DermoSegDiff and DermKEM for Comprehensive Dermatology AI

Notebook for the H3N1 Lab at CLEF 2025

Nguyen Pham Hoang Le[1,2], Hoang Pham Duc Huy[1,2], Hien Thai Dinh Nhat[1,2], Hoang Thach Minh[1,2] and Thien B. Nguyen-Tat[1,2,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

The integration of AI into dermatological diagnostics is rapidly transforming clinical practice, with crucial applications in precise lesion segmentation and intelligent Visual Question Answering (VQA). Our team, H3N1, participated in ImageCLEF MAGIC 2025, tackling both these core challenges. We proudly announce our significant achievements: a top 4 finish in the Dermatology Segmentation Task and the winner position in the Dermatology VQA Task. These results validate the power of two advanced systems. Firstly, we use **DermoSegDiff**, a system proposed by A. Bozorgpour et al. which revolutionizes skin lesion segmentation by leveraging a Denoising Diffusion Probabilistic Model (DDPM) with novel boundary detection through weighted loss and 'Boundary Attention' for unparalleled contour precision (Jaccard: 0.514, Dice: 0.679). The modified U-Net with two-path feature extraction strategy helps capture different features, therefore providing the model with a more comprehensive view. Simultaneously, our developed **DermKEM** (**Derm**atology **K**nowledge-Enhanced **E**nsemble **M**odel) system for Dermatology VQA excels with a knowledge-augmented multi-model ensemble. It employs a Genetic Algorithm for image enhancement, enriches captions via BLIP and external knowledge (Gemini 2.5 Flash), and feeds this into an ensemble of baseline models such as MUMC, and Gemini 2.5 Flash for highly accurate, contextually rich answers (Accuracy: 0.758). These synergistic systems not only demonstrate state-of-the-art capabilities but also redefine intelligent clinical support in dermatology.

Keywords

Dermatology, Image Segmentation, Visual Question Answering, Multimodal, Large Language Model, Diffusion Model

1. Introduction

The integumentary system, primarily comprising the skin, stands as the human body's largest organ and serves as a crucial interface with the external environment. Its multifaceted roles include providing a protective barrier against pathogens and physical insults, regulating body temperature, facilitating sensory perception, and contributing to immune responses. Consequently, dermatological diseases, which encompass a wide array of conditions ranging from common inflammatory disorders such as eczema and psoriasis to potentially life-threatening malignancies like melanoma, represent a significant global health burden. The accurate and timely diagnosis of these conditions is paramount for effective management and improved patient outcomes. Traditionally, dermatological diagnosis heavily relies on visual inspection by trained clinicians, often augmented by non-invasive imaging techniques like dermatoscopy. While this approach remains fundamental, its efficacy can be influenced by factors such as inter-observer variability, the subtlety of early-stage lesion characteristics, and the clinician's experience level. Furthermore, the increasing demand for accessible dermatological care, particularly in remote or underserved regions, has spurred the growth of teledermatology, where automated clinical feedback is essential. In this context, objective and reliable automated decision support systems are

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
22520982@gm.uit.edu.vn (N. P. H. Le); 22520474@gm.uit.edu.vn (H. P. D. Huy); 22520424@gm.uit.edu.vn (H. T. D. Nhat); 22520477@gm.uit.edu.vn (H. T. Minh); thienntb@uit.edu.vn (T. B. Nguyen-Tat)
0009-0007-6448-5640 (N. P. H. Le); 0009-0008-3332-9909 (H. P. D. Huy); 0009-0001-6477-4924 (H. T. D. Nhat); 0009-0004-1125-3051 (H. T. Minh); 0000-0002-4809-7126 (T. B. Nguyen-Tat)

becoming increasingly vital. Artificial intelligence (AI) has demonstrated considerable potential in augmenting diagnostic capabilities within various medical imaging domains, including dermatology. Specifically, two key areas of AI research, dermatology image segmentation and Closed Visual Question Answering (CVQA) for dermatology, are poised to revolutionize dermatological image analysis. Medical image segmentation aims to precisely delineate regions of interest from surrounding healthy tissue by generating masks that identify the affected areas. This provides quantitative data crucial for lesion characterization and monitoring. Concurrently, VQA systems, which integrate computer vision with natural language processing, enable clinicians to pose specific questions about an image (e.g., "How much of the body is affected?") and receive contextually relevant, evidence-based answers. Such systems can enhance diagnostic accuracy, streamline workflows, and facilitate more effective communication between healthcare providers and AI tools. However, the development of robust AI models relies on the availability of large-scale, high-quality, and diverse datasets, as well as rigorous benchmarking to ensure quality and diversity of data. To address these challenges and foster advancements in AI for dermatology, the ImageCLEFmed MEDIQA-MAGIC 2025 [1] challenge has been established. This initiative includes tasks directly relevant to dermatology: Task 1: Segmentation of Skin Conditions and Task 2: Generative Closed-domain Question Answering on Dermatology Images. The competition aims to stimulate the development of systems that can automatically generate clinical feedback in a teledermatology context, leveraging both visual and textual data.

Our team, H3N1, participated in both dermatological tasks. For Task 1 (Segmentation), we investigated advanced deep learning architectures, potentially exploring diffusion-based models like DermoSegDiff [2] for precise lesion boundary delineation, complemented by robust preprocessing techniques including genetic algorithm-based image enhancement. For Task 2 (Closed VQA), we introduce DermKEM, a system designed to leverage sophisticated multimodal approaches, drawing inspiration from established frameworks such as MUMC [3] for medical VQA and the capabilities of large multimodal models like Gemini [4]. These will be augmented by strategic data preprocessing including BLIP-based additional image caption generation[5] and external knowledge linking using Gemini [4]. This paper details our methodologies, experimental setup, and results for the MEDIQA-MAGIC 2025 dermatological tasks, contributing to the growing body of research on AI-driven solutions for enhanced dermatological care.

2. Related Works

Recent works about AI in medical field have increased significantly[6][7][8]; especially in the field of visual question answering (VQA), existing efforts have largely concentrated on radiological images. VQA-Med 2019[9] specifically focused on radiology images and four main categories of questions. The top-performing systems in the contest mainly employed deep learning techniques, using CNNs such as VGGNet[10] and ResNet[11] to extract visual features, and models like BERT[12] or RNNs to encode the questions. Attention mechanisms and multimodal pooling methods such as MFB and MFH were then used to fuse image and text features for answer prediction.In the MEDIQA-M3G 2024 Shared Task [13], researchers explored solutions for dermatological consumer-health visual question answering, in which user-generated queries and images serve as input, and a free-text answer is produced as output. The top performance for the English results was achieved by CLIP[14] (and its fine-tuned variant), Claude[15] with prompt-based engineering, and PMC-VQA[16] (PMC-CLIP and PMC-LLaMA).

In the medical field, segmentation is a crucial step for identifying and delineating abnormal skin regions such as lesions, malignancies, or infected areas. A widely adopted architecture for this task is U-Net[17], which features an encoder–decoder structure with skip connections that help retain detailed spatial information. U-Net and its variants, such as UNet++[18] and UNet 3+[19], have demonstrated high effectiveness in segmentation.

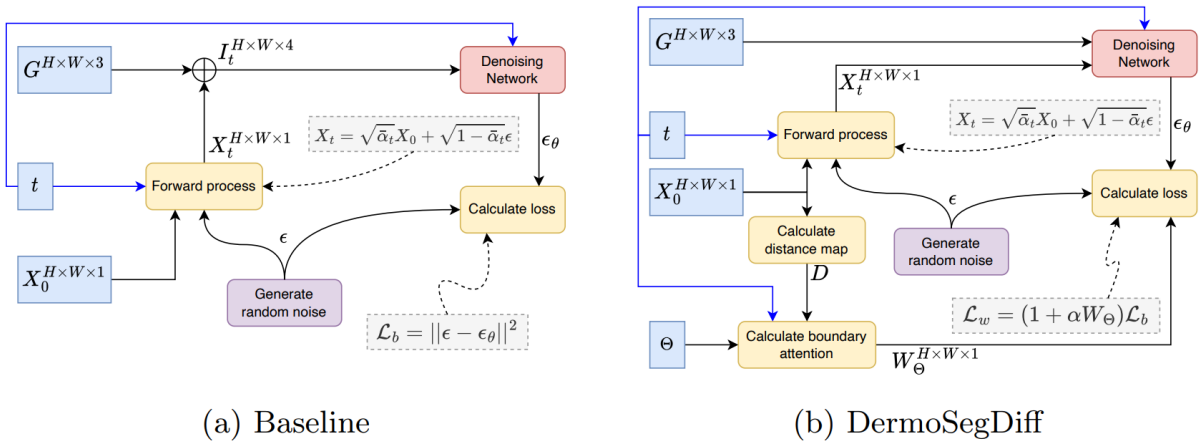

Figure 1: The DermoSegDiff architecture

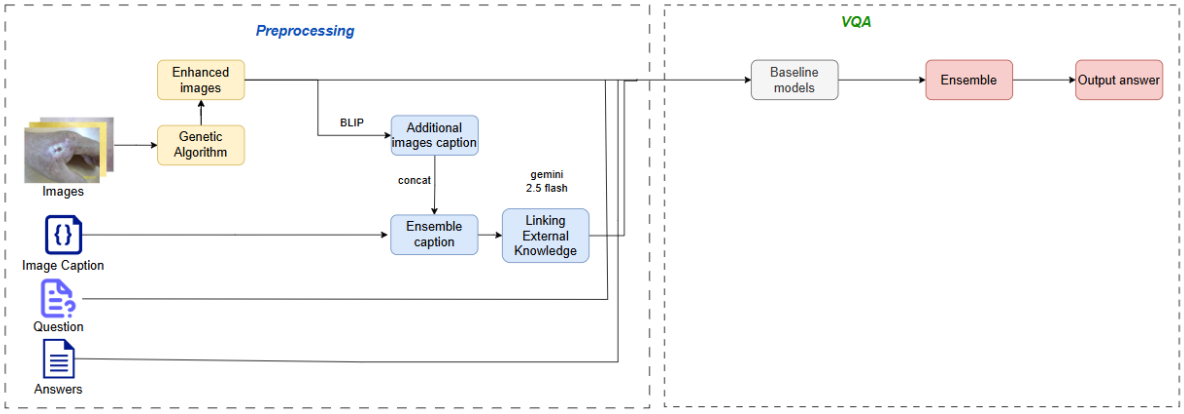

Figure 2: The DermKEM architecture

3. The Proposed Approach

In this work, we addressed two primary tasks. For Task 1, we employed DermoSegDiff,a boundary-aware system for skin lesions segmentation. For Task 2, we developed DermKEM, an advanced visual question answering (VQA) system tailored for dermatology. The architectures of these two systems are illustrated in Figure 1 and Figure 2, respectively, and will be elaborated upon in subsequent sections.

3.1. Data Preprocessing

3.1.1. Image Enhancement

We noticed that there is no image enhancement method in the proposed model DermoSegDiff, so we provided our own Genetic Algorithm-based image enhancement. We generate a population of 20 through 10 generations. For each image, we randomly modify the constrast α and the brightness β and calculate fitness using SSIM. Next, we select the top half by fitness and create new population by averaging α and β from two randomly selected top individuals, adding small random variation in the process. After 10 generations, we extract the best enhanced image based on SSIM with contrast α and brightness β. Figure 3.1.1 shows that there is significant improvement around skin lesions, making it clearer and easier to detect the boundary.

Figure 3: Image after proposed enhancement.

3.1.2. Image Resizing

Due to the inconsistency of the images provided in the dataset, we apply a resize image method proposed in DermoSegDiff[2] to ensure the uniformity of image size, therefore reducing training time due to resources restrain. Firstly, images are transformed into tensor. Next, we resize image tensors, adding interpolation to retain image information. We normalize image tensors to [0,1], replacing NaN values with 0.5.

3.1.3. Additional images caption

To generate descriptive image captions during preprocessing, we employed the BLIP (Bootstrapping Language-Image Pre-training) model [20]. To adapt this model for the dermatology domain, we fine-tuned the pre-trained BLIP on the SkinCap dataset [21], which contains 4,000 dermatology images with corresponding captions. The model was fine-tuned for 8 epochs, achieving a BLEU score of 0.16 on the SkinCap validation set. This confirmed the model's improved proficiency for generating domain-specific captions for our task.

3.1.4. Linking External Knowledge

In the Medical VQA task, image-generated captions often pose challenges for models that are not extensively trained on specialized medical contexts. These captions are typically short and may fail to accurately describe lesions or pathological signs. We leverage Gemini 2.5 Flash [4] via the Vertex AI API to enrich the captions by linking external knowledge from reputable open-access medical sources such as DermNet NZ (dermnetnz.org) and WikiDoc (wikidoc.org). This enriched captioning process helps the model better understand the medical image before answering the question.

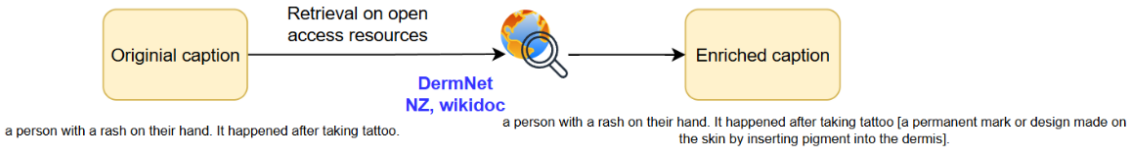

Figure 4: Linking External Knowledge

3.2. Methodology

3.2.1. Task 1: Segmentation

For the skin lesion segmentation task, we employed DermoSegDiff [2] by A. Bozorgpour et al., a diffusion-based segmentation model designed specifically for skin lesion segmentation. The model leverages the generative capabilities of Denoising Diffusion Probabilistic Models (DDPMs) while incorporating boundary-aware mechanisms to enhance segmentation precision, especially around

lesion boundary.

Diffusion Process: DermoSegDiff based on a standard DDPM framework that includes a forward process - where Gaussian noise is gradually added to the ground truth segmentation mask — and a reverse denoising process, which reconstructs the mask step-by-step with guiding image. Rather than predicting the mask directly, the network learns to estimate the noise added at each timestep, enabling more accurate reconstruction. **Loss Function:** To address the challenge of fuzzy and ambiguous lesion boundaries, DermoSegDiff introduces a novel boundary-aware loss function:

$$\mathcal{L}_w = (1 + \alpha W_\Theta) \left\| \epsilon - \epsilon_\theta(x_t, g, t) \right\|^2 \tag{1}$$

where $W_\Theta \in R^{(HxWx1)}$ is a dynamic weight map derived from a distance transform of the mask boundary. This term increases the weight of pixels near the lesion boundary and decreases when further from the boundary. The dynamic nature of W_Θ, based on the current time step t, ensures the progressive refinement of the boundary regions as the denoising proceeds. A gamma correction is applied to the distance map to control the sharpness of attention around the boundary.

Denoising Network Architecture: The denoising network is a modified U-Net [22] architecture with a two-path feature extraction mechanism:

- The image-conditioned path extracts semantic features from the input image g, providing contextual guidance throughout the reverse process.
- The latent-conditioned path processes the noisy segmentation mask x_t , learning to progressively reduce noise.

Each path contains ResNet blocks with separate time embeddings to capture distinct temporal characteristics. The features of both paths are fused at multiple stages of the U-Net[22], allowing an effective integration of semantic and noise-related representations. A dual-attention bottleneck module further enhances the model's ability to capture both spatial dependencies and long-range interactions.

The decoder reconstructs the estimated noise (ϵ_θ) by utilizing enriched skip connections from the encoder, which contains both semantic and boundary-focused information. An additional skip connection from the initial noisy input x_t to the output layer ensures the retention of noise characteristics critical for accurate reconstruction.

Inference Strategy with Sampling-based Ensemble: During inference, DermoSegDiff adopts a sampling-based ensemble strategy to enhance segmentation robustness. Specifically, the model generates nine segmentation predictions for each test image by running the diffusion sampling process multiple times. These predictions are then averaged pixel-wise, followed by thresholding to produce the final segmentation mask (threshold of 0). This ensemble method mitigates the stochastic nature of diffusion models and improves result consistency.

3.2.2. Task 2: Closed Visual Question Answering

We experiment two kinds of model: traditional model and Vision Language model to discover the performance of baseline models on dermatology visual question answering. For traditional model, we ultilize MUMC[3], the state-of-the-art (SOTA) model on medical visual question answering. For VLM, we use Gemini 2.5 Flash[4] through Vertex AI API.

MUMC: MUMC[3] uses a novel self-supervised pretraining method to efficiently learn to understand and associate information from medical images and texts through information masking and multi-level contrastive learning.

Gemini 2.5 Flash: Gemini 2.5 Flash[4] represents the latest advancement within the Gemini family of models. It is specifically engineered to optimize for speed and cost-effectiveness, thereby offering a substantially faster and more lightweight alternative for tasks demanding low latency and high throughput. While maintaining robust performance across diverse modalities including text, image, and audio (and potentially video, contingent upon its specific capabilities), Gemini 2.5 Flash excels in applications such as real-time interactive chat, text summarization, and on-device Artificial Intelligence.

This characteristic renders powerful multimodal understanding more accessible and practical for an expanded range of use cases.

Answer Shuffle: For Vision-Language Models (VLMs), we experimented with shuffling the order of answer options to evaluate model consistency. Typically, answer options are mapped as follows: option A to 1, option B to 2, option C to 3, and option D to 4. Subsequent to shuffling, an example mapping could be A to 3, B to 1, C to 4, and D to 2. The model is queried multiple times with these shuffled mappings, and the final selected option is determined by the answer choice that achieves the highest frequency of selection. In instances where multiple answer choices receive the same highest frequency, one is selected randomly, although this scenario was observed to occur infrequently.

Few-shot Learning: For VLMs, we implemented few-shot learning. For each input sample, we provided a set of ground-truth examples randomly selected from the training dataset as in-context learning prompts. The number of such examples was equivalent to the total number of questions in the evaluation set, as detailed further in Section 5.

Ensemble: We employed a hard-voting ensemble strategy, combining the outputs from multiple model inferences. The final output was determined as the answer most frequently selected by the constituent models. In cases of a tie for the most frequent answer, one was selected randomly.

4. Task and Dataset Descriptions

4.1. Dataset Descriptions

We used the official ImageCLEFmed MEDIQA-MAGIC 2025 dataset [23], which builds upon documentation from DermaVQA [24]. The dataset supports two tasks and is divided into training (2474 images), validation (157 images), and test (314 images) sets.

This dataset facilitates a segmentation task and is structured as follows: The training set, constituting 85%, consists of 2474 images, 7448 masks, and 842 queries. The validation set contains 157 images, 472 masks, and 56 queries, while the test set is composed of 314 images, 944 masks, and 100 queries. Mask files are stored as binary TIFF files, adhering to the naming convention `IMG_{ENCOUNTERID}_{IMAGEID}_mask_{ANNOTATOR#}.tiff`. Corresponding image files are available in PNG or JPG format, named as `IMG_{ENCOUNTERID}_{IMAGEID}.png` or `IMG_{ENCOUNTERID}_{IMAGEID}.jpg`. For the Closed QA task, the dataset includes closed questions, associated images, A dictionary of all possible closed questions and the option values associated with them, and a predefined list of 27 questions provided in the `closedquestions_definitions_imageclef2025.json` file, with both English and Chinese translations. Distribution of types of questions related to illnesses as show in Figure 5.

4.2. Task Definitions

The second edition of the MEDIQA-MAGIC task focuses on multimodal dermatology response generation. Building upon the previous year's challenge, this task introduces more complex reasoning by combining clinical narratives with associated dermatology images. The task is divided into two sub-tasks:

4.3. Segmentation

- **Definition:** Given a clinical history and an associated dermatological image, participants are required to generate segmentation masks that identify regions of interest related to the described dermatological condition.

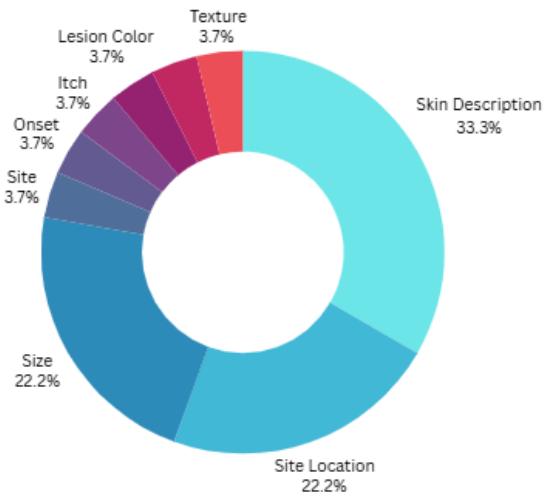

Figure 5: Question Distribution

4.4. Closed VQA

- **Definition:** Closed VQA Participants are provided with a dermatology-related query (clinical narrative), one or more related images, and a multiple-choice question. The goal is to select the correct answer from the provided options.

5. Experiment Results

5.1. Implementation Details

Our experiments for the ImageCLEF MAGIC 2025 challenge were conducted on the Kaggle platform, utilizing two NVIDIA T4 GPUs and one NVIDIA P100 GPU.

5.1.1. Task 1: Segmentation

For our segmentation task, we used the DermoSegDiff framework as the core model. To improve the quality of our training data, we applied a Genetic Algorithm (GA) before training the segmentation model. Since DermoSegDiff doesn't have built-in image enhancement, this step helped create clearer and more informative images. The enhanced images were used to build new versions of the training, validation, and test datasets. We trained the model with the following settings:

- **Batch size:** 8
- **Input image size:** 128×128 pixels
- **Diffusion settings:**
  - Timesteps: 250
  - Beta schedule: Linear
  - β_{start}: 0.0004
  - β_{end}: 0.08

For optimization, we used the Adam optimizer with:

- Learning rate: 0.0001
- Betas: (0.7, 0.99)
- Weight decay: 0.0

We also included a learning rate scheduler (ReduceLROnPlateau) to help the model train more efficiently. If the validation score didn't improve after 5 epochs, the learning rate was cut in half. Each run lasted up to 40,000 iterations, or ended earlier if the validation results stopped improving. No other parameters was tested due to time constraint, as we believed the original parameter is the most optimal for the task. For evaluation, we used an ensemble approach. Each test image was passed through the model 5 times, and we averaged the outputs pixel by pixel. The final result was turned into a binary mask using a threshold. This method helped reduce noise and made the predictions more stable and accurate.

5.1.2. Task 2: Closed Visual Question Answering

MUMC: We utilized the MUMC model. The model's pre-training stage involved a version pre-trained on three datasets: ROCO [25], MedICaT [26], and the ImageCLEF2022 Image Caption Dataset [27]. Subsequently, its fine-tuning stage employed a model pre-trained on three public medical VQA datasets: VQA-RAD [28], PathVQA [29], and SLAKE [30]. This model was then fine-tuned on our proprietary datasets using its default hyperparameter settings, with each training stage comprising 50 epochs.
Gemini 2.5 Flash: Gemini 2.5 Flash was accessed via the Vertex AI API, incorporating both few-shot learning and answer shuffling techniques. For few-shot learning, the number of in-context examples provided with each query was set to be equal to the total number of questions in the evaluation set. In this experiment, this translated to 27 few-shot examples accompanying each inference request. This approach aims to enhance prediction accuracy by providing relevant ground-truth examples as context. For answer shuffling, the order of answer options was permuted twice for each sample. Consequently, the model generated three outputs per sample: one with the original answer order and two with distinct shuffled orders.
Ensemble: The ensemble method aggregated outputs from multiple model inferences. Final predictions were determined using hard voting, where the answer option with the highest frequency among the collected outputs was selected. We experimented with three distinct ensemble configurations:

- MUMC combined with two Gemini 2.5 Flash inference runs (one using the original answer order, and one using a single shuffled answer order).
- MUMC combined with three Gemini 2.5 Flash inference runs (one original, and two using distinct shuffled answer orders).
- An ensemble of three Gemini 2.5 Flash inference runs (one original, and two using distinct shuffled answer orders).

5.2. Experimental Results

5.2.1. Evaluation Metrics

We used the official evaluation metrics defined by the ImageCLEFmed MEDIQA-MAGIC 2025 organizers for each task.

Task 1: Segmentation Performance was measured by the Jaccard Index (IoU) and Dice Coefficient.
$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \quad ; \quad \text{Dice} = \frac{2 \times |P \cap G|}{|P| + |G|}$
Where P is the predicted mask and G is the ground truth mask. The ground truth was constructed using a 'Majority Vote' from four annotators, where a pixel is considered positive if marked by at least two annotators. The final reported scores are macro-averaged IoU and Dice across the test set.

Task 2: Closed Visual Question Answering The task was evaluated using Overlap-based Accuracy, suitable for multi-label answers. The accuracy for each question i is calculated as: $\text{Accuracy\_}i = \frac{|G_i \cap P_i|}{\max(|G_i|, |P_i|)}$
Where G_i and P_i are the ground truth and predicted label sets, respectively. The final score is the mean of these per-question accuracies across the entire test set.

5.2.2. Results

Task 1: Segmentation This section presents the performance of our submission for the MEDIQA-MAGIC 2025 segmentation task. Our approach utilized the DermoSegDiff model, a state-of-the-art diffusion-based architecture for medical image segmentation. The model was evaluated on the unseen private test set using the official competition metrics. The official results are summarized in the Table 1.

Table 1
Private test results

Jaccard	0.5145	**Dice**	0.6794
Jaccard mean of max	0.6356	**Dice mean of max**	0.7430
Jaccard mean of mean	0.5472	**Dice mean of mean**	0.6591

Table 2 displays the official leaderboard of the MEDIQA-MAGIC 2025 Segmentation

Table 2
The official MEDIQA-MAGIC 2025 segmentation task results

Rank	Team Name	Jaccard	Dice
1	Anastasia	0.645774234	0.784766
2	IReL, IIT(BHU)	0.588132	0.7407
3	KLE1	0.540960473	0.702108
4	**H3N1 (Ours)**	**0.514468506**	**0.679405**
5	Kasukabe Defense Group	0.186594865	0.314505

Task 2: Closed Visual Question Answering For private test phase, we contributed 4 piplines, which detailed in Table 3. Gemini 2.5 Flash shows the efficientness that the ensemble contains three Gemini inference runs achieved the highest private test score.

Table 3
Overall Accuracy Results for the Evaluated Pipelines

ID	Pipeline	Overall Accuracy
1	Gemini 2.5 Flash (+Preprocessing)	0.7358
2	Ensemble (MUMC + two Gemini 2.5 Flash runs) (+Preprocessing) (+Shuffling)	0.7452
3	Ensemble (MUMC + three Gemini 2.5 Flash runs) (+Preprocessing) (+Shuffling)	0.7507
4	Ensemble (three Gemini 2.5 Flash runs) (+Preprocessing) (+Shuffling)	0.7580

Table 4 displays the official leaderboard of the MEDIQA-MAGIC 2025 Closed Visual Question Answering task. Our system achieved first place, outperforming all other teams. This highlights the robust and efficient performance of our proposed system.

6. Conclusion and Future Works

Limitations Despite the effectiveness of the proposed model, it exhibits notable limitations in handling very small lesions or subtle variations in texture and color, which may be imperceptible to the human eye or not be sufficiently captured by the model without specialized training. The model can also struggle to distinguish between lesion types with similar surface characteristics and may fail to fully capture multi-attribute lesions (e.g., those presenting multiple colors or patterns). Additionally, the encoder stage in the proposed diffusion network may not sufficiently extract comprehensive features

Table 4
The official MEDIQA-MAGIC 2025 Closed Visual Question Answering task results

Rank	Team Name	Overall Accuracy
1	**H3N1 (Ours)**	**0.7580**
2	DS@GT MEDIQA-MAGIC	0.7095
3	KLE1	0.5698
4	Kasukabe Defense Group	0.5366
5	Oggy	0.2223
6	IReL, IIT(BHU)	0.1731

from both the input images and corresponding segmentation masks, potentially limiting segmentation accuracy in complex cases.

Future Works Future improvements can be focused on combining MUMC and Gemini 2.5 Flash in a hybrid or ensemble model to improve both image analysis and language understanding, making the system more accurate and reliable. Fine-tuning the models on a larger and more diverse medical dataset, especially with rare skin conditions, could help improve generalization. Adjusting Gemini 2.5 Flash's prompt format to better handle medical terms may also boost performance. To reduce bias and increase realism, the dataset could be expanded with more context-based questions and images from people of various races and age groups. Evaluation can be strengthened by using metrics such as BLEU, ROUGE, and METEOR for free-form responses, as well as expert feedback from dermatologists. For the segmentation part, testing stronger encoder models like TransUNet [31] and applying more data augmentation can lead to better learning from the available data.

7. Appendices

A. MUMC Architecture and Training Details

The MUMC (Masked Vision and Language Pre-training with Unimodal and Multimodal Contrastive Losses) [3] framework operates in two main stages: pre-training and fine-tuning.

Stage 1: Pre-training This stage aims to learn robust and highly generalizable representations from large-scale medical image-caption datasets.

1. **Input Data Preparation:**
    - **Image:** Each image is divided into 16×16 patches. During training, 25% of the patches are randomly masked, and only the unmasked patches are input to the image encoder. This masked image modeling encourages the model to learn from partial visual information.
    - **Text (Caption):** TAssociated text descriptions are tokenized using a WordPiece tokenizer.
2. **Architecture:** A dual-encoder setup based on Momentum Contrast (MoCo) is employed, consisting of online and momentum encoders that stabilize contrastive training.
3. **Optimization Objectives:** The model is jointly optimized with four self-supervised losses:
    - **Contrastive Loss (UCL and MCL):** The core objective, which structures the latent space by pulling similar pairs together and pushing dissimilar pairs apart. Unimodal Contrastive Loss (UCL) operates within a single modality (image-to-image, text-to-text), while Multimodal Contrastive Loss (MCL) learns the alignment between modalities (image-to-text).
    - **Image-Text Matching (ITM) Loss**: A binary classification loss that determines whether a given image-caption pair is matched or randomly paired. This reinforces semantic alignment between modalities beyond just contrastive structure.

- **Masked Language Modeling (MLM) Loss**: A language modeling objective where 15% of input tokens in the caption are randomly masked and the model must predict them using the remaining text and associated image features, thereby strengthening contextual understanding

Stage 2: Fine-tuning The pretrained weights are transferred to a VQA model with a Transformer-based answering decoder (6 layers) that generates free-text answers from fused image-question embeddings. The model is fine-tuned using a standard cross-entropy loss over the target answer sequences.

B. Prompt for Gemini

This prompt utilizes a few-shot learning technique, wherein the model is provided with 27 complete examples from the training set before it is asked to process a new case. The primary objective of this structured prompt is to strictly constrain the model's output to a single integer corresponding to the chosen answer's index, thereby simplifying the results parsing process and increasing output consistency.

B.1. Prompt Structure

The prompt is constructed following a "guidance - examples - task" architecture. The full prompt template is provided below.

```
You are an expert dermatologist. Your task is to
answer a multiple-choice question about a clinical
case based on the provided clinical context and
image(s).

You will be shown several examples first, followed by
a new case to solve.

Your response MUST be a single integer representing
the index of the correct answer.

DO NOT include any other text, explanation, or
formatting. JUST THE NUMBER.

--- EXAMPLE 1 START ---
[IMAGE(S) ARE PROVIDED HERE]
Clinical Context: {Example 1 Context}
Question: {Example 1 Question}
Options:
0: {Option 0}
1: {Option 1}
...
Correct Answer Index:
{Example 1 Correct Index}
--- EXAMPLE 1 END ---

--- EXAMPLE 2 START ---
[IMAGE(S) ARE PROVIDED HERE]
Clinical Context: {Example 2 Context}
Question: {Example 2 Question}
```

```
Options:
0: {Option 0}
1: {Option 1}
...
Correct Answer Index:
{Example 2 Correct Index}
--- EXAMPLE 2 END ---

... (25 more examples follow the same structure) ...

--- EXAMPLE 27 START ---
[IMAGE(S) ARE PROVIDED HERE]
Clinical Context: {Example 27 Context}
Question: {Example 27 Question}
Options:
0: {Option 0}
1: {Option 1}
...
Correct Answer Index:
{Example 27 Correct Index}
--- EXAMPLE 27 END ---

--- YOUR TASK START ---
Now, analyze the following new case and provide your answer as
a single integer.
[IMAGE(S) ARE PROVIDED HERE]
Clinical Context: {Query Case Context}
Question: {Query Case Question}
Options:
0: {Query Option 0}
1: {Query Option 1}
2: {Query Option 2}
3: {Query Option 3}
Answer Index:
```

B.2. Explanation of Prompt Components

- **Role Definition**: The initial line, "You are an expert dermatologist," establishes a specific persona and domain expertise for the model. This helps the model to approach the problem from the perspective of a medical specialist, potentially activating more relevant reasoning paths.
- **Strict Output Instruction**: The lines "Your response MUST be a single integer..." are capitalized and emphasized to capture the model's attention and ensure it understands the precise output format required. This is the most critical change compared to previous prompt versions, as it eliminates conversational or explanatory text that would complicate automated evaluation.
- **Few-shot Examples**: A total of 27 examples (from ‘EXAMPLE 1‘ to ‘EXAMPLE 27‘) are included to "teach" the model the desired input-output format. Each example provides a complete instance, including the full context, the image (passed as an image object, represented by ‘[IMAGE(S) ARE PROVIDED HERE]‘), the question, the multiple-choice options, and the correct answer index. This in-context learning is crucial for guiding the model's behavior without updating its weights.
- **Query Task**: The final section, beginning with ‘— YOUR TASK START —‘, is where the actual data for the case to be predicted is inserted. Placeholders such as ‘Query Case Context‘ are dynamically replaced with the real data. The final line, ‘Answer Index:‘, acts as a direct cue for

the model to complete the sequence and provide its numerical answer.

C. Prompt for Knowledge Enrichment

The core of the enrichment process is to design sophisticated prompt for a powerful generative model (Gemini 2.5 Flash). The full prompt text is provided below to illustrate the detailed instructions given to the model.

```
ROLE: AI Medical Concept Enrichment Specialist

CONTEXT: You are tasked with processing text content,
typically image captions or descriptions
(query_content_en field) related to medical
observations, often within a Visual Question
Answering (VQA) context for medicine, especially
but not limited to Dermatology. Your goal is to
enhance this text by identifying specific
medical terms and appending concise, accurate
definitions. This prompt is designed for the Gemini
2.5 Flash model.

OBJECTIVE: To enrich the input text by identifying
relevant medical terms (including but not limited to
Dermatology) and appending their brief, accurate
definitions immediately after the term, formatted as
Term [Definition]. Definitions for
Dermatology-specific
terms should prioritize consistency with DermNet NZ
(site:dermnetnz.org). Definitions for other medical
terms (diseases, symptoms, findings, anatomical
locations, procedures, relevant medications) should
be consistent with standard medical knowledge and
ontologies like SNOMED CT
(site:https://www.snomed.org/)
or UMLS (site:
https://www.nlm.nih.gov/research/umls/index.html)
or wikidoc
(site:
https://www.wikidoc.org/index.php/Main_Page),
use your crawl skills to get information.
INPUT: A single string of text representing the value
of a query_content_en field or similar medical text
description.
OUTPUT: The modified string of text, with relevant
medical terms enriched as specified. The overall
structure and non-relevant parts of the original
text must remain unchanged.

CONSTRAINTS:
1. Scope: Enrich specific medical terms. This includes:
   - Names of diseases or conditions (e.g., psoriasis).
   - Specific symptoms or clinical findings (e.g.,
```

```
   macule, erythema).
   - Relevant anatomical locations (e.g., epidermis, dermis).
   - Medical or surgical procedures (e.g., biopsy).
   - Commonly referenced medications (e.g., methotrexate).
2. Exclusion: Do NOT enrich:
   - Highly general terms (e.g., 'disease', 'patient').
   - Common non-medical words (e.g., 'tired', 'left', 'right').
   - Terms already adequately explained by the context.
3. Source Prioritization:
   - For Dermatology: Prioritize DermNet NZ

   (site:dermnetnz.org).
   - For other medical terms: Use SNOMED CT, UMLS, Wikidoc.
4. Definition Format: Term [Concise, clear definition].
5. Definition Content: Brief, 1-2 short sentences.
6. Accuracy: Ensure definitions are medically accurate.
7. Case Sensitivity: Identify terms regardless of case,
but preserve
   original capitalization in the output.
8. No Modification Otherwise: Do not alter any other
part of the text.

INSTRUCTIONS:
1. Receive the input text string.
2. Scan the text to identify potential medical keywords.
3. For each term:
   a. Verify it meets enrichment criteria
   (CONSTRAINT 1 & 2).
   b. Determine if it is dermatological or general medical.
   c. Generate a concise definition per Source Prioritization.
   d. Format the definition as specified (CONSTRAINT 4 & 5).
   e. Append the formatted definition to the term.
4. If no relevant terms are found, return the original text.
5. Return the fully processed text string.

EXAMPLES:
Input Text (Dermatology Focus): The patient presented
with severe psoriasis and was prescribed methotrexate.
Output Text: The patient presented with severe psoriasis
[a common,
chronic inflammatory skin disease characterized by red,
itchy, scaly
patches] and was prescribed methotrexate [an
immunosuppressant drug...].
Input Text (General Medical): Image shows pitting edema on
the lower leg.
Output Text: Image shows pitting edema [swelling,
typically in the limbs, where pressing the skin
leaves a temporary indentation] on the lower leg.
Input Text (Exclusion): Doctors predict that he
has some kind of infection, he feels tired.
Output Text: Doctors predict that he has some kind of
```

```
infection [invasion and multiplication of
microorganisms...], he feels tired.
(Note: 'predict', 'tired' are not enriched).
Now, process the following input text based on these
instructions:
```

C.1. Analysis of Prompt Architecture and Effectiveness

The prompt's design is multi-faceted, aiming to transform a powerful but general-purpose language model into a precise and reliable medical annotation tool. Each component plays a strategic role in achieving this goal.

- **Role-Playing (‘ROLE‘):** The prompt assigns the model the persona of an ‘"AI Medical Concept Enrichment Specialist"‘. This technique is crucial for setting the context. It shifts the model from a general conversational mode to a professional, domain-specific one, encouraging it to access and utilize its training data related to medical science and formal writing styles.
- **Zero-Shot, Instruction-Following (‘INSTRUCTIONS‘):** The core of the prompt is a detailed, algorithmic set of instructions. Rather than relying on the model to infer the task from examples alone (few-shot), it explicitly defines the procedure: scan, identify, verify, generate, format, and return. This converts a potentially ambiguous creative task into a more deterministic, rule-based process, significantly increasing the reliability and consistency of the output.
- **Strict Constraints and Negative Logic (‘CONSTRAINTS‘):** A key to high-precision output is defining not only what to do, but also what *not* to do.
  - **Inclusion Scope:** By listing categories of terms to enrich (diseases, symptoms, anatomy), the prompt focuses the model's attention on high-value, specific medical concepts that carry significant diagnostic weight.
  - **Exclusion Scope:** The negative constraints (e.g., excluding 'patient', 'doctor', 'left') are vital for preventing "over-enrichment". Without these rules, the model might define common words, cluttering the output and diluting the importance of the truly significant medical terms. This improves the signal-to-noise ratio of the enriched caption.
- **Knowledge Grounding and Source Prioritization:** To prevent model "hallucination" or inaccurate definitions, the prompt explicitly grounds the required knowledge in authoritative external sources. By instructing the model to prioritize definitions consistent with **DermNet NZ** for dermatology and established ontologies like **SNOMED CT** or **WikiDoc** for general medicine, we guide the model to generate factually accurate and contextually appropriate information. The instruction to ‘"use your crawl skills"‘ leverages the model's ability to access and synthesize information from its vast training data, which includes these reliable web sources.
- **Format Enforcement via Examples:** While the prompt is primarily instruction-based, it includes a set of clear ‘EXAMPLES‘. These serve a critical function: they demonstrate the exact implementation of all the preceding rules, especially the strict output format of ‘Term [Definition]‘. The examples cover diverse cases, including dermatology-specific terms, general medical terms, and a case showing correct exclusion, leaving no ambiguity about the expected output. This combination of explicit instructions and illustrative examples is a powerful technique for ensuring the model adheres to the desired schema.

In conclusion, the effectiveness of our knowledge enrichment pipeline is not merely due to using a powerful LLM, but is a direct result of this carefully crafted prompt. It strategically combines role-playing, explicit instructions, positive and negative constraints, knowledge grounding, and clear examples to transform a general-purpose tool into a specialized, reliable, and highly effective component of our medical VQA system.

Declaration on Generative AI

During the preparation of this work, we used Gemini 2.5 Flash and ChatGPT-3.5 in order to: check grammar and sentence structure. After using these tools, we reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D4-2025-04.

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] A. Bozorgpour, Y. Sadegheih, A. Kazerouni, R. Azad, D. Merhof, Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation, in: I. Rekik, E. Adeli, S. H. Park, C. Cintas, G. Zamzmi (Eds.), Predictive Intelligence in Medicine - 6th International Workshop, PRIME 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings, volume 14277 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 146–158. URL: https://doi.org/10.1007/978-3-031-46005-0_13. doi:10.1007/978-3-031-46005-0_13.

[3] P. Li, G. Liu, J. He, Z. Zhao, S. Zhong, Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering, in: H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, R. Taylor (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Springer Nature Switzerland, Cham, 2023, pp. 374–383.

[4] S. B. J.-B. A. e. a. Gemini Team, Rohan Anil, Gemini: A family of highly capable multimodal models, 2025. URL: https://arxiv.org/abs/2312.11805. arXiv:2312.11805.

[5] J. Li, D. Li, C. Xiong, S. C. H. Hoi, BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 12888–12900. URL: https://proceedings.mlr.press/v162/li22n.html.

[6] T. B. Nguyen-Tat, H.-A. Vo, P.-S. Dang, Qmaxvit-unet+: A query-based maxvit-unet with edge enhancement for scribble-supervised segmentation of medical images, Computers in Biology and Medicine 187 (2025) 109762. URL: http://dx.doi.org/10.1016/j.compbiomed.2025.109762. doi:10.1016/j.compbiomed.2025.109762.

[7] T. B. Nguyen-Tat, T.-Q. T. Nguyen, H.-N. Nguyen, V. M. Ngo, Enhancing brain tumor segmentation in mri images: A hybrid approach using unet, attention mechanisms, and transformers, Egyptian Informatics Journal 27 (2024) 100528. URL: https://www.sciencedirect.com/science/article/pii/S1110866524000914. doi:https://doi.org/10.1016/j.eij.2024.100528.

[8] T. B. Nguyen-Tat, T. Q. Hung, P. T. Nam, V. M. Ngo, Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities, Alexandria Engineering Journal 119 (2025) 558–586. URL: https://www.sciencedirect.com/science/article/pii/S1110016825001176. doi:https://doi.org/10.1016/j.aej.2025.01.090.

[9] A. B. Abacha, S. A. Hasan, V. Datla, J. Liu, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in: Conference and Labs of the Evaluation Forum, 2019. URL: https://api.semanticscholar.org/CorpusID:198489641.

[10] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015. URL: https://arxiv.org/abs/1409.1556. arXiv:1409.1556.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385. arXiv:1512.03385.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[13] W.-w. Yim, A. Ben Abacha, Y. Fu, Z. Sun, F. Xia, M. Yetisgen, M. Krallinger, Overview of the MEDIQA-M3G 2024 shared task on multilingual multimodal medical answer generation, in: T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, D. Bitterman (Eds.), Proceedings of the 6th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 581–589. URL: https://aclanthology.org/2024.clinicalnlp-1.55/. doi:10.18653/v1/2024.clinicalnlp-1.55.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

[15] Anthropic, Claude 3 family, https://www.anthropic.com/news/claude-3-family, 2024. Accessed: 2024-04-24.

[16] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, W. Xie, Pmc-vqa: Visual instruction tuning for medical visual question answering, 2024. URL: https://arxiv.org/abs/2305.10415. arXiv:2305.10415.

[17] N. Siddique, S. Paheding, C. P. Elkin, V. Devabhaktuni, U-net and its variants for medical image segmentation: A review of theory and applications, IEEE Access 9 (2021) 82031–82057. URL: http://dx.doi.org/10.1109/ACCESS.2021.3086020. doi:10.1109/access.2021.3086020.

[18] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, 2018. URL: https://arxiv.org/abs/1807.10165. arXiv:1807.10165.

[19] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, Unet 3+: A full-scale connected unet for medical image segmentation, 2020. URL: https://arxiv.org/abs/2004.08790. arXiv:2004.08790.

[20] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, in: ICML, 2022.

[21] J. Zhou, L. Sun, Y. Xu, W. Liu, S. Afvari, Z. Han, J. Song, Y. Ji, X. He, X. Gao, SkinCAP: A Multi-modal Dermatology Dataset Annotated with Rich Medical Captions, 2024. arXiv:2405.18004.

[22] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, 2015. URL: https://arxiv.org/abs/1505.04597. arXiv:1505.04597.

[23] 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).

[24] W. wai Yim, Y. Fu, Z. Sun, A. B. Abacha, M. Yetisgen-Yildiz, F. Xia, Dermavqa: A multilingual visual question answering dataset for dermatology, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2024.

[25] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology objects in context (roco): A multimodal image dataset, in: D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, P. Jannin (Eds.), Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer International Publishing, Cham, 2018, pp. 180–189.

[26] S. M.-B. B. M. v. Z. S. P. S. S. M. G. Sanjay Subramanian, Lucy Lu Wang, H. Hajishirzi, MedICaT: A Dataset of Medical Images, Captions, and Textual References, in: Findings of EMNLP, 2020.

[27] J. Rückert, A. B. Abacha, A. G. S. de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of imageclefmedical 2022 - caption prediction and concept detection, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 1294–1307. URL: https://ceur-ws.org/Vol-3180/paper-95.pdf.

[28] J. J. Lau, S. Gayen, D. Demner, A. Ben Abacha, Visual question answering in radiology (VQA-RAD), 2022.

[29] X. He, Z. Cai, W. Wei, Y. Zhang, L. Mou, E. Xing, P. Xie, Towards visual question answering on pathology images, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 708–718. URL: https://aclanthology.org/2021.acl-short.90/. doi:10.18653/v1/2021.acl-short.90.

[30] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, X.-M. Wu, Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, 2021. URL: https://arxiv.org/abs/2102.09542. arXiv:2102.09542.

[31] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021. URL: https://arxiv.org/abs/2102.04306. arXiv:2102.04306.