

HealthAlignSumm: Utilizing Alignment for Multimodal Summarization of Code-Mixed Healthcare Dialogues

Akash Ghosh^{1*}, Arkadeep Acharya^{1*}, Sriparna Saha¹, Gaurav Pandey², Dinesh Raghu², Setu Sinha³

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

² IBM Research, India

³ Indira Gandhi Institute of Medical Sciences, Patna, India

{akash_2321cs19,arkadeep_2101ai41,sriparna}@iitp.ac.in, {gpandey1,diraghu1}@in.ibm.com, drsinhasetu@gmail.com

Abstract

As generative AI progresses, collaboration between doctors and AI scientists is leading to the development of personalized models to streamline healthcare tasks and improve productivity. Summarizing doctor-patient dialogues has become important, helping doctors understand conversations faster and improving patient care. While previous research has mostly focused on text data, incorporating visual cues from patient interactions allows doctors to gain deeper insights into medical conditions. Most of this research has centered on English datasets, but real-world conversations often mix languages for better communication. To address the lack of resources for multimodal summarization of code-mixed dialogues in healthcare, we developed the *MCDH* dataset. Additionally, we created *HealthAlignSumm*, a new model that integrates visual components with the BART architecture. This represents a key advancement in multimodal fusion, applied within both the encoder and decoder of the BART model. Our work is the first to use alignment techniques, including state-of-the-art algorithms like Direct Preference Optimization, on encoder-decoder models with synthetic datasets for multimodal summarization. Through extensive experiments, we demonstrated the superior performance of *HealthAlignSumm* across several metrics validated by both automated assessments and human evaluations. The dataset *MCDH* and our proposed model *HealthAlignSumm* will be available in this GitHub account <https://github.com/AkashGhosh/HealthAlignSumm-Utilizing-Alignment-for-Multimodal-Summarization-of-Code-Mixed-Healthcare-Dialogues>

Disclaimer: This work involves medical imagery based on the subject matter of the topic.

1 Introduction

In India, the glaring disparity in doctor-to-patient ratios among different states and sectors exacerbates the already complex healthcare landscape.

*These authors contributed equally.

This uneven distribution, coupled with the transformative impact of technological advancements and the lingering effects of the COVID-19 pandemic, has led to a significant surge in the adoption of telehealth services (Nittari et al., 2020). As Artificial Intelligence (AI) continues to evolve, there is an increase in collaboration between medical practitioners and AI researchers. Together, they are at the forefront of pioneering and automating various medical procedures. One essential task involves doctors engaging in one-on-one chat conversations with patients to discuss their medical conditions. In this context, effectively understanding previous patient-doctor interactions through dialogue summarization emerges as a critical solution for efficient time management amidst the escalating imbalanced doctor-patient ratio.

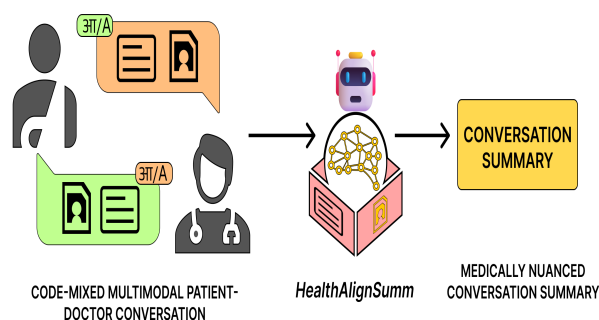


Figure 1: Our model *HealthAlignSumm* takes a textual doctor-patient dialogue and its corresponding image as input and generates the summary of the query that contains information from both the text and the image.

Previous research in this field has mainly focused on using text-only setups for patient-doctor conversations (Joshi et al., 2020),(Molenaar et al., 2020). However, in today’s healthcare landscape, it’s becoming increasingly common for patients to share images of their medical conditions (Sahoo et al., 2024b, 2023), especially when they’re unsure about

medical terms. Visual cues offer clarity in such situations. Recent advancements in visual language models (Ghosh et al., 2024c) have enabled the integration of both text and images for a more refined generation. Studies like (Ghosh et al., 2024a; Sahoo et al., 2024c) demonstrate the effectiveness of multimodality in healthcare applications.

Following this trend, Tiwari et al. (2023) introduced a multimodal dataset for this task of medical dialogue summarization. However, in countries like India, statistics show that people often switch between languages like English and Hindi during conversation for easier communication; a new challenge arises ¹.

Motivation: In our study, we address the gap in healthcare communication by creating *MCDH*, a codemixed Hindi-English dataset reflecting real-world doctor-patient interactions. Since there was no Hindi-English codemixed dataset available for this task, we have leveraged cutting-edge Large Language Models (LLMs) to fill in the gap.

We have created a novel technique called *Hintbag* for generating quality codemixed datasets for this task using a technique called *Hinting*. This dataset aims to enhance communication in healthcare settings by accommodating linguistic diversity. Our work is the first attempt at multimodal codemixed doctor-patient dialogue summarization, especially in the Indian setting. Regarding architecture, existing works like MedSumm (Ghosh et al., 2024b) concatenate image and text vectors for summary generation, whereas KM-CliConSummation (Tiwari et al., 2023) demonstrates the benefits of multimodal fusion into BART’s encoder. Ghosh et al. (2024d) is the most recent work in this domain where they have introduced multimodal attention in both the encoder and decoder of the BART model. However, no work is done in the utilization of recent post-processing alignment techniques like Direct Preference Optimization (DPO) (Rafailov et al., 2024) using synthetically generated preference datasets for this task of multimodal summarization. Our study introduces *HealthAlignSumm* to investigate this approach.

The overall working of *HealthAlignSumm* is shown in Figure 1.

Research Objectives: We have addressed the below research questions in our work:

R1) How does the performance of our proposed

model *HealthAlignSumm* compare to the baselines?

R2) What is the impact of visual cross-attention in the decoder of the BART model? What is the impact on performance when the order of image fusion is altered within the layers of the encoder and decoder?

R3) How much do alignment algorithms like DPO influence the quality of the summaries based on the preference dataset generated by Large Language Model (LLM)?

Contributions: Our research brings forth significant contributions, outlined as follows:

1) This study pioneers the Multimodal Dialog Summarization Task within a Hindi-English codemixed environment, representing a novel initiative in healthcare research. By embracing linguistic diversity, our work sets a precedent for enhanced personalization and seamless communication in healthcare settings.

2) We curated a novel dataset, namely Multimodal Codemixed Dialogue Summarization in Healthcare (*MCDH*), to advance research in this direction. We utilized a novel automated approach, utilizing LLM, to simplify the creation of the Hinglish dataset. Furthermore, we conducted a comprehensive analysis of the dataset to ensure its linguistic quality and suitability for research purposes.

3) We introduce *HealthAlignSumm*, an innovative model that incorporates alignment algorithms like DPO for multimodal summarization. *HealthAlignSumm* demonstrates the effectiveness of using synthetically generated preference datasets, with a large language model (LLM) acting as a judge. This is particularly valuable in healthcare, where obtaining preference datasets, especially in low-resource languages, is challenging. Additionally, we have shown the benefits of using multimodal attention in both the encoder and decoder of the BART model for this task. This enhanced model *HealthSumm* is also used to generate quality synthetic datasets for alignment purposes.

4) We conducted comprehensive human evaluations, complemented by detailed qualitative analyses and risk assessments. This meticulous examination guarantees the safety and reliability of our model’s performance, confirming its suitability for deployment in real-world healthcare applications with utmost confidence.

¹ <https://theconversation.com/the-rise-and-rise-of-hinglish-in-india-53476>

2 Related Works

The following works have been relevant to the following two research areas, namely (a) medical Question Summarization and (b) the role of Multi-modality in Summarization.

Medical Dialogue Summarization: In 2020, Joshi et al. (2020) introduced medical dialogue summarization, optimizing the pointer generator network to capture unique patient history structures and model negations effectively. Song et al. (2020) proposed a hierarchical encoder-tagger model (HET) for generating high-quality summaries by identifying problem statements and treatment recommendations. Chintagunta et al. (2021) developed an algorithm using GPT-3 to create synthetic training data, improving both medical accuracy and coherence. Additionally, Tiwari et al. (2023) introduced multimodal dialog summarization in healthcare, integrating images for better comprehension of conditions.

Multimodal Summarization: Zhu et al. (2020) introduced MSMO, a novel task combining multimodal summarization with multimodal output, employing a dataset, multimodal attention model, and novel evaluation method (MMAE) to effectively generate and assess summaries. Zhu et al. (2020) introduced a multimodal objective function with the guidance of multimodal guidance function to avoid the modality bias problem. Kumar et al. (2023) showed multimodality helps in summarizing news articles. Delbrouck et al. (2021) showed that integrating images leads to better summarization of radiology reports. Ghosh et al. (2024b) marks the pioneering effort in summarizing medical queries by incorporating shared images, enhancing the depth of summary generation for medical professionals.

3 MCDH Dataset

In this study, we used the *MMCliConsumm* dataset (Tiwari et al., 2023), the only publicly available multimodal medical dialog summarization dataset. Each data point includes a medical conversation along with visual cues and their corresponding summaries. Due to the lack of suitable code-mixed healthcare datasets for this task, we created a high-quality synthetic dataset for our study.

We aimed to create synthetic code-mixed datasets that should convincingly mimic real-world doctor-patient conversations. Extensive brainstorming with doctors and linguists revealed that com-

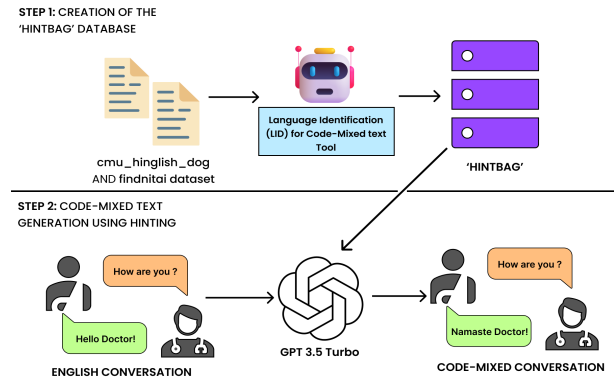


Figure 2: The overall framework for the generation of *MCDH* dataset. We have used prompting using a few shot prompting, which is powered by *Hinting*.

plex medical terms should remain in English, as patients generally understand these better. For instance, "cancer" is preferred over "arbud". Following this, we established a high-level template for our synthetic code-mixed medical data. Linguists annotated around 5% samples, converting English conversations into Hinglish, which were verified by a doctor². Using GPT-3.5 Turbo in a few-shot setting³, we converted the conversations into code-mixed Hindi-English dialogues, leveraging the model's ability to follow language patterns. While initial results were promising, further analysis showed some words were unnaturally forced into Hindi, making the conversations awkward. To address this, we developed a novel prompting technique called *Hintshot*, providing hints on which words should remain in English based on context.

3.1 Framework for developing the *MCDH* Dataset utilizing *Hintshot*

The steps followed in developing the generation of the Hinglish *MCDH* Dataset are mentioned below:

a) We employed two extensive Hinglish conversational datasets, namely *cmu_hinglish_dog* (Zhou et al., 2018) and *findnital (dat)* dataset. Within these datasets, we extracted all English words and compiled them into a set named 'Engbag'. This approach aims to capture English words commonly used in Hindi-English codemixed conversations. To achieve this, we utilized the LID tool.⁴

² We have compensated the volunteers in accordance with their workload, following Government guidelines.

³ We have tried out a few ways of prompting as suggested by Sahoo et al. (2024d) but in our case, few-shot prompting worked best

⁴ LID-tool is an opensource tool from Microsoft for word-level language identification. The link for this:

b) For every data point in *MMClinConSumm*, we have cataloged words used as nouns and verbs, constituting a set called POS (part of speech). We then intersect this set with Engbag, containing English words from codemixed conversations, resulting in a new set named *Hintbag*.⁵

d) After preprocessing, we employed the prompt below to generate the final Hinglish dataset, referred to as the *MCDH* dataset.

The complete framework of our *Hinting* approach for Hinglish dialogue conversion is shown in Figure 2.

Prompt used for codemixed text generation

You are a linguistic expert whose task is to convert the English passages into corresponding Hinglish codemixed ones. <Labelled Examples>: English: {text} Hinglish: {text} . Given the English passage: {text}, convert it into the corresponding Hinglish passage shown in the <Labelled Examples>. Keep all the occurrences of the words present in the set named {Hintbag} in Roman scripts and change the other words to Hindi based on context.

3.2 Analysis of *MCDH* Dataset

We conducted a comprehensive analysis of around 80 samples extracted from the *MCDH* dataset, leveraging the expertise of two linguists proficient in both Hindi and English, who also possess a deep understanding of clinical terminology. They rated *Hinglish* dialogs for fluency and coherence, and we used a code-mixing index (Gambäck and Das, 2014) (CMI) to assess the quality of the synthetic dataset. Fluency and coherence, scored out of 5, are defined in Appendix A.7. *MCDH* dataset samples with Fewshot and *Hintshot* are shown in Appendix A.5.

	Fewshot	Hintshot
Fluency	3.1	3.45
Coherence	3.3	3.5
CMI	25.5	31.5

Table 1: Comparison of Fewshot and *Hintshot* techniques on different metrics

<https://github.com/microsoft/LID-tool>

⁵ We have used GPT-3.5 turbo for the final prompting and extracting the nouns and verbs.

3.3 Statistical analysis of *MCDH* Dataset

Statistics related to *MCDH* are as follows:

1) The dataset comprises 1668 conversations with a total of 5483 utterances. It encompasses a vocabulary of 3512 unique words.

2) The dataset contains 1668 unique images and covers 266 distinct symptoms.

3) The number of diseases covered is 90, and there are a total of 10 medical departments.

4 Methodology

This section elaborates on the working of different modules of our proposed *HealthAlignSumm* model.

4.1 Problem Formulation

In our approach, we address the task of generating nuanced summaries of a multimodal doctor-patient conversation by integrating both textual medical dialogue (M) represented by $M = \{m_0, m_1, \dots, m_n\}$ (where n represents the number of text token in the dialogue exchanges present in the conversation) and visual information from corresponding images (denoted as V). Thus it can be formally represented as:

$$\text{Summary} = f(M = \{m_0, m_1, \dots, m_n\}, V)$$

The construction of our model, termed *HealthAlignSumm*, can be broadly divided into 4 major modules: i) *Modality Representation Module*, ii) *Multimodal contextualized Fusion based Encoder*, iii) *Multimodal Cross-Attention based Decoder*, and iv) *Alignment using DPO*. The architecture and its components are depicted in Figure 3. In the subsequent sections, we discuss each module individually.

4.2 Modality Representation Module

We employed two different types of information to represent the doctor-patient conversation: textual discourse, illustrating the patient’s clinical dialogue, paired with a visual image to enrich contextual understanding. It should be noted that in this study, the text comprises Hindi-English code-mixed utterances in "Romanized" characters (refer to Figure 5 for example). We employ BART tokenizers to segment the text, which is subsequently fed into the BART encoder model for generating contextualized 768-dimensional embeddings. As for images, we utilize existing embeddings generated by ResNet 152, as previously followed by Tiwari et al. (2023).

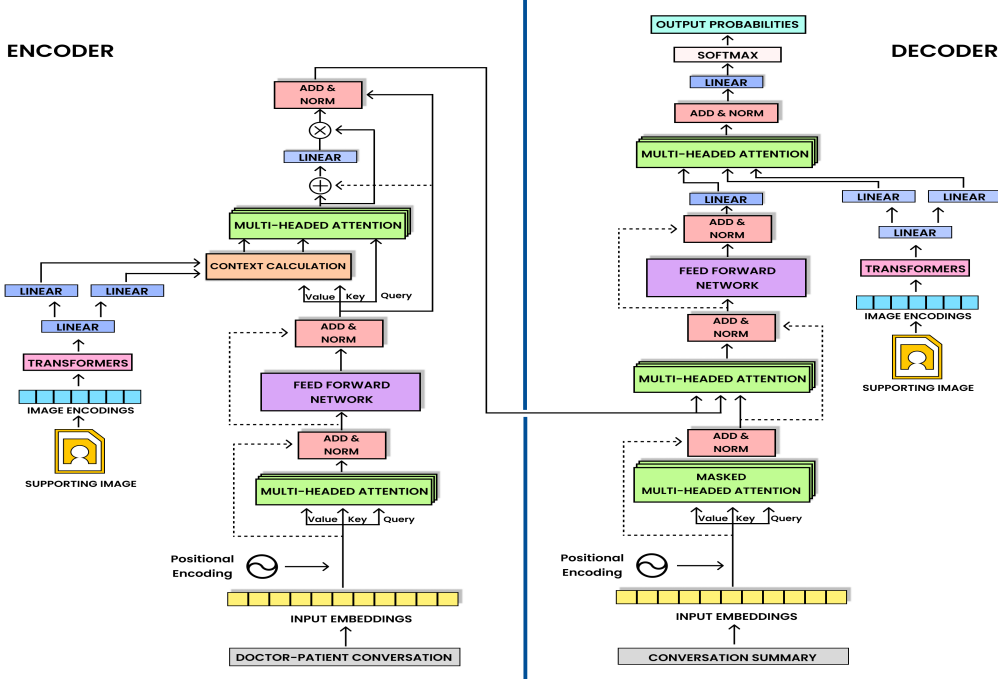


Figure 3: Architecture of our model *HealthSumm*. It consists of a multimodal contextual fusion at the encoder and multimodal cross-attention at the decoder of the BART model.

4.3 Multimodal contextualized Fusion based Encoder

The subsequent points provide a detailed explanation of the components involved in fusing contextual information from images within the Encoder.:

a) Visual Context-Aware Self Attention: The efficacy of the aggregated representation is based upon the efficient and coherent integration of diverse information modalities. Guided by insights from previous studies by Yang et al. (2019) and Kumar et al. (2022), we introduce a novel "multimodal context-aware self-attention" mechanism where we aim to generate conditional key (\hat{K}) and value (\hat{V}) vectors, essential for tailoring attention vectors to each modality. Figure 3 provides a visual representation of this process.

Upon obtaining the intermediate latent representation of the dialogs represented as $Z = \{z_0, z_1, \dots, z_n\}$ from the BART encoder at a specific layer, we compute query (Q), key (K), and value (V) vectors, each with dimensions $\mathbb{R}^{n \times d}$, where d denotes the model dimension (768 for BART encoder), as outlined in Equation 1.

$$[Q, K, V] = Z[W_Q, W_K, W_V] \quad (1)$$

where W_Q, W_K , and W_V represent the learnable parameters associated with the query, key, and value vectors, respectively.

To leverage the latent information inherent in the relationship between textual dialogues and visual cues, we design conditioned key (\hat{K}) and value (\hat{V}) vectors⁶ tailored to both textual and visual contexts. These vectors adapt the query vector as previously obtained from the hidden text representation, producing a unified information vector enriched with visual features. The computation of the key and value pairs is detailed in Equation 2. Additionally, we utilize a transformer-based model to process the visual representation, aligning its sequence length with that of the textual data to enable seamless multimodal fusion.⁷

$$\begin{bmatrix} \hat{K} \\ \hat{V} \end{bmatrix} = (1 - \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix}) \begin{bmatrix} K \\ V \end{bmatrix} + \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} (J \begin{bmatrix} L_k \\ L_v \end{bmatrix}) \quad (2)$$

$$\begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} = \sigma \left(\begin{bmatrix} K \\ V \end{bmatrix} \begin{bmatrix} W_{k1} \\ W_{v1} \end{bmatrix} + J \begin{bmatrix} U_k \\ U_v \end{bmatrix} \begin{bmatrix} W_{k2} \\ W_{v2} \end{bmatrix} \right) \quad (3)$$

where $J \in \mathbb{R}^{n \times d}$ indicates visual representation, L_k and L_v are the learnable parameters. The parameter λ is a learned factor that regulates the extent to which information from the visual modality is preserved. This is computed using Equation 3

⁶ The definition of conditioned keys values are explained in Appendix FAQ section

⁷ We adjust the sequence length of the image, initially set to one, to match the sequence length of the textual data, denoted as n . This is necessary because each textual query corresponds to a single image, hence the sequence length of the image is inherently one.

⁸. $W_{k_1}, W_{k_2}, W_{v_1}$, 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.

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

where M_v represents visual information fused vector. This is represented by context calculation in Figure 3.

b) Modality Fusion Gate: We propose the modality-fusion gate to gain fine-grained control over the flow of visually-enhanced information. The contextual information is transmitted through these gates in accordance with Equation 5:

$$\text{MoFu} = \sigma([Z \oplus Z_v]W_v + b_v) \quad (5)$$

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

$$\hat{H} = Z + \text{MoFu} \odot Z_v \quad (6)$$

The contextualized vector, enriched with image information (\hat{H}), is propagated to the higher layer of the transformer before being directed to the decoder.

4.4 Multimodal Cross-Attention based Decoder

We introduce a visual cross-attention block to the BART decoder to facilitate the inclusion of image-specific nuances during the decoding. Firstly we obtain the image representation by passing the image tokens through a transformer which adjusts the sequence length of the image to match the textual sequence length, resulting in the image representation $I \in \mathbb{R}^{m \times d}$, where m signifies the sequence length at the decoder. The intermediate hidden representation denoted as $L \in \mathbb{R}^{m \times d}$ is fed as input to the cross-attention block. We obtain the query (Q_d) from L , and the key (K_d) and value (V_d) vectors from I , as illustrated in Equation 7.

$$[Q_d, K_d, V_d] = [LW_{Q_d}, IW_{K_d}, IW_{V_d}] \quad (7)$$

The W_{Q_d}, W_{K_d} , and $W_{V_d} \in \mathbb{R}^{d \times d}$ are learnt during model training. Subsequently, we obtain a visually-enhanced representation L via multi-head cross-attention as detailed in Equation 8.

⁸ Here U_k and U_v are learnable parameters.

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

We combine the visually-informed representation L_I with intermediate representation L via a gating mechanism highlighted in Equation 9 and Equation 10.

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

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

$$\hat{L} = L + g_d \odot L_I \quad (10)$$

4.5 Aligning using DPO

The model *HealthSumm* that was developed using multimodal fusion in both the encoder and decoder of the BART is aligned using Direct Preference Optimization (DPO)(Rafailov et al., 2024) to improve the quality of summaries. To apply DPO, we created a synthetic preference dataset using GPT-3.5 turbo, as no existing preference dataset was available. So first for every training data point, we create two responses using the same encoder-decoder multimodal-infused BART model with varying temperatures.

Prompt used for generation of synthetic preference dataset

As a Natural Language Processing Expert, your task is to determine the superior summary of a Hindi-English codemixed conversation {dialogue} between a doctor and a patient. The summary should encompass all the medical concepts discussed in the conversation. Given two options, Option 1 {summary1} and Option 2 {summary2}, select the summary that most accurately captures the essence of the medical conversation {dialogue}. The output summary should be strictly in Roman Script and no additional information should be present apart from one of the mentioned options

Then we used the prompt above to generate the chosen summary and the rejected one using GPT-3.5 Turbo. To ensure the quality of GPT-3.5 Turbo as a judge for creating the preference dataset, we

conducted a human evaluation of the dataset, as detailed in Question 7 of the FAQ section.

This synthetic dataset was used to train our proposed model *HealthAlignSumm* for generating more nuanced summaries.

The DPO loss function, where a large language model (LLM) is used to select the preferred output y_w and less preferred output y_l , can be expressed as:

$$L_{\text{DPO}}(\pi_{\theta}; \pi_{\text{LLM}}) = -\mathbb{E}_{(x, y_w, y_l) \sim D_{\text{LLM}}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{LLM}}(y_w | x)} - \beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{LLM}}(y_l | x)} \right) \right] \quad (11)$$

where:

- $L_{\text{DPO}}(\pi_{\theta}; \pi_{\text{LLM}})$ is the DPO loss function.
- $\mathbb{E}_{(x, y_w, y_l) \sim D_{\text{LLM}}}$ is the expectation over the dataset D_{LLM} , where y_w is the preferred output and y_l is the less preferred output based on the LLM’s judgment.
- $\pi_{\theta}(y_w | x)$ and $\pi_{\theta}(y_l | x)$ are the probabilities assigned by the model’s policy π_{θ} to the preferred and less preferred outputs, given the input x .
- $\pi_{\text{LLM}}(y_w | x)$ and $\pi_{\text{LLM}}(y_l | x)$ are the probabilities assigned by the LLM to the preferred and less preferred outputs, respectively, given the input x .
- β is a scaling factor that adjusts the sharpness of preference between the preferred and less preferred outputs.
- σ is the sigmoid function, defined as $\sigma(z) = \frac{1}{1+e^{-z}}$, which scales the difference in log probabilities between 0 and 1, indicating the likelihood that y_w is preferred over y_l .

The overall pipeline of our proposed approach is shown in Figure 4.

5 Experimental Results and Analysis

In the subsequent section, we discuss the experimental framework and assess the effectiveness of the proposed model *HealthAlignSumm* using a wide range of evaluation metrics concerning various baselines, encompassing both automated and

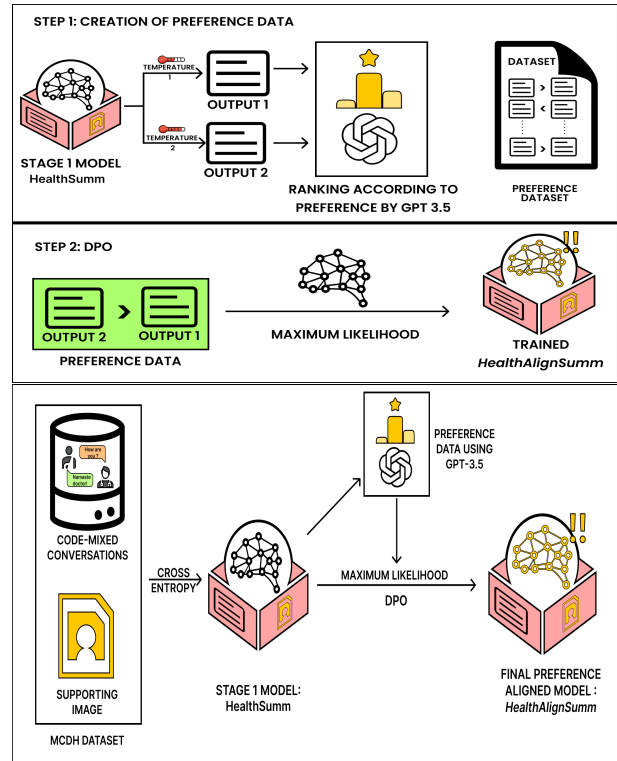


Figure 4: Our proposed model, *HealthAlignSumm*, operates through a comprehensive pipeline comprising two key stages. Initially, we train HealthSumm, followed by its utilization in aligning with the synthetic preference dataset to generate the output summaries.

human-centric assessments. Furthermore, we conduct a qualitative examination of the generated summaries across different model configurations.⁹

5.1 Experimental Setup

During our experimentation, we employed an RTX 3090 GPU, with each model completing its runtime within an average window of 30-40 minutes. Our proposed model *HealthAlignSumm* uses BART as its foundation model¹⁰. Dataset partitioning involved creating training, validation, and test sets in an 80:5:15 ratio. The execution of models was guided by a set of hyperparameters, as outlined in Table 2.

To thoroughly evaluate our proposed model’s effectiveness, we conducted an in-depth analysis. This involved exploring various configurations of multimodal fusion within both the encoder and decoder components of the model. Additionally, we compared the performance against both

⁹ The qualitative analysis of the generated summaries has been added in the appendix section.

¹⁰ The reason for choosing BART model for *HealthAlignSumm* has been explained in the FAQ section of Appendix.

Hyperparameters	Value
Maximum epochs	30
Maximum Sequence Length	360
Visual embedding size	786
Optimizer	Adam
Learning rate	5e-05
Rank	16
Lora Alpha	16
Lora dropout	0.06
Target Modules	Q,K,V,up, down,gate

Table 2: Hyperparameters used in *HealthAlignSumm*

textual and multimodal baseline models to provide a comprehensive assessment. We selected Llama-2 (Touvron et al., 2023), m-BART-base (Liu et al., 2020), BART-base (Lewis et al., 2019), and T5-base (Raffel et al., 2020) as our unimodal baselines. Additionally, for multimodal baselines specifically designed for summarization tasks, we chose GPT-4v (OpenAI, 2024), Gemini-1 pro vision (Team et al., 2023), MedSumm (Ghosh et al., 2024b), EDI-Summ (Ghosh et al., 2024d) and KM-ClinConSummation (Tiwari et al., 2023) as they are proposed for the task of multimodal summarization. MedSumm, KM-CliConSummation and EDI-Summ¹¹ are the proposed models for the task of multimodal summarization in Ghosh et al. (2024b), Tiwari et al. (2023) and Ghosh et al. (2024b), respectively.¹²

To study the effectiveness of our *HealthAlignSumm* and the influence of alignment, we have also considered the case where DPO alignment is not included, and we refer to that model as *HealthSumm*, a key focus of our analysis. Notably, to the best of our knowledge *HealthAlignSumm* is the first work where the DPO Alignment has been performed on the task of multimodal summarization. Furthermore, we explored an configuration where multimodal fusion is done in both encoder and decoder namely *HealthSumm* and also where its implemented only in the encoder, termed *HealthSumm(encoder)*.

To evaluate the performance of generated summaries, we utilized automated metrics like ROUGE (Lin, 2004) scores, namely, ROUGE-1, ROUGE-2,

¹¹ Both EDI-Summ and *HealthSumm* has almost similar architecture. In the above we have used the pretrained version of EDI-Summ.

¹² GPT-4V and Gemini pro-vision 1.0 pro are used in few shot setting while all the other baselines are finetuned end to end. *HealthSumm* is finetuned without LORA. But *HealthAlignSumm* is finetuned using DPO and used LORA adapters for finetuning.

and ROUGE-L scores alongside human evaluation metrics. For human evaluation, we partnered with a healthcare professional and engaged volunteers from the medical community. Our methodology involves evaluating three distinct and medically intricate metrics: clinical assessment score, Factual Recall (Abacha et al., 2023) and Omission Rate (Abacha et al., 2023)

5.2 Automated Evaluation

Table 3 delivers comprehensive findings that illuminate the performance of various models, showcasing the effectiveness of various strategies for this task of multimodal clinical query summarization.

R1) Comparison with baselines: From Table 3 we can conclude among textual baseline models, T5 base exhibited the poorest performance, with Llama-2 emerging as the top performer. In terms of multimodal baselines, EDI-Summ surpassed MedSumm in all metrics. But our proposed model *HealthAlignSumm* comes out as the best performer.

R2) Impact of Decoder Visual Cross Attention : Through meticulous experimentation, we have determined that while the addition of decoder attention does yield some enhancements over models employing only encoder fusion these improvements do not reach statistical significance. To further investigate the impact of encoder-decoder fusion in *HealthSumm*, we conducted an ablation study present in Appendix section A.3 by varying the modality fusion order in both the encoder and decoder of the *HealthSumm*.¹³

R3) Impact of DPO in the generated Summaries: In our automated tests, we saw a big improvement in performance after using DPO for alignment. This shows that synthetic preference datasets can be really helpful for alignment, especially in areas with limited data like low-resource healthcare. Also, our detailed analysis, which is shared in Appendix section A.4, shows that DPO has greatly reduced the chances of hallucinations.

5.3 Human Evaluation

A team of medical students, under the guidance of a doctor(who is also a co-author of the paper), conducted the human evaluation on 80 data samples (around 35% of the test samples) randomly selected for this purpose. The evaluation metrics

¹³ It is to be noted that (Ghosh et al., 2024d) findings are also on the same line .

Model	ROUGE-1	ROUGE-2	ROUGE-L
mBART-base	31.25	14.33	25.18
T5-base	40.21	16.31	29.20
BART-base	40.36	19.71	31.9
LLAMA-2	42.14	22.75	29.88
GPT-4v	45.97	20.10	34.76
Gemini-1.0 pro-vision	47.68	24.13	37.85
MedSumm	48.61	23.9	37.24
EDI-Summ	48.91	24.6	37.95
KM-CliConSummation	54.70	31.00	44.96
<i>HealthSumm(encoder)</i>	54.14	30.70	44.65
<i>HealthSumm</i>	54.70	31.00	44.96
<i>HealthAlignSumm</i>	60.2	38.54	50.46

Table 3: Performances of different models for multi-modal codemixed dialog summarization task on *MCDH* dataset. ***HealthAlignSumm*** performance is superior to all the baselines and versions of *HealthSumm*. Among unimodal baselines, LLAMA works best, and KM-CliConSummation achieved the best results among multimodal baselines. The best results for each subsection are shown in bold.

Model	Factual Recall	Omission Rate	Clinical-Eval Score
BART	0.68	0.38	3.31
<i>HealthSumm</i>	0.78	0.26	3.75
<i>HealthAlignSumm</i>	0.85	0.21	4.12
Annotated Summary	0.96	0.08	4.74

Table 4: Human Evaluation of our proposed ***HealthAlignSumm*** concerning various baselines and golden summaries in different human evaluation metrics

comprised: **Clinical Evaluation Score**, wherein ratings ranging from 1 (poor) to 5 (good) were provided by doctors and the team to assess summaries based on overall relevance, consistency, fluency, coherence and level of hallucination (Sahoo et al., 2024a); and **Medical Fact-Based Metrics**, such as Factual Recall (Abacha et al., 2023) and Omission Rate metrics (Abacha et al., 2023), which gauged the accuracy of the generated summary in capturing medical facts compared to the gold standard annotated summary. Table 4 illustrates the results of ***HealthAlignSumm***, demonstrating its significant outperformance of all baselines in the chosen human evaluation metrics. This underscores the critical role of alignment using synthetically generated data.

6 Conclusion

This study presents a new task called Multimodal Clinical Dialogue Summarization in a mix of Hindi and English, marking the first attempt to work with codemixed languages in the healthcare field. We present a novel approach utilizing *Hinting* to generate Hinglish text and introduce the *MCDH* dataset. Our proposed pipeline architecture, ***HealthAlignSumm***, demonstrates superior performance over

all baselines in both human and automated evaluation metrics. The findings highlight the efficacy of synthetically generated datasets for alignment, resulting in a more refined and nuanced summary generation.

7 Limitations

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

1) With a modest 1668 samples, our dataset forms the foundation. However, before the real-world deployment of the model, a crucial task lies in its expansion. This expansion aims to incorporate a wider range of medical conditions tailored to the demographic characteristics of the target deployment region.

2) Currently, our dataset is limited to English and Hindi-English codemixed languages. This hinders the deployment of our model in regions where Hindi and English are not commonly understood. To address this constraint, our future efforts will focus on expanding our dataset to include more multi-lingual languages, catering to the diverse linguistic needs of various regions for the development of personalized healthcare models.

3) While our current model is designed to focus on images as an additional modality, we acknowledge the dynamic landscape of data-sharing practices. Particularly in the medical domain, we observe a rising prevalence of videos and voice recordings. In light of this evolution, our future strategy entails expanding the dataset to encompass these additional modalities.

8 Ethics Statement

Our project utilized the *MCDH* dataset, which is a refined version of the publicly available *MMCLI-ConSumm* dataset. Throughout the process, we collaborated with a medical expert, who is also a co-author of this paper, to ensure accuracy and quality control at every stage, from data collection to validation. To uphold ethical standards, we compensated all volunteers in alignment with India’s minimum wage regulations. Privacy was paramount in our approach, and we diligently ensured the dataset was free of any images that could compromise individuals’ privacy. Additionally, we are currently seeking approval from the Institutional Review Board (IRB) to further reinforce the ethical integrity of our work. Notably, our proposed model is exclusively designed for summarization tasks, deliberately avoiding any predictive functions that could have unintended consequences for users. This decision reflects our deep commitment to ethical principles and responsible use of AI technologies.

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

- Hugging face datasets library. Accessed: April 16, 2024.
- 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*.
- 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.
- Björn Gambäck and Amitava Das. 2014. On measuring the complexity of code-mixing. In *Proceedings of the 11th international conference on natural language processing, Goa, India*, pages 1–7.
- 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*.
- Akash Ghosh, Mohit Tomar, Abhishek Tiwari, Sriparna Saha, Jatin Salve, and Setu Sinha. 2024d. From sights to insights: Towards summarization of multimodal clinical documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13117–13129.
- 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*.
- Raghvendra 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.
- 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.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and

- Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- 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. 2024. [Gpt-4v system card](#). Accessed: 2024-06-09.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- 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. *Journal of machine learning research*, 21(140):1–67.
- 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, Manjeevan Seera, Saksham Kumar Sharma, and Manish Kumar. 2023. Enhancing computer-aided cervical cancer detection using a novel fuzzy rank-based fusion. *IEEE Access*.
- Pranab Sahoo, Saksham Kumar Sharma, Sriparna Saha, Deepak Jain, and Samrat Mondal. 2024b. A multi-stage framework for respiratory disease detection and assessing severity in chest x-ray images. *Scientific Reports*, 14(1):12380.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Aman Chadha, and Samrat Mondal. 2024c. Enhancing adverse drug event detection with multimodal dataset: Corpus creation and model development. *arXiv preprint arXiv:2405.15766*.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024d. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.
- Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- 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.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrusti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Aditya Vavre, Abhirut Gupta, and Sunita Sarawagi. 2022. Adapting multilingual models for code-mixed translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7133–7141.
- Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *arXiv preprint arXiv:2103.13309*.
- 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.
- Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.
- Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9749–9756.

A Appendix

A.1 Risk Analysis

Through an extensive array of automated, human, and qualitative assessments, we’ve determined that incorporating visual cues significantly enhances the depth of clinically nuanced summaries. However,

in critical sectors such as healthcare, the potential for misinformation or misinterpretation poses substantial risks. Our human evaluation revealed instances where the model, influenced by visual data, overlooked essential keywords within textual queries. Similarly, qualitative analysis identified instances of hallucinations in the model’s output. Yet, employing DPO alignment has notably mitigated the occurrence of hallucinated information in the final summaries. Moreover, we’ve observed that image quality plays a crucial role in generating accurate summaries, as poor quality images can introduce noise into the output. Consequently, while *HealthAlignSumm* enhances the generation of nuanced summaries, it’s imperative to involve medical professionals in high-stakes scenarios. Our model serves as an aid rather than a substitute for the expertise of experienced physicians.

A.2 Analysis of various infusion orders across layers of encoder-decoder of BART:

To further investigate the impact of encoder-decoder fusion in *HealthSumm*, we conducted an ablation study by varying the modality fusion order in both the encoder and decoder of the *HealthSumm* model as shown in Table 5. Our experiments revealed that optimal performance is achieved when fusion occurs at layer 3 of the encoder and layer 4 of the decoder.

Encoder layer	Decoder layer	ROUGE-1	ROUGE-2	ROUGE-L
2	2	53.57	28.58	43.09
3	3	54.55	29.43	43.76
3	4	54.70	31.00	44.96
4	4	54.02	28.45	42.66

Table 5: Evaluation of the proposed *HealthSumm* model under diverse scenarios involving variations in the infusion order of modalities. The best results are obtained when infusion is done at layer 3 of the encoder and layer 4 of the decoder.

A.3 Qualitative Analysis of Generated Summaries

In Figure 5, we compare the gold summary with those generated by the *HealthSumm* and *HealthAlignSumm* models. Our analysis reveals that the *HealthSumm* model often overlooks valuable information. For instance, in the example shown at the top, it fails to include the medical term concept "Acariasis" in the summary, whereas *HealthAlignSumm* includes it. Further-

more, in example-1, *HealthSumm* erroneously adds the word "Dermatitis" to its summary, which is absent in the golden summary. Similar patterns are evident in example at the bottom. The hallucinated text in both examples is highlighted in red.

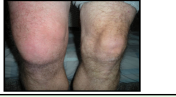
DOCTOR-PATIENT CONVO	
	<p>Patient: Good morning doctor, mujhe skin rash ho gaya hai. Mere paas agli hafte ek swimming competition hai aur mujhe dar hai ki main use miss kar dunga/n</p> <p>Doctor: Main apki pareshani samajh sakta hoon. Lekin kya aapko skin me itching ho rahi hai?/n</p> <p>Patient: Bilkul sahi kaha, doctor.</p> <p>InDoctor: Kya aapke andkosh me sujan hai?/n</p> <p>Patient: Aap bilkul sahi samajh rahe hai/n</p> <p>Doctor: Kya aapko skin me lesion hai?/n</p> <p>Patient: Haan, mujhe haan/n</p> <p>Doctor: Kya aapko ulti jaisa lag raha hai?/n</p> <p>Patient: Sahi hai, doctor./n</p> <p>Doctor: Symptom ke hisab se, aapko <MASK> ho sakta hai.</p>
TARGET SUMMARY	
The conversation is between a patient and a doctor about the patient's skin rash. (skin rash) .The patient is concerned about possibly missing a swimming competition because of it. The doctor asks about symptoms and determines that the patient may have acariasis, which causes itching, swelling, skin lesions, and possible vomiting.	
Summary with encoder-decoder attention without alignment	Summary with encoder-decoder attention along with DPO alignment
The patient is suffering from skin rash and is worried about missing a swimming competition. The doctor asks about other symptoms such as itching, swelling, skin lesion, and ultimatium . Based on the symptoms, the doctor diagnoses the patient with dermatitis due to sun exposure.	The conversation is between a patient and a doctor regarding the patient's skin rash. The patient is concerned about missing a swimming competition because of it. The doctor asks about other symptoms such as itching, swelling, and possibly a skin lesion. Based on the symptoms, the doctor suggests that the patient may have Acariasis.
DOCTOR-PATIENT CONVO	
	<p>Patient: Hi doctor, mujhe ghutno mein stiffness hai. Ise hilane mein dard hota hai, yeh kyun ho raha hai?/n</p> <p>Doctor: Yeh ek painful experience ho sakta hai. Kya aapke ghutne mein dard bhi hai?/n</p> <p>Patient: Haan doctor, mujhe ghutno mein dard hai. Yeh mujhe sidhe chalne mein mushkil kar raha hai/n</p> <p>Doctor: Main apki takleef samajh raha hoon. Kya aap ghutno mein kamzori bhi mehsoos kar rahe hai?/n</p> <p>Patient: Haan, mujhe ghutno mein kamzori mehsoos hoti hai/n</p> <p>Doctor: Mujhe lagta hai aapko <MASK> hai.</p>
TARGET SUMMARY	
The patient has knee stiffness and pain, as well as difficulty moving through stairs and weakness in the knee. The doctor suspects that the patient have Chondromalacia of the patella.	
Summary with encoder-decoder attention without alignment	Summary with encoder-decoder attention along with DPO alignment
The conversation is between a patient and a doctor regarding the patient's knee lump or mass . The doctor asks about other areas of pain and determines that the patient may have Chondromalacia of the patella.	The conversation between the patient and the doctor revolves around the patient's knee pain and its possible causes. The doctor provides a diagnosis of Chondromalacia of the patella based on the given symptoms.

Figure 5: Analysis of summaries generated by our proposed model concerning various baselines.

A.4 Samples from MCDH with fewshot and Hintshot

To better compare the influence of *Hintshot* we have shown below samples of MCDH with fewshot and Hintshot in Table-6.

A.5 FAQs

1) Why BART was chosen as the base model?

Ans: This study delves into exploring the influence of multimodal cross-attention within encoder-decoder architectures. We excluded LLAMA-2, a decoder model, due to its lack of support for this feature. Our experimentation revealed that both the T5-base and BART-base showcased similar performance for our task. Given BART’s lighter weight compared to T5, we selected BART-base as the foundational model for constructing our *HealthAlignSumm* framework.

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

Ans: To ensure adherence to ethical guidelines, oversight was provided by a seasoned medical practitioner, an associate professor of medicine affiliated with a government medical college, who also

Fewshot	Hintshot
<p>Patient: Doctor, mujhe aankhon mein spots ya clouds najar aate hain. Maine kuch bhi saaf nahi dekh pa raha hoon. Mujhe nirash hone ka ehsaas hai. Doctor: Mai aapki bhaavanaon ko samajh sakta hoon. Mujhe aapki bimari ka diagnosis karne ke liye kuch sawaal puchne honge. Kya aapki drishti kamjor ho gayi hai? Patient: Sach hai, mere vision kamjor ho gayi hai. Doctor: Kya aapko aankhon ke lakshan hai? Patient: Haan, mujhe aankhon ke lakshan hai. Doctor: Kya aapko aankh mein dard mehsoos hota hai? Patient: Zaroor, mujhe aankh mein dard hai. Doctor: aapke lakshan batate hain ki aapko <MASK> hai.</p>	<p>Patient: Doctor, mujhe aankhon mein spots ya clouds dikhte hain. Main kuch clear dekh nahi paa raha hoon. Main helpless feel kar raha hoon. Doctor: Main samajh sakta hoon aap kaise feel kar rahe hain. Mujhe aapki condition samajhne ke liye kuch questions puchhne hain. Kya aapki vision weak ho gayi hai? Patient: Haan, mera vision weak ho gaya hai. Doctor: Kya aapko aankhon ke aur koi symptoms hain? Patient: Haan, mujhe aur bhi symptoms hain. Doctor: Kya aapko aankhon mein pain hota hai? Patient: Haan, mujhe aankhon mein pain hota hai. Doctor: Aapke symptoms se lagta hai ki aapko <MASK> ho sakta hai.</p>
<p>Patient: Doctor, mujhe aankhon mein spots ya clouds najar aate hain. Maine kuch bhi saaf nahi dekh pa raha hoon. Mujhe nirash hone ka ehsaas hai. Doctor: Mai aapki bhaavanaon ko samajh sakta hoon. Mujhe aapki bimari ka diagnosis karne ke liye kuch sawaal puchne honge. Kya aapki vision kamjor ho gayi hai? Patient: Sach hai, mere vision kamjor ho gayi hai. Doctor: Kya aapko aankhon ke lakshan hai? Patient: Haan, mujhe aankhon ke lakshan hai. Doctor: Kya aapko aankh mein dard mehsoos hota hai? Patient: Zaroor, mujhe aankh mein dard hai. Doctor: aapke lakshan batate hain ki aapko <MASK> hai.</p>	<p>Patient: Doctor, mujhe aankhon mein spots ya clouds dikhai dete hain. Main kuch bhi clearly nahi dekh pa raha hoon. Mujhe helpless feel ho raha hai. Doctor: Main samajh sakta hoon aap kaisa feel kar rahe hain. Mujhe aapki condition diagnose karne ke liye kuch questions puchhne padhenge. Kya aapki vision weak ho gayi hai? Patient: Haan, mera vision weak ho gaya hai. Doctor: Kya aapko aankhon ke aur koi symptoms hain? Patient: Haan, mujhe aur bhi symptoms hain. Doctor: Kya aapko aankhon mein pain hota hai? Patient: Haan, mujhe aankhon mein pain hota hai. Doctor: Aapke symptoms se lagta hai ki aapko <MASK> ho sakta hai.</p>

Table 6: Samples from MCDH with Fewshot and Hintshot

served as a co-author on our paper. The research team comprised three volunteers who received diligent supervision from the aforementioned doctor throughout the entirety of the project.

3) Why do we use cross-attention instead of contextual cross-attention in the decoder of *HealthSumm*?

Ans: In *HealthSumm*, we opt for cross-attention over contextual cross-attention in the decoder. This choice stems from the dynamic nature of token generation in the decoder, wherein the sequence length increases with each token generated. Consequently, the calculation of contextualized vectors, as outlined in Equation 2, becomes unfeasible. This is due to the requirement of adding key and value vectors from both text and image, a task hindered by the increasing sequence length of the text tokens juxtaposed with the fixed length of the image

representation.

4) What is the significance of this study in comparison to previous works for this task?

Ans: In the realm of healthcare, the exploration of datasets within Indic settings remains relatively limited. Thus, our dataset presents an opportunity for pioneering research, potentially paving the way for further exploration in other underrepresented low-resource Indic languages. Regarding our model, *HealthAlignSumm*, its novelty rests on two key fronts. Firstly, while prior endeavors predominantly focus on multimodal fusion within the encoder of the BART model, we pose the question of whether decoder attention contributes significantly to our task. Secondly, to our knowledge, our work stands as the first to showcase the efficacy of alignment through synthetically generated data in the domain of summarization.

5) Why multilingual pre-trained language models like mBART was not chosen as the base model instead of BART?

Ans: In this work, we considered the input text to be code-mixed (Hinglish) but in Roman script so Hindi language tokens will not help. Please refer to the examples in Figure 5. In multilingual contexts, languages are not mixed within a single context. However, in code-mixed contexts, the speaker freely mixes languages within the same context. MBART is not inherently built to handle code-mixed languages and requires adaptation or fine-tuning to effectively manage code-mixed text. Previous studies suggest the same Vavre et al. (2022), Winata et al. (2021). To make sure the same thing holds true for our usecase we have fine-tuned mbart on MCDH dataset with the same hyperparameters as BART and we got the results as shown in Table 5 which suggests BART is a better model than mBART for our usecase.

6) Are the results in Table 3 statistically significant?

Ans: Five runs were conducted for each of the finetuned models. Statistical testing revealed a p-value of 0.004, with a confidence level of 95%. Thus, the observed findings are statistically meaningful.

7) How did we make sure the synthetic dataset used for training HealthAlignSumm is of good quality?

Ans: We evaluated 300 samples from the DPO positive and negative samples before running the final DPO model, *HealthAlignSumm*. Our evaluation indicates that in approximately 94% of cases, GPT-3.5 correctly judged the positive summary. In 4% of cases, the two summaries were so close that human evaluators deemed both could be considered positive. In a few samples (around 2%), there was a contradiction between the human evaluator and GPT-3.5. In the last two days, we evaluated another random 80 samples and observed similar trends. Our evaluation suggests that GPT-3.5 is highly reliable in distinguishing between good and bad summaries, contributing to the improved performance of *HealthAlignSumm*. We will include these details in the final camera-ready version.

8) Why was HealthAlignSumm built starting from BART instead of using a clinical model like ClinicalT5?

Ans: In our experiments, we found BART to perform better than T5 for clinical summarization,

which is why we did not use T5 or ClinicalT5 initially. For completeness, we conducted experiments with ClinicalT5, and the results were as follows:

- **ClinicalT5:** ROUGE-1: 39.8, ROUGE-2: 15.7, ROUGE-L: 29.8
- **HealthAlignSumm (with ClinicalT5):** ROUGE-1: 57.4, ROUGE-2: 36.62, ROUGE-L: 47.2

From these results, we conclude that ClinicalT5 does not offer any additional advantage for this task. We believe one possible reason is that the MCDH dataset is code-mixed, so the inherent clinical knowledge of ClinicalT5 does not provide any significant benefit.

9) What do you mean by conditioned key and value?

Ans: Here, the conditioned key (\hat{K}) and value (\hat{V}) refer to learnable parameters that depend on another set of parameters, represented as $[\lambda_k \lambda_v]$. These parameters, λ_k and λ_v , control the contextual attention between different modalities of information.

A.6 Parameters for human evaluation for synthetic MCDH dataset

We have collaborated with two linguists who are familiar with clinical terminology. Upon discussion with the doctor who is also the coauthor of the work we have decided two factors on which the generated text will be evaluated. The parameters for human evaluation of synthetic MCDH dataset are shown below:

- 1) The overall text makes sense as a whole. This we call as coherence.
- 2) Ease of understanding without awkward transitions between languages. This we call as fluency.

A.7 Statistical analysis of MCDH Dataset

The distribution of lengths of input dialogs and the output summary are shown in Figure 7 and in Figure 6.

5) The word cloud distribution of dialogues is shown in Figure 8 and the word cloud distribution of summaries is shown in Figure 9.

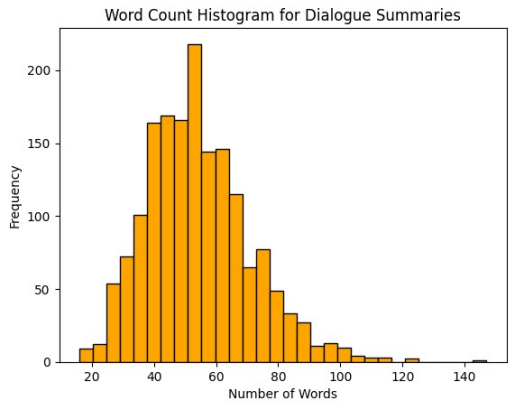


Figure 6: Distribution of Words in Summaries

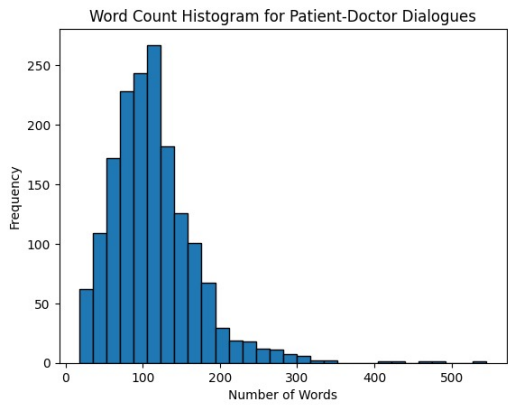


Figure 7: Distribution of Words in Dialogs

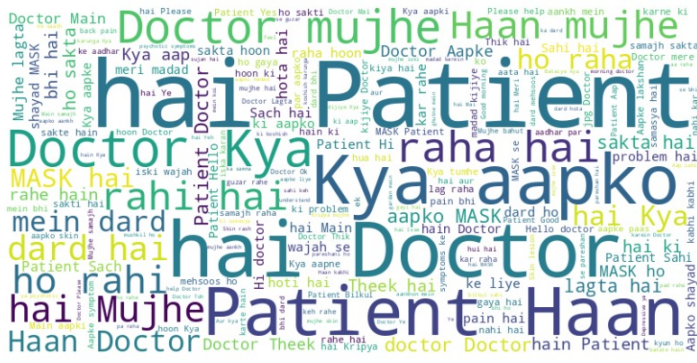


Figure 8: Word Cloud for Dialogs

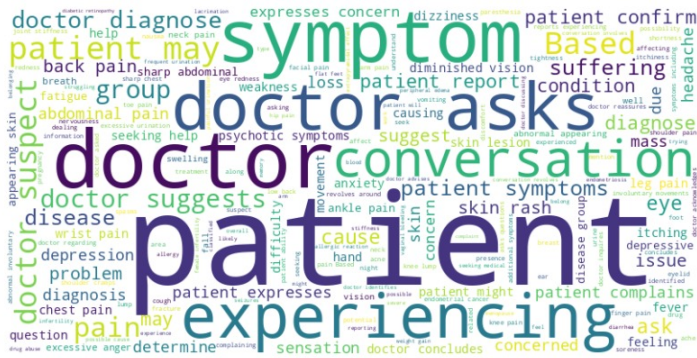


Figure 9: Word Cloud for Summaries