encoded using the Microsoft's SWIN [4] vision transformer, yielding patch-level embeddings which are mean-pooled to obtain a global image vector.

Each pair of textual and image embeddings is projected via separate linear layers into a shared latent space. These projected embeddings are concatenated and passed through a feedforward classification head, which outputs a scalar compatibility score:

$$s_j = \text{FFN}([\mathbf{v}_{\text{text}}; \mathbf{v}_{\text{img}}]).$$

For each question, the option with the highest score is selected as the predicted answer:

$$\hat{y} = \arg\max_{j=1}^{k} s_j.$$

4.2.4. Training Objective and Optimization

The model is trained as a k-way classifier using the standard cross-entropy loss:

$$\mathcal{L} = -\log\left(\frac{\exp(s_{j^*})}{\sum_{j=1}^{k}\exp(s_j)}\right),$$

where j^* is the index of the correct answer. Training is conducted using the AdamW optimizer with a learning rate of $5e^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01.

Each batch contains 2 question instances, where each instance includes k fused input pairs (one for each answer option). Training is performed for 10 epochs, with early stopping monitored on the validation F1-score. Dropout is applied with $p = 0.1$ after projection layers. Gradient clipping is employed with a max norm of 1.0 to ensure training stability. Mixed-precision training is enabled using PyTorch AMP for memory and computational efficiency.

All components are implemented in PyTorch using the Hugging Face Transformers and Datasets libraries. Data loading is parallelized using 4 workers. Visual and textual inputs are managed using custom `Dataset` and `DataCollator` classes. The final model is checkpointed using validation-based saving, and deterministic training is enforced via fixed random seeds.

5. Results

This section presents the experimental results for the proposed approach and the evaluation metrics used corresponding to both subtasks.

5.1. Evaluation Metrics

Performance of the segmentation task is quantitatively assessed using two commonly adopted overlap-based similarity measures: the Dice coefficient and the Jaccard index (Intersection over Union).

- **Dice Coefficient:** Also known as the F1 score for segmentation, this metric quantifies the overlap between the predicted mask and the ground truth. It is calculated as:

  $$\text{Dice} = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

  where A and B are the sets of pixels in the predicted and ground-truth masks, respectively. The Dice score ranges from 0 (no overlap) to 1 (perfect overlap) [16].
- **Jaccard Index:** Also referred to as Intersection over Union (IoU), this metric measures the proportion of shared elements between the predicted and ground-truth masks relative to their union. It is defined as:

  $$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|}$$

  with the same definitions for A and B as above. Like the Dice coefficient, the Jaccard index ranges from 0 to 1, with higher values indicating more accurate segmentations [17].

The performance of the CVQA task is evaluated using metrics that capture overall accuracy.

- **Accuracy:** This metric quantifies the proportion of correct predictions by measuring the overlap between the predicted and gold answer sets for each question. For each question instance, the intersection over maximum length (IoM) between the predicted and ground truth answer sets is computed. The final accuracy is the average of these IoM scores across all instances. Formally, for a set of n instances:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\text{Pred}_i \cap \text{Gold}_i|}{\max(|\text{Pred}_i|, |\text{Gold}_i|)}$$

where Pred_i and Gold_i denote the predicted and ground truth answer sets for the i^{th} instance, respectively.

5.2. Results

Table 1 summarizes the segmentation results across all submitted runs from different teams. The evaluation metrics include Jaccard and Dice similarity coefficients reported as mean-of-max (M(Max)) and mean-of-mean (M(Mean)) values, along with additional segmentation-specific metrics such as Jaccard, Dice.

Table 1
Segmentation Results across submissions by different teams

Team	Jacc. M(Max)	Jacc. M(Mean)	Dice M(Max)	Dice M(Mean)	Jaccard	Dice
Anastasia	0.677	0.591	0.783	0.705	0.646	0.785
Anastasia	0.677	0.591	0.783	0.705	0.646	0.785
Anastasia	0.677	0.591	0.783	0.705	0.646	0.785
Anastasia	0.631	0.550	0.742	0.666	0.611	0.759
IReL, IIT(BHU)	**0.655**	**0.569**	**0.765**	**0.686**	**0.588**	**0.741**
KLE1	0.638	0.554	0.751	0.671	0.541	0.702
KLE1	0.638	0.554	0.751	0.671	0.541	0.702
KLE1	0.638	0.554	0.751	0.671	0.541	0.702
KLE1	0.638	0.554	0.751	0.671	0.541	0.702
H3N1	0.636	0.547	0.743	0.659	0.514	0.679
H3N1	0.636	0.547	0.743	0.659	0.514	0.679
H3N1	0.636	0.547	0.743	0.659	0.514	0.679
H3N1	0.636	0.547	0.743	0.659	0.514	0.679
Anastasia	0.521	0.411	0.633	0.525	0.321	0.485
Anastasia	0.523	0.411	0.635	0.525	0.313	0.477
Kasukabe Defense Group	0.162	0.135	0.224	0.191	0.187	0.315

The M(Max) metric captures the highest similarity score obtained per sample and averages these values across all cases, reflecting the best achievable segmentation accuracy under optimal conditions. Conversely, the M(Mean) metric averages the scores over all relevant regions or slices within each sample before averaging across samples, representing the model's overall consistency and stability.

Our team, IReL, IIT(BHU), achieves a Jaccard M(Max) score of 0.655 and Dice M(Max) score of 0.765, ranking competitively among all participants. These peak performance metrics indicate our model's capability to segment lesion regions with high precision on the best-performing cases. Furthermore, the M(Mean) scores (Jaccard 0.569, Dice 0.686) demonstrate that our model maintains reliable segmentation performance consistently across the dataset.

Beyond these, the per-sample Jaccard and Dice scores for our team are 0.588 and 0.741, respectively, reflecting strong agreement between predicted and ground truth segmentations on individual evaluation points.

Comparatively, the NaiveNotNice and Anastasia teams achieved slightly higher M(Max) Jaccard scores around 0.677 and Dice scores near 0.783, indicating somewhat better peak segmentation in their

models. Overall, the high similarity scores places IReL, IIT(BHU) among the top-performing teams. This demonstrates the strength of our approach, which leveraged fine-tuning of a segmentation decoder with clinical context embedding, leading to both accurate and clinically meaningful segmentation results.

Figure 3 shows a histogram that illustrates the distribution of the Jaccard Index (Intersection over Union, IoU) computed per image on the validation dataset (see Figure X). This visualization provides insight into segmentation performance on an individual image level, revealing the variance across samples. The distribution indicates that while a substantial number of images achieve high IoU scores (above 0.75), reflecting strong segmentation accuracy, there also exists a long tail of lower-performing cases. These lower IoU values may be attributed to images with complex anatomical structures, occlusions, or limited contrast, which pose inherent challenges to the model. By analyzing this distribution, we demonstrate both the robustness of our segmentation method on a majority of cases and identify avenues for potential improvement in challenging scenarios.

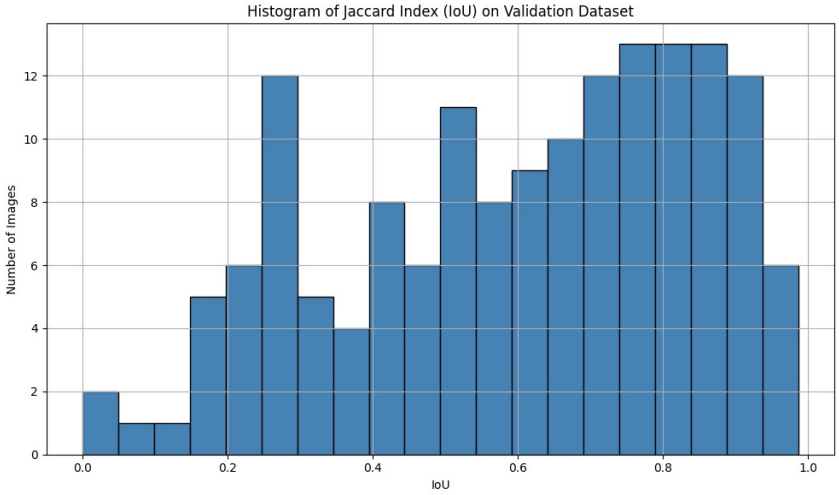

Figure 3: Histogram of per-image Jaccard Index (IoU) on the validation set, showing variability in segmentation performance.

Table 2 reports the CVQA performance of different participating teams, evaluated using the `accuracy` metric. This metric captures the model's ability to correctly select the appropriate answer from multiple options based on an input image and corresponding clinical question, reflecting joint visual-linguistic reasoning performance.

The top-performing team, Hoangwithhisfriends, achieved accuracy scores above 0.74 across multiple submissions, indicating a robust pipeline capable of extracting and reasoning over fine-grained clinical and visual cues. DS@GT MEDIQA-MAGIC and Kasukabe Defense Group followed with several strong runs, with accuracy levels ranging from 0.71 to 0.66 in their best submissions.

In contrast, our team, IReL, IIT(BHU), attained an accuracy of 0.1731, placing significantly lower than other submissions. A potential reason for this performance gap lies in the preprocessing phase, specifically the construction of question-image-option triplets. Our pipeline may not have aligned the multiple-choice options with the image-question pairs effectively, impacting the model's ability to learn discriminative patterns. Additionally, errors or misalignments in candidate option formatting could have weakened training supervision. Post-submission work will include addressing the identified limitations by refining the alignment of image-question-option triplets and improving the formatting consistency of candidate options, with the goal of enhancing training supervision and overall model performance.

Error Analysis and Limitations

QID Parent-Level Aggregation: Our current implementation evaluates each QID independently, without grouping or aggregating predictions for questions that belong to the same QID parent. Given

Table 2
CVQA performance across submissions by participating teams

Team Name	accuracy
H3N1	0.7580
H3N1	0.7507
H3N1	0.7452
H3N1	0.7452
H3N1	0.7358
DS@GT MEDIQA-MAGIC	0.7095
DS@GT MEDIQA-MAGIC	0.7062
DS@GT MEDIQA-MAGIC	0.6924
DS@GT MEDIQA-MAGIC	0.6752
DS@GT MEDIQA-MAGIC	0.6527
KLE1	0.5698
KLE1	0.5530
KLE1	0.5530
Kasukabe Defense Group	0.5366
Kasukabe Defense Group	0.5259
Kasukabe Defense Group	0.4637
DS@GT MEDIQA-MAGIC	0.3743
Oggy	0.2223
IReL, IIT(BHU)	**0.1731**

that several clinical questions in the dataset are semantically related variants, this lack of aggregation could obscure meaningful patterns or degrade accuracy. Future versions of the model will introduce:

- Aggregation of predictions across sibling QIDs under a common parent.
- Majority-vote or consensus fusion strategies for unified parent-level responses.
- Hierarchical loss functions to optimize predictions jointly at child and parent levels.

Encounter ID Alignment and Formatting Consistency: We also investigated potential issues with data alignment. Although our evaluation script confirmed matching `encounter_id` sets between ground truth and predictions, we suspect that inconsistencies in how input triplets (question, option, image) were constructed may have contributed to low performance. For example: Candidate options may not have been correctly mapped to corresponding images. To mitigate these issues, we have implemented stricter ID alignment checks and refined our preprocessing to enforce consistent formatting. These refinements are expected to enhance supervision quality and improve overall model performance in future work.

6. Conclusion and Future Work

This paper presented our approach for the MEDIQA-MAGIC 2025 challenge, focusing on lesion segmentation and visual question answering. We developed a vision-language segmentation model, aligning with clinical interpretation by integrating textual and visual data, which improved lesion identification. Although the question answering module underperformed, it highlighted challenges in multimodal alignment, guiding future work on better fusion and domain-specific prompt design. Our findings demonstrate the promise of multimodal learning to enhance automated dermatological analysis. By advancing model design and clinical adaptation, we aim to support diagnostic accuracy and hope to inspire further research in this evolving field.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. Yim, A. Ben Abacha, N. Codella, R. A. Novoa, J. Malvehy, Overview of the mediqa-magic task at imageclef 2025: Multimodal and generative telemedicine in dermatology, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Span, 2025.

[2] T. Lüddecke, A. Ecker, Image segmentation using text and image prompts, arXiv preprint arXiv:2112.10003 (2021). URL: https://arxiv.org/pdf/2112.10003. arXiv:2112.10003.

[3] E. Alsentzer, J. Murphy, W. Boag, W. H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv preprint arXiv:1904.03323 (2019).

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.

[5] A. Ben Abacha, C. Shivade, D. Demner-Fushman, Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in: CLEF (Working Notes), 2019.

[6] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, R. M. Summers, Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2016, pp. 2497–2506. URL: https://doi.org/10.1109/CVPR.2016.274. doi:10.1109/CVPR.2016.274.

[7] N. C. F. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. A. Marchetti, H. Kittler, A. Halpern, Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), arXiv preprint arXiv:1902.03368 (2019). URL: https://arxiv.org/abs/1902.03368.

[8] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015, pp. 234–241.

[9] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, in: EMNLP-IJCNLP, 2019.

[10] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), volume 139, PMLR, 2021, pp. 5583–5594.

[11] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, O. Oktay, Making the most of text semantics to improve biomedical vision–language processing, in: Computer Vision – ECCV 2022, Springer, 2022, pp. 1–21. doi:10.1007/978-3-031-20059-5_1.

[12] T. Karras, S. Laine, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110–8119.

[13] X. Liu, F. Zhao, H. Wu, Text2skin: A conditional generative model for dermatological image synthesis from clinical text, IEEE Transactions on Medical Imaging (2023).

[14] W. Yim, Y. Fu, A. Ben Abacha, M. Yetisgen, N. Codella, R. A. Novoa, J. Malvehy, Dermavqa-das: Dermatology assessment schema (das) and datasets for closed-ended question answering and segmentation in patient-generated dermatology images, CoRR (2025).

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep

learning library, in: Advances in Neural Information Processing Systems (NeurIPS), volume 32, 2019.
[16] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302.
[17] P. Jaccard, The distribution of the flora in the alpine zone, New Phytologist 11 (1912) 37–50.