

From Sights to Insights: Towards Summarization of Multimodal Clinical Documents

Akash Ghosh¹, Mohit Singh Tomar¹, Abhisek Tiwari¹, Sriparna Saha¹, Jatin Avinash Salve¹ Setu Sinha²

¹Department of Computer Science And Engineering, Indian Institute of Technology Patna, India

² Indira Gandhi Institute of Medical Sciences, Patna, India

{akash_2321cs19,sriparna}@iitp.ac.in, {mohitsinghtomar9797, abhisektiwari2014,jatinsalve.work,drsinhasetu}@gmail.com

Abstract

The advancement of Artificial Intelligence is pivotal in reshaping healthcare, enhancing diagnostic precision, and facilitating personalized treatment strategies. One major challenge for healthcare professionals is quickly navigating through long clinical documents to provide timely and effective solutions. Doctors often struggle to draw quick conclusions from these extensive documents. To address this issue and save time for healthcare professionals, an effective summarization model is essential. Most current models assume the data is only text-based. However, patients often include images of their medical conditions in clinical documents. To effectively summarize these multimodal documents, we introduce **EDI-Summ**, an innovative Image-Guided Encoder-Decoder Model. This model uses modality-aware contextual attention on the encoder and an image cross-attention mechanism on the decoder, enhancing the BART base model to create detailed visual-guided summaries. We have tested our model extensively on three multimodal clinical benchmarks involving multimodal question and dialogue summarization tasks. Our analysis demonstrates that **EDI-Summ** outperforms state-of-the-art large language and vision-aware models in these summarization tasks.

Disclaimer: The work includes vivid medical illustrations, depicting the essential aspects of the subject matter.

1 Introduction

One of the most impactful sectors where AI advancements can bring a significant revolution is healthcare. The latest WHO report highlights a drastic doctor-to-patient ratio, emphasizing the need for automation in various healthcare tasks. This imbalance, coupled with strides in information and communication technologies (ICTs) and the impact of COVID-19, has led to a significant rise in telehealth practices (Nittari et al., 2020). Recent

studies (Sahoo et al., 2022a), (Abacha et al., 2023), (Xue et al., 2018), (Sahoo et al., 2022b) demonstrate that AI models can be effectively employed in various healthcare settings and can drastically increase the productivity of healthcare professionals.

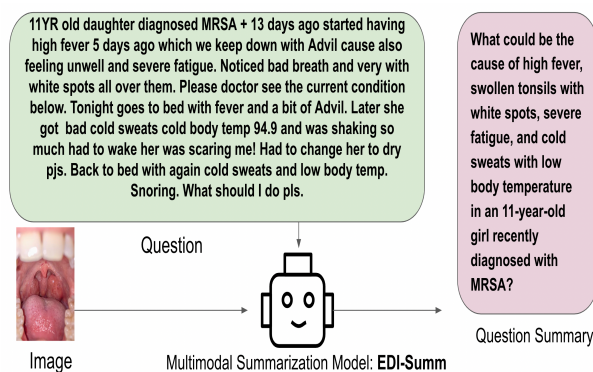


Figure 1: Our model **EDI-Summ** takes a multimodal clinical document as input and generates the corresponding summary.

In this regard, a significant development is that a considerable number of patients now discuss their medical history and conditions in various clinical forums before attending in-person appointments. Distilling key insights from these clinical documents not only boosts the efficiency of AI-based healthcare systems but also accelerates responses to patients, providing valuable time savings for healthcare professionals and facilitating a more prompt and beneficial exchange of information between both patients and doctors (Abacha and Demner-Fushman, 2019).

Most research in the field of medical document summarization has focused on unimodal, text-based approaches. Wei et al. (2023) explored medical query summarization using entity-driven contrastive learning. Similarly, (Joshi et al., 2020) addressed text-based medical dialogue summarization, while (DeYoung et al., 2019) focused on summarizing multiple medical documents. Recognizing the adage that **an image is worth a thousand**

words, a discernible shift is underway, with patients increasingly incorporating images of their medical conditions alongside their queries. This inclination arises from the fact that a considerable portion of the population may not be well-versed in medical terminology. This challenge becomes more pronounced when dealing with related symptoms, such as distinguishing between eye redness and eye inflammation. Integrating text and images provides a more comprehensive context, enriching the analysis beyond what textual examination alone might capture, which is underscored in the work (Kumar et al., 2018). Motivated by the research gap (Ghosh et al., 2024a), (Tiwari et al., 2023), and (Ghosh et al., 2024b) have contributed datasets along with their proposed models, utilizing techniques such as prompting and end-to-end fine-tuning to incorporate visual information alongside textual content. Each of these works has established its own model, but no single model has been developed that is well-suited for all these multimodal clinical summarization benchmark datasets. Our exploration culminates in proposing a model named *EDI-Summ*, designed to work across all the mentioned multimodal clinical summarization datasets. *EDI-Summ* integrates contextual multimodal fusion within the encoder and employs a cross-attention mechanism in the decoder of the BART architecture.

Research Questions: We aim to investigate the following research questions in this work:

R1) How is the performance of *EDI-Summ* compared to the baselines?

R2) How different pre-trained vision models like ResNet (He et al., 2016), VGGNet (Simonyan and Zisserman, 2014), VIT (Dosovitskiy et al., 2020) affect the quality of generated summary in terms of various evaluation metrics?

R4) How much does the performance vary if we change the order of image fusion in the encoder and decoder layers?

R5) How effective is *EDI-Summ* in handling Hindi-English codemixed text ?

R6) Do we need an Image Fusion module in both the encoder and decoder? What is the impact of Decoder Visual cross-attention?

Contributions: The key contributions of our work are stated below:

1) We present an innovative encoder-decoder visual-infused transformer model: *EDI-Summ*, which includes a contextual image fusion layer

on the encoder side and an image cross-attention layer on the decoder side. Our proposed model exhibits remarkable performance gains, substantially outperforming baseline models and unimodal architectures in comprehensive evaluations.

2) We undertook an exhaustive automated and human evaluation, complemented by in-depth qualitative analysis and risk assessments. This meticulous scrutiny was essential to guarantee the safety and reliability of the model's outputs, affirming its readiness for deployment in real-world healthcare scenarios with high confidence.

2 Related Works

The following works have been relevant to the following two research areas, namely (a) Works on Medical Document Summarization and (b) Works on Multimodal Summarization.

Works on Medical Document Summarization: Wang et al. (2020) was the first to conduct research on the summarization of COVID-19 documents. DeYoung et al. (2021) introduced a multi-document biomedical document summarization dataset. Wallace et al. (2021) introduced a dataset on the summarization of factual queries. In 2019, the MeQ-Sum dataset (Abacha and Demner-Fushman, 2019) revolutionized Medical Question Summarization (MQS), establishing itself as a dedicated resource for the field. There was a competition on the task in 2021 by Abacha (Abacha et al., 2021). Researchers initially utilized various pre-trained models like BART (Lewis et al., 2019), Pegasus (Zhang et al., 2020), and Prophetnet (Qi et al., 2020). Goff and Loehfelme (2018) proposed a novel method of using pre-trained models for summarizing the impression section of radiology reports. Joshi et al. (2020) proposed a method to exploit local structures for summarizing medical dialogues. Chintagunta et al. (2021) showed how GPT-3 can be utilized for generation of quality training data for medical dialogue summarization. Poornash et al. (2023) proposed a novel solution for layout summarization of biomedical articles.

Works on Multimodal Data: Previous research has demonstrated the benefits of incorporating multimodal information in various tasks (Ghosh et al., 2024c). For instance Jha et al. (2022) showed how combining both visual and language representations can help in patch-based identification of lexico-semantic relations. Suman et al. (2022) showed how multimodal information can be used

for personality prediction. Jha et al. (2024) showed multimodal explanations can be useful for enhanced cyberbullying understanding. Maity et al. (2022) showed how multimodal information from memes helps in combating cyberbullying. Tiwari et al. (2022) highlighted how multimodal information improves the performance of Disease Diagnosis Virtual Assistants. Delbrouck et al. (2021) showed that integrating images leads to better summarization of radiology reports. Kumar et al. (2023) showed multimodality helps in summarizing news articles. Verma et al. (2023) developed a large multilingual multimodal dataset. Kumar et al. (2023) illustrated how multimodal summarization can be useful for extracting insights from opinions. Joshi et al. (2020) and Molenaar et al. (2020) worked on the task of handling dialogues between patients and doctors and how to summarize their conversations. Goff and Loehfelm (2018) and Cai et al. (2021) are some of the first to propose the task of radiology report summarization.

3 Methodology

This section delineates the problem statement and outlines the various components of our proposed model.

3.1 Problem Formulation

Formally given an input textual document (D) represented by $D = \{d_0, d_1, \dots, d_n\}$ where n is the sequence length of the document, and a visual image corresponding to a textual query represented by (V), the task is to generate a clinically nuanced summary guided by the visual cue. The proposed architecture and its components are shown in Figure 2. The proposed model **EDI-Summ** has three main components, namely: i) *Representation of different Modalities* ii) *Encoder Contextual Image Fusion* iii) *Decoder Image Cross Attention*. A thorough explanation of the functionality of each section is provided in the following sections.

3.2 Representation of Different Modalities

Visual Representation: Our experimentation involved ResNet, VGG-Net, and ViT for generating image embeddings for the multimodal fusion layer. Since these pre-trained models produced image embeddings of dimension 2048, we introduced an additional neural network atop them to transform its model dimension to align with the textual dimension. The weights of all layers, excluding the

last three, were frozen, and we derived a vector representation of the image by pooling the output from the last layer, ultimately achieving a reduced embedding dimension of 768.

Textual Document Representation: In our study, a patient’s medical document is represented as a sequence of words. We utilize the pre-trained BART-base language model to generate 768-dimensional embeddings for textual data. With six layers in its encoder and decoder, BART captures bidirectional contextual information, producing a fixed-size textual representation. BART tokenizers are instrumental in breaking down input text into tokens, ensuring effective encoding for our multimodal clinical query summarization.

3.3 Encoder Contextual Image Fusion

The various components of the Encoder Contextual Image Fusion are elucidated in the following mentioned points:

a) Multimodal Context Aware Self Attention:

The effectiveness of the aggregated representation hinges on the seamless integration of multiple information sources. Inspired by the findings of Yang et al. (2019), we proposed a multimodal context-aware self-attention that produces conditioned key (K) and value (V) vectors that become instrumental in generating modality-aware attention vectors. This is elaborated in Figure 2.

Upon obtaining the intermediate representation $H = \{h_0, h_1, \dots, h_n\}$ generated by the BART encoder at a particular layer, we proceed to compute the query (Q), key (K), and value (V) $\in \mathbb{R}^{n \times d}$ vectors, where d is the model dimension. This calculation is shown in Equation 1.

$$\begin{pmatrix} Q \\ K \\ V \end{pmatrix} = H \begin{pmatrix} W_Q \\ W_K \\ W_V \end{pmatrix} \quad (1)$$

Here $W_Q, W_K,$ and $W_v \in \mathbb{R}^{d \times d}$ are the learnable parameters corresponding to the query, key, and value vectors, respectively.

In assessing the interrelation between textual query and the visual information, we create conditioned key (\hat{K}) and value (\hat{V}) vectors tailored to textual and visual contexts. These attention vectors transpose the query vector derived from the dialogue transcript, generating a versatile, image-infused information vector. The key and value pairs are calculated as shown in Equation 2. Here, we pass visual representation through a transformer

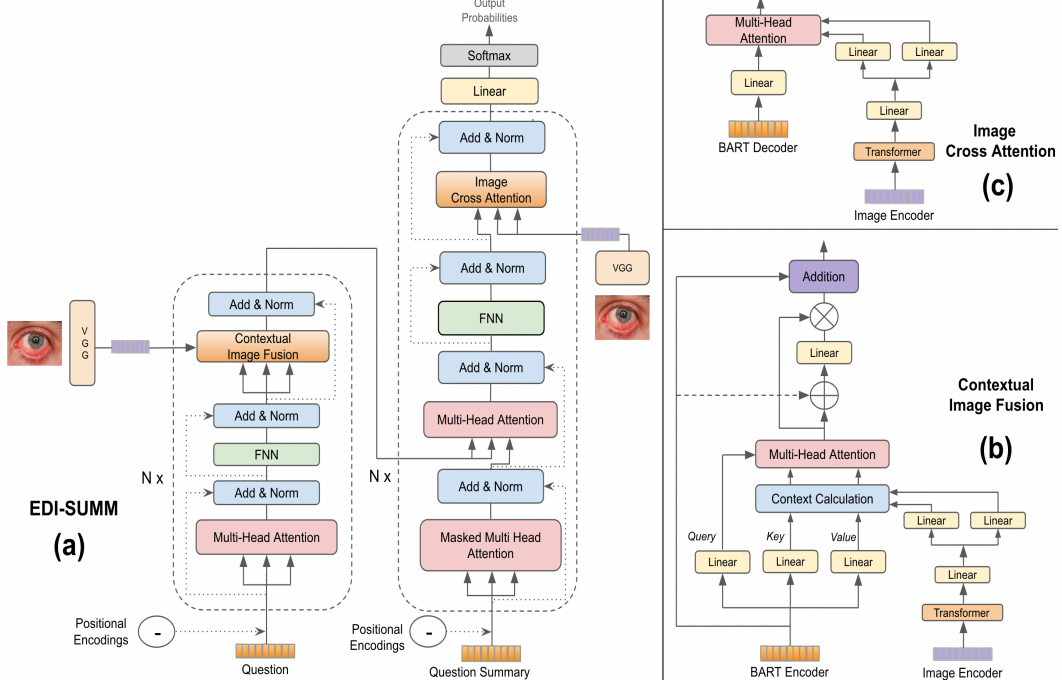


Figure 2: Architecture of our model (a) *EDI-Summ*, (b) Contextual Image fusion, (c) Image Cross Attention. Here, we obtain textual representation from BART and visual representation from VGG-Net. We perform image fusion in BART Encoder and Decoder and generate the question summary at the decoder.

and project its sequence length¹ to be equal to textual sequence length in order to perform multi-modal fusion.

$$\begin{bmatrix} \hat{K} \\ \hat{V} \end{bmatrix} = \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} (G \begin{bmatrix} U_k \\ U_v \end{bmatrix}) + (1 - \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix}) \begin{bmatrix} K \\ V \end{bmatrix} \quad (2)$$

where $G \in \mathbb{R}^{n \times d}$ indicates visual representation, U_k and $U_v \in \mathbb{R}^{d \times d}$ are the learnable parameters.

The vector $\lambda = \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix}$ controls how much information to retain from visual modality. This parameter can be calculated by the Equation 3.

$$\begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} = \sigma \left(\begin{bmatrix} K \\ V \end{bmatrix} \begin{bmatrix} W_{k_1} \\ W_{v_1} \end{bmatrix} + G \begin{bmatrix} U_k \\ U_v \end{bmatrix} \begin{bmatrix} W_{k_2} \\ W_{v_2} \end{bmatrix} \right) \quad (3)$$

where $W_{k_1}, W_{k_2}, W_{v_1},$ and $W_{v_2} \in \mathbb{R}^{d \times 1}$ are learnable parameters. The final scaled dot product attention can be calculated as shown in Equation 4.

$$H_v = \text{Softmax} \left(\frac{Q \hat{K}^T}{\sqrt{d_k}} \right) \hat{V} \quad (4)$$

where H_v represents visual information fused vector.

b) Unified Multimodal Integration: To regulate and modulate the flow of the visual-infused

¹We transform image sequence length (1) to textual sequence length (n). Image sequence length is one because it is a single image corresponding to a given textual query.

information, we utilize visual gates. The transmission of contextual information occurs through these gates according to Equation 5:

$$g_v = \sigma([H \oplus H_v]W_v + b_v) \quad (5)$$

Here, $W_v \in \mathbb{R}^{2d \times d}$ and $b_v \in \mathbb{R}^{d \times 1}$ are learnable parameters. The ultimate amalgamated representation, denoted as \hat{H} , is obtained according to the Equation 6.

$$\hat{H} = H + g_v \odot H_v \quad (6)$$

The contextualized image-infused vector (\hat{H}) is passed to the upper layer of the transformer and sent to the decoder.

3.4 Decoder Image Cross Attention

In the BART Decoder, we inject an Image cross-attention block for performing image fusion. We pass the image representation obtained from an image encoder to a transformer, then project its sequence length to the textual sequence length and obtain image representation $I \in \mathbb{R}^{m \times d}$, m is the sequence length at the decoder. Let the intermediate representation at the input of Image cross-attention be $F \in \mathbb{R}^{m \times d}$. We obtain query (Q_d) from F and key (K_d) and value (V_d) vectors from I as shown in Equation 7.

$$\begin{pmatrix} Q_d \\ K_d \\ V_d \end{pmatrix} = \begin{pmatrix} FW_{Q_d} \\ IW_{K_d} \\ IW_{V_d} \end{pmatrix} \quad (7)$$

Here W_{Q_d} , W_{K_d} , and $W_{V_d} \in \mathbb{R}^{d \times d}$ are trainable parameters. Then, we perform multi-head cross-attention as shown in Equation 8 to obtain a new image-infused vector F_I .

$$F_I = \text{Softmax}\left(\frac{Q_d K_d^T}{\sqrt{d_k}}\right) V_d \quad (8)$$

We combine the image-infused vector F_I with intermediate representation F using a gate mechanism shown in Equation 9 and Equation 10.

$$g_d = \sigma([F \oplus F_I] W_{F_d} + b_{F_d}) \quad (9)$$

Here $W_{F_d} \in \mathbb{R}^{2d \times d}$ and $b_{F_d} \in \mathbb{R}^{d \times 1}$ are trainable parameters. We pass the fused representation \hat{F} to the upper layers of the decoder and generate the summary.

$$\hat{F} = F + g_d \odot F_I \quad (10)$$

4 Experimental Results and Analysis

The following section outlines the datasets used for analysis, experimental setup, and evaluation discussion of the performance of the proposed model across a spectrum of both automatic and human evaluation metrics. A qualitative analysis of the generated summaries is conducted under various model configurations.

4.1 Datasets Used

To ensure the effectiveness of the proposed model, we conducted experiments on three multimodal clinical datasets: MMQS (Ghosh et al., 2024a), MMCQS (Ghosh et al., 2024b), and MM_CliConSummation (Tiwari et al., 2023). These datasets cover topics related to multimodal question summarization and multimodal dialogue summarization in the context of healthcare. The MMQS and MMCQS datasets contain approximately 3,015 data points and cover 18 medical symptoms. In comparison, the MM_CliConSummation dataset includes around 1,668 samples and encompasses 266 symptoms.

4.2 Experimental Setup

Our experimentation employed RTX 2080 Ti GPU, with each model requiring an average runtime of 20-30 minutes. Leveraging the PyTorch (Paszke

et al., 2019) and Hugging Face (Wolf et al., 2019) libraries, we executed both baselines and our proposed architecture. The foundation of our proposed model rests on the BART (Lewis et al., 2019). The dataset was partitioned into training, validation, and test sets in an 80:5:15 ratio. Hyperparameters included maximum epochs (30), maximum sequence length (360), visual embedding size (786), Adam optimizer, learning rate (5e-05), and batch size (32). We have performed the most extensive experiments in the case of the MMCS dataset. For the case of MMCS, our textual baselines featured T5 (Raffel et al., 2020), GPT-3.5 (OpenAI, 2023a), and BART (Lewis et al., 2019). For multimodal baselines, we used KM-CliConSummation (Tiwari et al., 2023) and GPT-4V (OpenAI, 2023b). Both GPT-3.5 Turbo and GPT-4V are used in the few-shot setting. In (Sahoo et al., 2024b), it is clearly stated that shot prompting is more effective in general than zero shot. Our proposed multimodal model, named *EDI-Summ*, is built on top of BART, where image fusion is done in both the encoder and decoder parts of the BART model. It is the first work where image fusion is performed in both the encoder and decoder portions of the BART model. *EI-Summ* is a simpler version of *EDI-Summ* where image fusion is done only in the encoder part of BART, inspired by (Kumar et al., 2022). We experimented with ResNet, VGG-Net, and ViT to generate image embeddings for our proposed multimodal model². For the task of multimodal dialogue summarization, the chosen baselines were GPT-3.5, GPT-4V, and KM-CliConSummation (Tiwari et al., 2023). To evaluate the capability of handling code-mixed text, we used the MMCQS dataset. The baselines selected for this task were GPT-3.5, GPT-4, and the MedSumm models (Ghosh et al., 2024b) of LLAMA2 and Zephyr versions. We utilized automated metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) scores alongside human evaluation metrics. We collaborated with a healthcare professional and several medical students to conduct human evaluations. Our approach encompasses the assessment of three unique and medically nuanced metrics: clinical evaluation score, factual recall (Abacha et al., 2023), and omission rate (Abacha

²We conducted experiments with varying image embeddings on the MMQS dataset. However, we observed that the results did not differ significantly across these embeddings. Therefore, we did not extend similar experiments to the MM-CQS and MM_CliConSummation datasets.

et al., 2023).

4.3 Automated Evaluation

Tables 1, 3, and 4 present the comprehensive evaluation of our proposed model, *EDI-Summ*, across three different multimodal medical summarization datasets: MMQS, MMCQS, and MM_CliConSummation. MMQS and MMCQS focus on multimodal medical query summarization, while MM_CliConSummation focuses on multimodal medical dialogue summarization.

R1) Comparison with baselines: From Table 1, we can infer that T5’s performance lagged behind the BART base model among the range of textual baseline models, leading to the conclusion that BART is a more advantageous choice for multimodal extension. In the case of multimodal baselines, KM-CliConSummation performed the best, followed by GPT-4V. For the task of medical dialogue summarization, as shown in Table 3, KM-CliConSummation achieved the highest performance. MedSumm(LLAMA-2) version was the top performer for code-mixed query summarization. The selection of these baselines is based on the works by (Ghosh et al., 2024b), (Ghosh et al., 2024a), and (Tiwari et al., 2023).

R2) Influence of Visual Cues: The impact of incorporating visual cues with textual queries for generating more nuanced summaries is evident across all metrics from Table 1, 3 and 4. The encoder image fusion model: *EI-Summ* and *EDI-Summ* exhibit significant superiority over models that rely solely on textual queries as input.

R3) Impact of Varied Pre-trained Vision Models for Generating Embeddings: In our exploration, as shown in Table 2, we investigated the impact of employing different pre-trained vision models to generate image embeddings for integration into the BART model. Our findings indicate that embeddings from ResNet and VGG tend to exhibit slightly superior performance compared to ViT, albeit the distinction is not markedly significant.

R4) Ablation study with different modality infusion orders: Table 2 elucidates how performance varies with changes in the infusion order within the encoder and decoder of the BART model employing VGG embeddings. Our experimental findings underscore a consistent trend, revealing that optimal results are consistently achieved when fusion occurs at layer 3 in both the encoder and

decoder of *EDI-Summ*. This pattern holds true across all pre-trained vision embeddings. In the scenario of the *EI-Summ*, the most favorable outcomes are attained through fusion at layer 3 of the encoder.³

R5) Effective in handling code-mixed text : Table-4 clearly suggests the effectiveness of *EDI-Summ* in handling Hinglish text. It performed much better than the baseline models that were proposed in (Ghosh et al., 2024b).

R6) Impact of Decoder Visual Cross Attention: A consistent trend emerges with the enhancement of ROUGE, BLEU, and METEOR scores when using *EDI-Summ* instead of *EI-Summ*, as shown in Tables 1, 3, and 4. Our evaluation across these three multimodal clinical summarization datasets supports our intuition that decoder attention improves the alignment of multimodal information.

4.4 Human Evaluation

As hallucination is a big problem in multimodal generation (Sahoo et al., 2024a), we have conducted a rigorous human-level analysis of the generated summaries. A team of medical students, under the guidance of a doctor, conducted the human evaluation on 55 data samples from each of the generated summaries from the three datasets. For the baseline for human evaluation we decided to go with the generated summaries of GPT-4v as it is the state-of-the-art multimodal baseline. To evaluate the significance of decoder attention, we compare the results of *EI-Summ* and *EDI-Summ* across all three different datasets. The evaluation metrics included: **Clinical Evaluation Score:** The doctor and the team assigned ratings ranging from 1 (poor) to 5 (good), evaluating the summaries based on overall relevance, consistency, fluency, and coherence, **Medical Fact-Based Metrics:** Factual Recall (Abacha et al., 2023) and Omission Recall metrics (Abacha et al., 2023) were employed to measure how well the generated summary captured medical facts compared to the gold standard annotated summary. Table-5 presents the results of *EDI-Summ*, which outperforms all baseline methods like GPT-4V across the selected human evalua-

³In Table 2, we present our experiments on the MMQS dataset. We conducted similar experiments on the MMCQS and MM_CliConSummation datasets and observed consistent trends across all three datasets.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
T5	45.28	28.12	41.98	39.2	27.9	22.91	18.9	41.96
BART	47.87	30.91	44.62	41.67	31.62	25.67	22.09	44.62
GPT-3.5 (few shot)	44.59	24.07	36.00	21.18	15.47	11.68	9.03	31.42
KM-CliConSummation	49.46	30.58	46.51	42.59	32.29	23.42	17.10	47.10
GPT-4V (few shot)	48.44	25.26	39.90	29.36	20.40	14.59	10.22	37.16
EI-Summ (RESNET)	54.22	30.92	46.72	43.24	30.94	22.96	17.06	49.43
EI-Summ (VGG)	53.88	30.41	46.77	43.12	30.45	22.75	16.62	50.24
EI-Summ (ViT)	53.62	30.12	46.34	43.10	30.55	22.55	16.55	49.16
EDI-Summ (RESNET)	54.50	30.89	47.38	44.27	31.55	23.64	17.64	51.09
EDI-Summ (VGG)	54.74	31.35	47.14	44.39	31.72	23.46	17.14	51.52
EDI-Summ (ViT)	53.91	30.53	46.95	43.77	31.32	23.11	16.92	49.50

Table 1: Performances of different models for multi-modal clinical query summarization in MMCS Dataset. *EDI-Summ*'s performance is superior to all the baselines and encoder-only fusion models. Our experiment shows *EDI-Summ* performs best with VGG embeddings, and EI-Summ performs best with RESNET embeddings. Among unimodal baselines, BART worked best, and KM-CliConSummation achieved the best results among multimodal baselines. The best results for each subsection are shown in bold.

Encoder layer	Decoder layer	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
2	3	54.28	30.71	47.25	44.25	31.70	23.76	17.88	50.64
2	4	53.56	30.89	46.9	43.54	31.30	23.54	17.6	51.36
3	3	54.74	31.35	47.14	44.39	31.72	23.46	17.14	51.55
3	4	53.43	29.91	46.61	43.54	30.85	22.98	16.63	51
4	4	54.23	30.88	46.65	43.7	31.38	23.29	16.68	49.9

Table 2: Evaluation of the proposed *EDI-Summ* model under diverse scenarios involving variations in the infusion order of modalities. Here, VGG-Net embeddings are used for multimodal fusion. We achieved the best results when fusion is done at layer 3 of both the encoder and decoder which are shown in bold.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
GPT-3.5(few shot)	51.79	27.84	47.75	40.6	28.7	19.97	14.18	42.13
GPT-4V(few shot)	52.3	27.98	48.2	41.39	29.21	20.18	14.30	42.52
KM-CliConSummation	60.27	37.90	50.87	48.77	37.37	28.44	22.16	56.70
EI-Summ	62.00	39.12	53.16	50.00	38.54	29.85	23.37	59.55
EDI-Summ	62.33	40.40	53.37	49.61	38.71	30.23	23.93	61.09

Table 3: Evaluation of different models on the MM_CliConSummation dataset. ResNet is utilized for generating image embeddings. The performances of *EI-Summ* and *EDI-Summ* are quite similar.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
GPT-3.5(few shot)	39.95	14.93	27.95	23.50	13.66	8.34	5.26	25.67
GPT-4V(few shot)	43.12	18.02	31.79	28.47	17.81	12.06	8.03	29.94
MedSumm (LLAMA-2)	46.75	25.59	38.41	32.5	22.55	17.56	14.88	35.74
MedSumm (Zephyr)	44.55	25.37	34.97	27.05	19.48	15.84	13.6	33.37
EI-Summ	53.20	30.87	44.92	45.00	33.81	27.11	23.52	47.54
EDI-Summ	53.32	30.64	45.01	45.88	34.13	27.33	23.51	48.46

Table 4: Performance of different models on MMCQS dataset. Surprisingly, *EDI-Summ*'s performance was way superior to the multimodal baselines like MedSumm and GPT-4V, which shows the effectiveness of our proposed model in handling codemixed text.

tion metrics on all three datasets.⁴

4.5 Qualitative Analysis of Generated Summaries

In Figure 3, we examine two examples from MMQS dataset comparing the gold summary with

⁴One key observation is that GPT-4V sometimes misses important medical concepts in the generated summaries.

the BART (Text only) model, *EI-Summ* and *EDI-Summ*. In both instances, the text-only model exhibits the poorest performance. Among the remaining two models, it is evident that the Encoder-only model overlooks certain intricate clinical terms in the final summary, which is potentially crucial for analyzing a patient's case. On the other hand, *EDI-Summ* consistently generates the most clinically

Dataset	Model	Factual recall	Omission rate	Clinical eval score
MMQS	GPT-4v	0.76	0.22	3.23
	EI-Summ	0.78	0.19	3.45
	EDI-Summ	0.81	0.17	3.51
MMCQS	GPT-4v	0.72	0.28	3.12
	EI-Summ	0.76	0.22	3.28
	EDI-Summ	0.79	0.18	3.47
MM_CliConsummation	GPT-4v	0.77	0.24	3.29
	EI-Summ	0.81	0.19	3.44
	EDI-Summ	0.81	0.20	3.43

Table 5: Results of human evaluation across three different datasets. A consistent trend was observed, indicating that **EDI-Summ** outperformed the selected baselines. Notably, GPT-4V’s performance was even weaker than that of EI-Summ.

nuanced summaries, as shown in the highlighted text in red color in Figure 3. *More qualitative analysis is shown in the Appendix section*



Image		
Question	Good day, I have a swollen lump on my tummy (right hand side), I consulted the doctor before and he said it was hernia. Few months after that I was diagnosed with ulcers in my mouth. Please see the image of mouth below and they are very painful. I went to another doctor and I explained to him all the diagnosis but he had a different opinion, he ruled out hernia and suspects gastrointestinal tract. I have both symptoms of ulcers, hernia and gastrointestinal tract. Now the lump gets sore and painful and sometimes feels itchy, please advice what could be the problem. Many thanks.	Hi, I am a 41 yr old female, for past three months I have had a heavy cold with bad catarrh, headaches, feverish and constantly tired. For the past 3 weeks I have had a hard lump to the side of my neck about 1 inch in circumference. Doctor Please look at the image of patient tonsils below. It's hard and sore to touch, its on bottom right hand side on the bone.
Gold Response	What could be causing the patient's swollen and painful lump, mouth ulcers, and conflicting diagnoses of hernia and gastrointestinal tract issues?	What is the potential diagnosis for a 41-year-old female with a heavy cold, headaches, fever, fatigue, and a hard, sore lump on the side of her neck?
BART (Text)	What causes swollen lump and symptoms of ulcers?	What could be the cause of hard lump on neck, on bottom right?
KM-CliConSummation	What could be the cause of a painful, swollen lump on the tummy?	What could be the cause of a hard lump to the back of the neck?
EI-Summ	What could be the cause of a swollen lump on the tummy that is sore, painful, and itchy?	What could be the cause of a hard, sore lump to the side of the neck in a 41-year-old female patient?
EDI-Summ	What could be the cause of a painful, swollen lump on the tummy, along with ulcers, hernia, and gastrointestinal issues?	What could be the cause of a 41-year-old female's heavy cold, catarrh, headaches, fever, fatigue, and hard, sore lump in the neck?

Figure 3: Analysis of summaries generated by our proposed model concerning various baselines. Our proposed model outperformed various baselines in capturing essential medical concepts, as evident in the highlighted text.

4.6 Statistical Analysis of the Generated Summaries

The median number of words in gold summary is 24, and in generated summaries is around 23, so they are pretty close. The golden and predicted summaries’ word clouds are shown below in Figure 4.

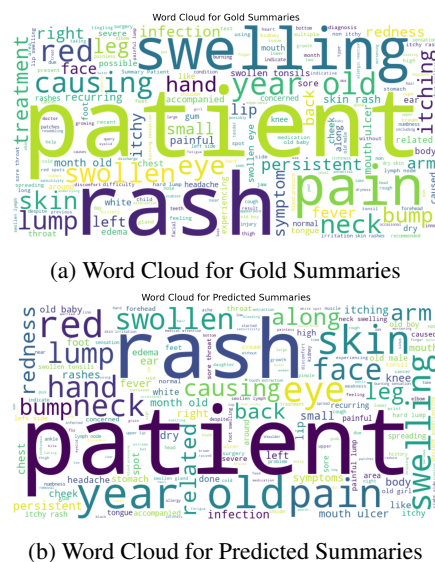


Figure 4: Comparison of Word Clouds for Gold and Predicted Summaries for MMCS dataset

5 Risk Analysis

Through comprehensive automated, human, and qualitative analyses, we have discerned that visual cues significantly contribute to generating more clinically nuanced summaries. However, particularly in sectors as critical as healthcare, the stakes are high for any potential misinformation or misinterpretation. In our human evaluation, instances were identified where the model, due to visual information, missed crucial keywords from the textual query. This occasional divergence underscores the need to carefully balance visual and textual inputs to ensure a more comprehensive summary. Moreover, our model is currently limited to recognizing 18 main symptoms. If presented with an image con-

taining symptoms outside this set, the model might inadvertently generate misinformation in the final summary. Additionally, the quality of the image plays a pivotal role in generating accurate summaries. Our experimentation with a low-quality image revealed a lack of proper detail capture. Hence, the involvement of a medical expert becomes imperative, particularly in high stakes, to mitigate the risks associated with potential inaccuracies or oversights.

6 Conclusion

In this paper, we introduced a multimodal model *EDI-Summ*, which incorporates contextual multimodal attention in both the encoder and decoder of the BART model. To evaluate the effectiveness of our proposed model, we conducted automated and human evaluations on three clinical multimodal summarization benchmark datasets. Our analysis suggests that incorporating contextual attention in the encoder and cross-attention in the decoder enhances the performance of multimodal summarization.

7 Limitations

There are some noticeable limitations of our work. Those are enumerated in the points below:

1) Although our comprehensive evaluation indicates that incorporating attention in both the encoder and decoder of the BART model achieves state-of-the-art performance for multimodal summarization, this claim lacks strong theoretical support.

2) In our experiments, we primarily focused on English and code-mixed settings. However, the effectiveness of this framework in handling multilingual text remains an open question. We plan to explore this research question in future work.

3) All the datasets MMQS, MMCQS, and MM_CliConSummation are multimodal, handling only text and images as inputs. However, the effectiveness of our proposed *EDI-Summ* in handling other modalities, such as videos and speech, remains to be explored in future work.

8 Ethics Statement

Our collaborative efforts extend to close coordination with a medical expert, who also holds the position of co-authorship on this paper. The execution of the entire task, spanning from experiment

design to human validation and qualitative analysis, involved the dedicated participation of three MBBS students, who volunteered for the project. As a testament to ethical considerations, these students were compensated following the prevailing minimum wage guidelines in India as outlined⁵. Furthermore, to reinforce the ethical integrity of our work, we are in the process of obtaining Institutional Review Board (IRB) approval for this project. It is important to note that the proposed model is designed solely for the task of summarization and does not include any predictive functionalities that could unfairly impact the user. This deliberate design choice underscores our commitment to ethical practices and responsible model usage.

9 Acknowledgements

Akash Ghosh and Sriparna Saha extend their sincere appreciation to the SERB (Science and Engineering Research Board) POWER scheme, Department of Science and Engineering, Government of India, for generously funding this research.

References

- Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234.
- Asma Ben Abacha, Yassine M'rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediq 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85.
- Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation metrics for automated medical note generation. *arXiv preprint arXiv:2305.17364*.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Xiaoyan Cai, Sen Liu, Junwei Han, Libin Yang, Zhenguo Liu, and Tianming Liu. 2021. Chestxraybert: A pretrained language model for chest radiology report summarization. *IEEE Transactions on Multimedia*, 25:845–855.

⁵<https://www.india-briefing.com/news/guide-minimum-wage-india-2023-19406.html/>

- Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR.
- Jean-Benoit Delbrouck, Cassie Zhang, and Daniel Rubin. 2021. QIAI at MEDIQA 2021: Multimodal radiology report summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 285–290, Online. Association for Computational Linguistics.
- Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024a. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.
- Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024b. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *European Conference on Information Retrieval*, pages 106–120. Springer.
- Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024c. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.
- Daniel J Goff and Thomas W Loehfel. 2018. Automated radiology report summarization using an open-source natural language processing pipeline. *Journal of digital imaging*, 31:185–192.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Prince Jha, Gaël Dias, Alexis Lechervy, Jose G Moreno, Anubhav Jangra, Sebastião Pais, and Sriparna Saha. 2022. Combining vision and language representations for patch-based identification of lexico-semantic relations. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4406–4415.
- Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. *arXiv preprint arXiv:2401.09899*.
- Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.
- Raghendra Kumar, Ratul Chakraborty, Abhishek Tiwari, Sriparna Saha, and Naveen Saini. 2023. Diving into a sea of opinions: Multi-modal abstractive summarization with comment sensitivity. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1117–1126.
- Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.
- Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. *arXiv preprint arXiv:2203.06419*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1739–1749.
- Sabine Molenaar, Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper. 2020. Medical dialogue summarization for automated reporting in healthcare. In *Advanced Information Systems Engineering Workshops: CAiSE 2020 International Workshops, Grenoble, France, June 8–12, 2020, Proceedings 32*, pages 76–88. Springer.

- Giulio Nittari, Ravjyot Khuman, Simone Baldoni, Graziano Pallotta, Gopi Battineni, Ascanio Sirignano, Francesco Amenta, and Giovanna Ricci. 2020. Telemedicine practice: review of the current ethical and legal challenges. *Telemedicine and e-Health*, 26(12):1427–1437.
- OpenAI. 2023a. Gpt3.5. <https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates>.
- OpenAI. 2023b. Gpt4. <https://openai.com/research/gpt-4>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- AS Poornash, Atharva Deshmukh, Archit Sharma, and Sriparna Saha. 2023. Aptsumm at biolaysumm task 1: Biomedical breakdown, improving readability by relevancy based selection. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 579–585.
- Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024a. Unveiling hallucination in text, image, video, and audio foundation models: A comprehensive survey. *arXiv preprint arXiv:2405.09589*.
- Pranab Sahoo, Sriparna Saha, Samrat Mondal, Sujit Chowdhury, and Suraj Gowda. 2022a. Computer-aided covid-19 screening from chest CT-scan using a fuzzy ensemble-based technique. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Pranab Sahoo, Sriparna Saha, Samrat Mondal, and Suraj Gowda. 2022b. Vision transformer based covid-19 detection using chest CT-scan images. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 01–04. IEEE.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024b. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Chanchal Suman, Sriparna Saha, Aditya Gupta, Saurabh Kumar Pandey, and Pushpak Bhattacharyya. 2022. A multi-modal personality prediction system. *Knowledge-Based Systems*, 236:107715.
- Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarba-jeet Tiwari. 2022. Dr. can see: towards a multi-modal disease diagnosis virtual assistant. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 1935–1944.
- Abhisek Tiwari, Anisha Saha, Sriparna Saha, Pushpak Bhattacharyya, and Minakshi Dhar. 2023. Experience and evidence are the eyes of an excellent summarizer! towards knowledge infused multi-modal clinical conversation summarization. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2452–2461.
- Yash Verma, Anubhav Jangra, Raghvendra Verma, and Sriparna Saha. 2023. Large scale multi-lingual multi-modal summarization dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3602–3614. Association for Computational Linguistics.
- Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605.
- Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Douglas Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.
- Sibo Wei, Wenpeng Lu, Xueping Peng, Shoujin Wang, Yi-Fei Wang, and Weiyu Zhang. 2023. Medical question summarization with entity-driven contrastive learning. *arXiv preprint arXiv:2304.07437*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. 2018. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I*, pages 457–466. Springer.

Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 387–394.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

A Appendix

A.1 Dataset and Code

The code of our proposed model and dataset will be shared in this GitHub account <https://github.com/AkashGhosh/From-Sights-to-Insights-Towards-Summarization-of-Multimodal-Clinical-Documents>.

B Qualitative analysis

We conducted more analysis of the samples from the test set. Our exploration revealed that among the various baselines examined, EDI-Summ stood out remarkably in its ability to encapsulate intricate medical concepts. Notably, it excelled in accurately capturing the precise medical terminology derived from the images (highlighted in red), thus providing a more nuanced summary generation.



Image		
Question	Hello, Doctor. I hope you're having a pleasant day. I'm reaching out because I'm quite concerned about my 24-year-old daughter, Sarah. She's been experiencing monthly episodes of pain in her tonsils, accompanied by fever. Please see the attached image. Interestingly, we've noticed that high doses of vitamins seem to provide temporary relief, but the symptoms return each month. Our family has a history of autoimmune disorders. Given this, we're worried about potential underlying causes for Sarah's recurring symptoms. Could you please advise us on what might be causing this and what steps we should take next? Thank you very much for your time and expertise.	Good morning, Doctor. I hope this message finds you well. I'm writing to seek your guidance regarding my father, a veteran who has been experiencing problems in his foot. Please see the image attached. He served in Vietnam and was exposed to Agent Orange during his time there. Additionally, he has a history of anemia and low B12 levels. He also suffers from diabetes. We're concerned that these factors may be contributing to his foot swelling. Could you please advise us on whether Agent Orange exposure, anaemia, and low B12 levels could indeed be related to his symptoms, and what steps we should take next to address this issue? Your expertise would be greatly appreciated.
Gold Response	What is the cause of swollen tonsils , and fever, relieved temporarily by high doses of vitamins?	Can Agent Orange exposure, anaemia, and low B12 levels contribute to foot swelling in a veteran with diabetes?
KM-ClICoN Summation	What could be causing recurring pain in Sarah's tonsils, accompanied by fever	Does Agent Orange exposure, anaemia, low B12 levels, diabetes?
EI-Summ	What could be the cause of in the tonsils, along with fever, and tenderness?	What could be causing persistent issues in foot, anemia with a history of service in Vietnam?
EDI-Summ	What could be the cause of recurrent swollen tonsils on the neck, along with fever, tonsillitis, and tenderness?	What could be the cause of foot swelling , anaemia, low B12 levels, and swollen feet in a patient with a history of service in Vietnam?

Figure 5: Some more examples of summaries generated by our proposed model with respect to various baselines.

B.1 FAQs

1. Why BART is chosen over T5 as the base model? Ans: We chose BART-Base as the foundational model for our proposed method after thorough experimentation. Our extensive evaluations consistently revealed that BART-Base outperforms T5-Base across various metrics and perspectives. We favored BART-Base for its encoder-decoder architecture, allowing the encoding of diverse information before summary generation. Additionally, to align with our available computing resources, we concentrated on the Base versions of all models for our comparative analysis.

2. What are the qualifications of the doctor and the other annotators? How senior are they?

Ans: To uphold ethical standards, the entire process was conducted under the guidance of a senior medical doctor—an additional professor at a government medical college and a co-author of our paper. The team included three medical postgraduate students who were closely supervised by the doctor throughout the entire process.

3. Why do we use cross-attention instead of contextual cross-attention in the decoder of EDI-Summ?

Ans: We use cross-attention instead of contextual cross-attention in the decoder of *EDI-Summ* because while generating the tokens, the sequence length at the decoder increases by one every time we generate a single token. Due to this, we can't calculate contextualized vector as given in Equation 2 because it requires the addition of key and value vectors from text and image, but since the sequence length of text keeps on increasing while generating tokens and the sequence length of the image representation is fixed, we can't add these two vectors.

4. Are the results in our experiments statistically significant?

Ans: We conducted five runs for each of the top-performing text baseline (BART), multimodal baseline (EI-Summ), and our proposed model, *EDI-Summ*, followed by a t-statistical test. The resulting p-value, calculated at 0.004, signifies statistical significance with a confidence level of 95%. Consequently, we concluded that the observed differences are statistically significant.

5. How EDI is different from other multimodal baselines?

Ans: In our paper, EDI-Summ is constructed on the foundation of the BART model, where we

have integrated Multimodal Context-Aware Attention in the encoder and Image Cross Attention in the decoder. The *EI-Summ* model implements image fusion solely in the encoder, a methodology similar to that of Kumar et al. (). We conducted extensive experiments utilizing various pre-trained image embedding techniques, including VGG, ViT, and Resnet, observing performance enhancements compared to our text and other multimodal baselines (*EI-Summ*). Similarly, *EDI-Summ* performs the best in human evaluation metrics too. To the best of our knowledge, our work represents the inaugural attempt at incorporating image fusion in both the encoder and decoder of the BART model.

6. Why only multimodal baselines are used for analysis of results of MMCQS and MM_CliConSummation?

Ans :The insights gained from the experiments on the MMCS dataset allowed us to streamline the experiments for the MMCQS and MM_CliConSummation datasets. For these later datasets, we only experimented with multimodal baselines, as text-based baselines were not a good baseline for our model.

7. Definition of relevance, consistency and fluency in our work.

Ans: Certainly, here's how each criterion would be defined in the context of human evaluation for medical passage summarization:

1. **Fluency:** The smoothness and clarity of the summary, evaluating how effectively the medical information is communicated coherently and understandably, without jargon or ambiguity.

2. **Adequacy:** The accuracy and comprehensiveness of the summary in encapsulating all crucial details and main points of the medical passage, ensuring that no vital information is omitted or misrepresented.

3. **Informativeness:** The degree to which the summary provides valuable and relevant medical insights, assessing its ability to convey essential information succinctly and clearly while avoiding extraneous or redundant details.

4. **Persuasiveness:** The convincing presentation of medical findings or recommendations in the summary, evaluating its ability to influence the reader's understanding or decision-making process regarding the medical topic, potentially leading to further action or consideration.