Experience and Evidence are the eyes of an excellent summarizer! Towards Knowledge Infused Multi-modal Clinical Conversation Summarization

Abhisek Tiwari
Indian Institute of Technology, Patna
Patna, India
abhisek_1921cs16@iitp.ac.in

Anisha Saha
Indian Institute of Technology, Patna
Patna, India
anisha0325@gmail.com

Sriparna Saha
Indian Institute of Technology, Patna
Patna, India
sriparna@iitp.ac.in

Pushpak Bhattacharyya
Indian Institute of Technology, Bombay
Bombay, India
pb@cse.iitb.ac.in

Minakshi Dhar
All India Institute of Medical Sciences, Rishikesh
Rishikesh, India
minakshi.med@aiimsrishikesh.edu.in

ABSTRACT

With the advancement of telemedicine, both researchers and medical practitioners are working hand-in-hand to develop various techniques to automate various medical operations, such as diagnosis report generation. In this paper, we first present a multi-modal clinical conversation summary generation task that takes a clinician-patient interaction (both textual and visual information) and generates a succinct synopsis of the conversation. We propose a knowledge-infused, multi-modal, multi-tasking medical domain identification and clinical conversation summary generation (*MMCliConSummation*) framework. It leverages an adapter to infuse knowledge and visual features and unify the fused feature vector using a gated mechanism. Furthermore, we developed a multi-modal, multi-intent clinical conversation summarization corpus annotated with intent, symptom, and summary. The extensive set of experiments, both quantitatively and qualitatively, led to the following findings: (a) critical significance of visuals, (b) more precise and medical entity preserving summary with additional knowledge infusion, and (c) a correlation between medical department identification and clinical synopsis generation. Furthermore, the dataset and source code are available at https://github.com/NLP-RL/MMCliConSummation.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Applied computing** → *Health care information systems.*

KEYWORDS

Multimodal Medical Dialogue Summarization, Online Counselling, Text Generation, Multimodal Infusion

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614870

ACM Reference Format:
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 (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3583780.3614870

1 INTRODUCTION

In the past few years, tele-health has grown immensely with the advancement of information & communication technologies (ICTs) and artificial intelligence-based applications for healthcare activities [19]. With the COVID-19 pandemic, internet utilization for healthcare activities has reached its peak in the last two decades and has become a new normal [31]. On the other hand, many recent healthcare surveys and the World Health Organisation (WHO) found an uneven doctor-to-population ratio, estimating a deficit of more than 12 million healthcare workers by 2030. Thus, tele-health usage is being actively encouraged by healthcare providers, and patients are adopting it at the same pace [28]. One such manifestation that has become popular in both research and industry communities is automatic disease diagnosis (ADD) [28]. ADD aims to assist doctor by conducting primary symptom and sign investigations, allowing them to focus on diagnosis and treatment. A few hospitals have already implemented diagnosis assistants like Ada[1], and Mayo Clinic[2] for clinical assistance.

When we consult with doctors, they often conduct a preliminary investigation by analyzing the patient's self-report and investigating other pertinent symptoms and signs. Even though, they may require some lab reports to confirm a medical condition, the initial investigation helps to decide on some lab tests. With this motivation, Wei et al. [30] formulated a conversational artificial intelligence-based symptom investigation and diagnosis assistant. In our conversations, we often show our visual medical conditions, such as skin rash, to doctors for precise diagnosis and accurate treatment. Driven by the motivation, Tiwari et al. [26] first proposed a multi-modal automatic disease diagnosis virtual assistant

[1]https://ada.com/
[2]https://www.mayoclinic.org/

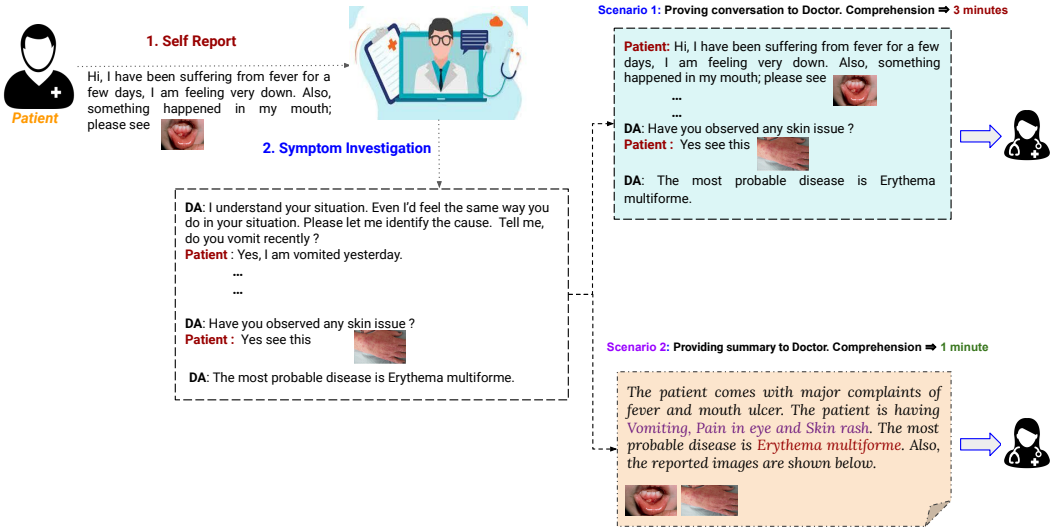


Figure 1: An illustration of an autonomous symptom investigation and disease diagnosis assistant with and without conversation summary generation. The second scenario is evidently more comprehensible and time-efficient.

called MDD-VA, which demonstrated the critical impact of considering the visual form of symptom reporting on diagnosis efficacy and end-user satisfaction. The diagnosis dialogue framework is illustrated in Figure 1 (left side). Certainly, it aids doctors by automatically collecting primary investigations. However, the diagnosis assistant forwards the entire conversation to doctor, which takes a significant amount of time to comprehend the same (Figure 1, right top). On the other hand (Figure 1, right bottom), a summary of the dialogue has been provided, which took significantly less time to comprehend the case. Furthermore, the summary could be utilized for storing the cases efficiently and thus enhancing its re-usability.

It is often stated that an image conveys an idea more effectively than thousand words. In today's digital media era, we are compelled to use images and visuals in our daily lives. We frequently rely on visuals in conversations, especially in clinical discussions, where we aim to convey medical conditions accurately. A recent survey conducted by Popov et al. [21] highlighted the substantial market value of medical imaging, which reached 32 billion in 2022 and is projected to grow at a compound annual growth rate of 4.9%. This clearly underscores the importance of images and visuals in clinical dialogue settings. In real-life scenarios, if two individuals without medical expertise were asked to summarize two medical dialogues, the older individual would likely perform better. The key differentiating factor is the additional knowledge possessed by the older person. A corpus serves as a representation of behavior, and therefore, learning from a corpus, along with additional relevant knowledge, incorporates a global perspective [27]. Hence, inspired by the effectiveness of multi-modality and external knowledge, we explore fundamental research questions related to multimodal dialogue summarization and propose a novel transformer-based adapter-driven clinical conversation summarization framework.

Research Questions We aim to investigate the following three research questions related to multimodal clinical conversation summary generation in the paper: **(i)** How does the inclusion of visual cues, such as visual signs and patients' expressions, impact the process of clinical patient-doctor interaction summarization? **(ii)** Can the inclusion of external knowledge offer more relevant context, thereby enhancing the quality of generated medical dialogue summaries? Does the fusion mechanism of visual/knowledge information with text have any influence on the overall quality of the summary? **(iii)** Is there a correlation between the identification of medical departments and the summarization of medical dialogues?

In order to build a multi-modal clinical conversation summary generation model and validate the research questions, we take the first attempt to build a multi-modal clinical conversation summary (*MM-CliConSumm*) dataset. The dataset bridges the following gaps: (i) Textual-Visual aided clinical interactions annotated summary. (ii) Each patient utterance is annotated with the medical entities contained in it and each dialogue with the concerned medical department and disease. (iii) We have also provided two additional executive summaries for interactions: (a) Medical Concern Summary (MCS), which is a concise one-line summary that captures the primary concern expressed by the patient during the discussion, (b) Doctor Impression (DI) encapsulates the final reaction and impression of the doctor following the conversation with the patient.

Key Contributions The key contributions of the work are fourfold, which are enumerated below.

- Motivated by the tremendous efficacy of visuals in clinical conversation settings, we first propose an autonomous task of multi-modal clinical conversation summarization (MM-CCS) and medical concern summary (MCS) generation.
- We first created a multimodal medical conversation summarization dataset, named *MM-CliConSumm*, which contains clinical conversation annotated with medical vitals such as medical department, patient side summary, one-line summary, and overall summary.
- We propose a multitasking knowledge-infused medical department identification and multi-modal clinical conversation summary generation (*MM CliConSummation*) model incorporated with an adapter-based contextualized M-modality

fusion mechanism that evaluates visual abnormalities and infuses additional knowledge in conjunction with patient-doctor interaction.
- The proposed *MM CliConSummation* model outperforms existing state-of-the-art uni-modal medical summarization models and baselines across all evaluation metrics, including human evaluation.

2 RELATED WORK

The proposed work is relevant to the following three research areas: dialogue summarization, multi-modal summarization, and knowledge-infused text generation. In the following paragraphs, we have summarized the relevant works.

Dialogue Summarization Dialogue summarization has been a longstanding and fundamental problem in Natural Language Processing (NLP). Over the past two decades, the field of dialogue summarization has progressed in the following directions [7]: (i) Feature guided Extractive Summarization [3], (ii) RNN-based summary generation [16], (iii) Pre-trained large language model (PLM) based summarization [35]. In the last few years, the focus has been on the aspect (domain/intent/keyword) guided dialogue summarization and, synthetic data creation with few shot settings. In [11], the authors have proposed a summarization model based on a pointer network generator. The model takes dialogues as input and generates a summary for each turn (doctor-patient) of the interaction. The work [24] proposed a hierarchical encoder-tagger for summarizing medical patient-doctor conversations by identifying important utterances.

Multi-modal Summarization Multi-modal summarization aims to generate coherent and important information from data having multiple modalities [36]. In the last few years, the main focus of multi-modal summarization has been to find co-relation among different modalities: text, audio, and image for video data [1]. An important segment of a video is a subjective concern and may also vary among consumers. In [10], the authors have proposed a new task of user constraint-based summarization and proposed an attention mechanism to summarize the query-relevant content. To generate a coherent summary, synchronizing different modalities is crucial. Shang et al., [23] proposed a time-aware multi-modal transformer (TAMT) that leverages time stamps across image, text, and audio to generate an adequate and coherent video summary.

Knowledge-Infused Text Generation and Summarization Knowledge-infused text generation [17] incorporates external knowledge or information into the process of generating text, allowing the model to generate more accurate and relevant content [8]. In [29], the authors have proposed a novel knowledge-infused dialogue generation model that infuses additional knowledge provided by ConceptNet [25] for query-type utterances, with dialogue context, demonstrating improved generation quality over traditional models. In dialogue, all utterances are not equally important. Motivated by the observation, Manas et al., (2021) [18] proposed PHQ-9 lexicon-guided clinical text-based conversation. They showed their proposed unsupervised model that infuses the knowledge and performs superior in terms of informativeness and underlying interview theme. However, the summaries produced are primarily template-driven and consist of a compilation of turn-specific synopses, resulting in a tedious and lengthy summary.

3 DATASET

We first extensively scrutinized the existing benchmark clinical conversational datasets, and the summary is presented in Table 1. We found Vis-MDD [26] as the most relevant dataset for the proposed multi-modal clinical conversation summarization task. Motivated by the unavailability and the efficacy of clinical conversation summary, we first take the move to develop a multimodal clinical conversation summary generation (*MM-CliConSumm*) dataset. We curated the dataset based on the Vis-MDD corpus under the guidance of two medical professionals.

3.1 *MM-CliConSumm*

We, along with the two medical experts, first analyzed a few medical dialogues of Vis-MDD dataset. We have provided a subset of 100 dialogue samples across ten medical departments, each having 10 samples to the clinicians for summary writing. It contained both text and image-based utterances in each conversation. They wrote three different kinds of summaries for each interaction: overall summary, medical concern summary/MCS (patient side short summary), and doctor impression (DI). The objective of annotating two new kinds of summaries was inspired by telemedicine. MCS helps online healthcare users to locate relevant information effectively, whereas doctor impression aims to help doctors and healthcare systems for effective reference and action points. We further asked them for annotation guidelines for summary writing and provided the sample dataset to three annotators (biology graduate students) for scaling up the corpus. In order to ensure annotation agreement among the annotators, we calculated the kappa coefficient (k). It was found to be 0.73, indicating a significant uniform annotation. The *MM-CliConSumm* dataset statistics are provided in Table 2.

3.2 Qualitative Aspects

The objective of summarization is to represent the essence of a large document in a precise yet concise manner without losing any critical characteristics. To generate an adequate summary of a clinical conversation, we analyze different qualitative characteristics of clinical interactions and incorporate them accordingly.

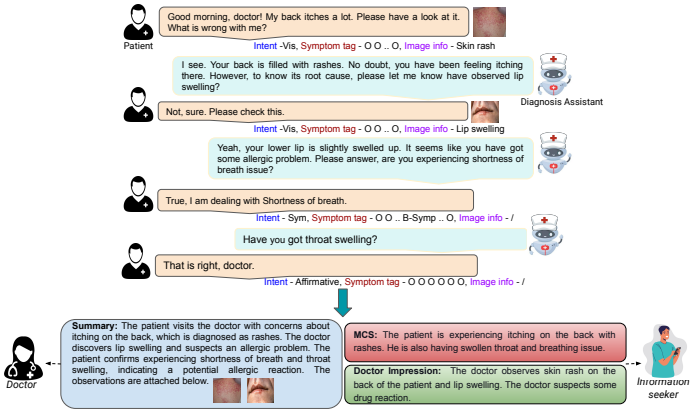


Figure 2: A data sample from the *MM-CliConSumm* corpus

Importance of Visual Descriptions The medical domain is highly specialized and sensitive, and many individuals are unfamiliar with

Dataset	Language	Conversation	Image	Sign's Severity	Intent & Symptom	Medical Department	Summary	Patient Concern Summary	Doctor Impression
RD [30]	Chinese	×	×	×	×	✓	×	×	×
DX [32]	Chinese	×	×	×	×	✓	×	×	×
M^2 [33]	Chinese	✓	×	×	✓	✓	×	×	×
MedDialog-EN [34]	English	✓	×	×	×	×	×	×	×
SD [14]	English	×	×	×	×	✓	×	×	×
Dr. Summarize [11]	English	✓	×	×	×	×	×	×	×
GPT3-ENS SS [6]	English	✓	×	×	×	×	✓	×	×
Vis-MDD [26]	English	✓	✓	×	✓	✓	×	×	×
MM-CliConSumm	English	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Statistics of the existing publicly available medical datasets for disease diagnosis task

Entries	Value
# number of conversations	1668
# of utterances	5483
# of unique words	3512
# of unique images	1668
# number of symptoms	266
# number of diseases	90
# number of medical departments	10
# of diseases in each department	9
avg. length of overall summary (# of words)	48
avg. length of MCS (# of words)	16.64
avg. length of Doctor impression (# of words)	16.86
tags	intent, symptom, visual information, overall summary, MCS, and doctor impression

Table 2: *MM-CliConSumm* dataset statistics

various medical terms, such as "mouth ulcer" and "skin growth". Furthermore, we make an effort to communicate our medical conditions as accurately as possible. Describing something like a skin rash and its intensity through text can be challenging (as shown in Figure 2), so presenting the actual medical condition visually offers an easier and more precise means of communication.

Importance of Medical Department Labeling clinical conversations with medical departments can be beneficial for both clinicians and online healthcare users, as it allows for easy referencing and reusability. Moreover, the medical department label can assist in generating domain-guided responses and summaries of the conversation, ensuring that the information provided is more relevant and tailored to the specific medical field. The distribution of different medical departments in the curated dataset is provided in Figure 3. There are 9 medical departments, and each group contains 10 different diseases. The division is determined as per International Classification of Diseases (ICD-10-CM).

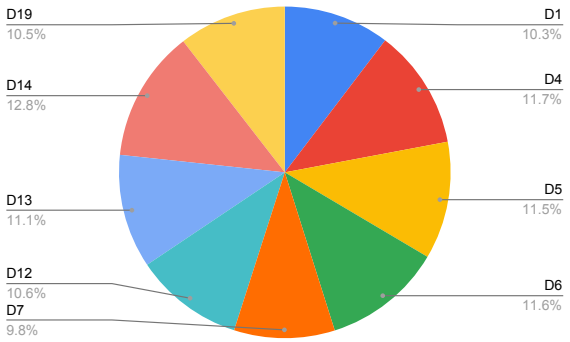


Figure 3: Distribution of conversations across different medical departments in the corpus

Role of Medical Concern Summary (MCS) The Medical Concern Summary (MCS) is a concise and focused summary of a patient's main concern, which is discussed during the interaction with a clinician. Its purpose is to assist online healthcare users in quickly identifying whether a clinical conversation contains the information they are seeking. As an example, the MCS presented in Figure 2 encapsulates the essence of the entire conversation, allowing users to easily determine whether they should refer to the content or not.

Importance of Doctor Impression (DI) In the process of clinical diagnosis and treatment, a patient's journey typically involves multiple interactions rather than a single visit. Consequently, reviewing the entire transcript of a previous lengthy conversation can be time-consuming. Therefore, having access to the patient's Medical Concern Summary (MCS) along with the doctor's impression (as shown in Figure 2) serves as a helpful synopsis/action points of the case for different healthcare stakeholders, reducing the need to refer to the lengthy transcript.

Ethical Consideration We strictly followed the medical research's legal, ethical, and regulatory guidelines during the dataset curation process. With this in mind, we have not added or removed any utterances from the conversation. The curated dataset does not reveal users' identities, such as their names and demographic information. The annotation guidelines are provided by the clinicians, and the dataset is thoroughly checked and corrected by them. Furthermore, we have also obtained approval from our institute's healthcare committee and institutional ethical review board (ERB) to employ the dataset and carry out the research.

4 PROPOSED METHODOLOGY

We anticipate that multi-modal clinical interaction summarization has crucial importance of the following in addition to text/speech of clinicians and patients: (a) patient's visual reporting during an interaction, (b) additional relevant knowledge, and (c) concerned medical department. Thus, we propose a multi-tasking, multi-modal, knowledge-infused medical department identification and dialogue summary generation framework. The proposed architecture is illustrated in Figure 4. We introduce the novel concept of Contextualized M-modality fusion, which utilizes an adapter-based module in a transformer to effectively integrate order-driven visuals and external relevant knowledge for dialogue summarization. There are three key stages: (i) Discourse, Visual and Knowledge representation, (ii) Contextualized M-modality fusion, and (iii) Clinical department identification and Summary generation. The working of each stage and the involved module is explained and illustrated in the subsequent sections.

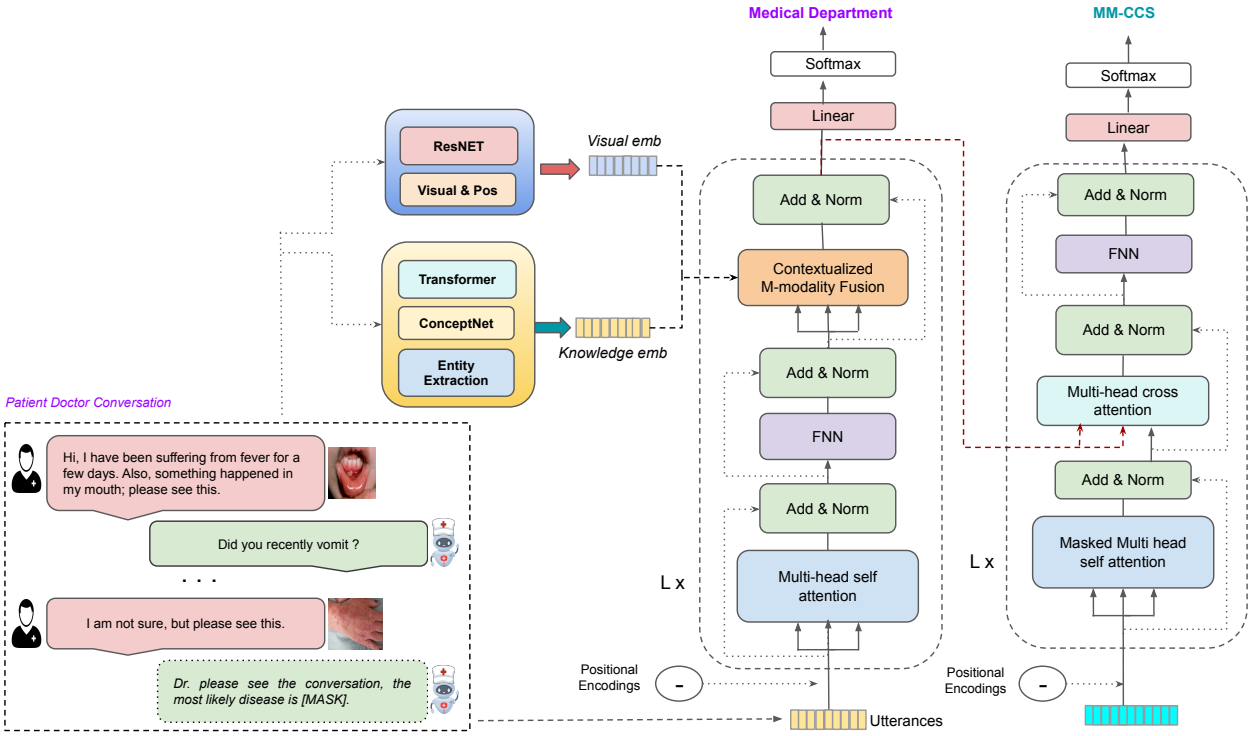


Figure 4: Architecture of the proposed multi-tasking, multimodal medical department identification, and summary generation (*MM-CliConSummation*) model

4.1 Discourse, Visual and Knowledge Representation

We have employed three types of information to encode a patient-clinician interaction: discourse text (utterances from both the clinician and the patient), visuals that were discussed during the interaction, and supplementary discourse-related information obtained from general knowledge. The process of encoding the entities is described below.

Discourse Representation A discourse consists of a sequence of utterances from both the patient and the doctor, where the patient explains his/her main concerns and the doctor conducts further inquiries to aid in diagnosis. We tokenize the combined patient-doctor texts, segmented by turns, using BART tokenizers [13], and extract the transcript embedding.

Visual Features To leverage the pre-training of a state-of-the-art model, we opted for ResNet 152 [9], a widely used visual model, to represent the images. We began by fine-tuning the ResNet model using our labeled dataset of 1700 images, each with corresponding symptom or sign labels. For the symptom identification task, we added a neural network on top of the ResNet and froze the weights of all layers except the last three. Eventually, we extracted the vector representation of the image by pooling the output of the last layer of the ResNet. In cases where multiple images were present within a single interaction, we computed the mean of the image embeddings for visual representation.

Knowledge Infusion The selection of content to include in a summary is a crucial aspect of summarization. This decision-making process is influenced by factors such as the number of samples in the dataset and their diversity. However, in the medical domain, the dataset size is not extremely large due to its sensitivity. In such cases, prior experience and relevant additional knowledge can be particularly valuable. Hence, we utilize additional knowledge to assist the generation model in emphasizing pertinent content.

Algorithm 1 Discourse aware Knowledge Distillation (DKD)

Input Context ($C : p_1, d_1, p_2, d_2, ...d_n$) where p_i and d_i represents i^{th} utterances of patient and doctor, respectively
Output Context relevant Knowledge Graph (KG_C)
Initialization n_k (7): threshold for the number of keywords from a conversation, n_r (5): threshold for the number of concepts for an entity

KG_C = []
K[1, 2, ...n_k] = YAKE(C, n_k) ⇒ K: list of entities
for entity in K **do**
  KGT_{entity} =[] ⇒ KGT_{entity}: KG triplet for "entity"
  for j in range(0, n_r) **do**
    <r_j, h_j, t_j> = ConceptNet(entity, KGT_{entity}) ⇒ r: relation, h: head and t: tail
    KGT_{entity} = KGT_{entity} + [r_j, h_j, t_j]
  end for
  KG_C = KG_C + KGT_{entity}
end for
return KG_C

We choose ConceptNet [25] for knowledge infusion, which is one of the largest knowledge graphs (8 million nodes and 21 million edges) that contains concepts of various domains, such as healthcare. While knowledge is crucial, focusing on relevant knowledge is more significant while solving a task. Thus, infusing the entire ConceptNet knowledge with the proposed summarization setup would be ineffective and may even deteriorate the performance because a

large chunk of it would be irrelevant in a very large number of cases. We propose to distill the external knowledge based on discourse and inject a subset of the knowledge graph dynamically depending on the context. It first extracts essential words (keywords) from the dialogue using an unsupervised statistical-based keyword extractor called YAKE [5]. The extracted entities are passed to the ConceptNet, which identifies relevant concepts associated with them as described in the Algorithm 1.

4.2 Contextualized M-modality Fusion

The manner in which multiple pieces of information are integrated together holds substantial importance for the effectiveness of the combined representation. Therefore, it is vital to merge them in a manner that transforms them into a unified embedding space, ensuring the coherence of the combined representation. Motivated by this, we propose an adapter-based infusion mechanism called contextualized M-modality fusion for combining text, image, and knowledge, which is effective to incorporate with transformer models. The contextualized M-modality generates contextualized modality-conditioned key and value vectors and produces a scaled dot product attention vector. The contextualized modality attention vector is being utilized for calculating the global information attended over visual and knowledge information, which is being utilized for medical domain identification and clinical summary generation. The infusion mechanism is illustrated in Figure 5. It takes the hidden state (H) and calculates the contextualized modality attention as follows:

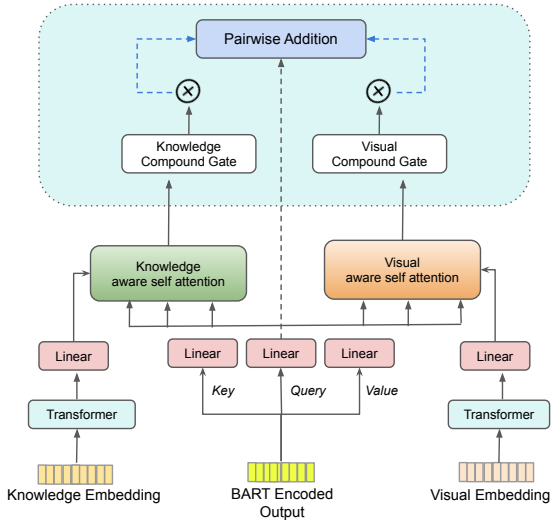


Figure 5: Proposed modality driven knowledge infused modality fusion technique

$$[QKV] = H[W_Q, W_K, W_V] \tag{1}$$

where $Q, K, V \in \mathbb{R}^{lxd}$ are query, key, and value, respectively. Here, l and d denote the sequence length and the dimension of the hidden state (H). The term W_Q, W_K, and W_v are the learnable parameters corresponding to the key vector, having the dimension of $\mathbb{R}^{dxd}$. To determine the co-relation of visuals and additional knowledge with patient-doctor interaction discourse, we generate visual and relevant knowledge conditioned key ($\hat{K}$) and value ($\hat{V}$) vectors. The attention vectors transpose the query vector (dialogue transcript) to generate a multi-modal, knowledge-aware information vector. The key and value pairs are calculated as follows:

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

where $\lambda \in \mathbb{R}^{lx1}$ is the learnable parameter that determines how much information from the textual modality should be retained and how much other modality information should be integrated. Here, E could be evidence (visual feature) or experience (additional relevant knowledge). U_k and U_v are the learnable parameters. The modality controlling parameters (λ) are calculated using the gating mechanism as follows:

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

where $W_{k_1}, W_{k_2}, W_{v_1}$ and W_{v_2} ($\in \mathbb{R}^{dx1}$) are trainable weight matrices. Finally, the visual and knowledge aware attentions (H_v, and H_{kn}) and the final attended vector ($\hat{H}$) are calculated as follows:

$$\begin{aligned} H_v &= Softmax(\frac{Q\hat{K}_v^T}{\sqrt{d_k}})\hat{V}_v \\ H_{kn} &= Softmax(\frac{Q\hat{K}_{kn}^T}{\sqrt{d_{kn}}})\hat{V}_{kn} \end{aligned} \tag{4}$$

Fusion In order to infuse and control the amount of information transmitted from the different modalities (visual and external knowledge), we build two compound gates: visual (g_v) and world-knowledge (g_{kn}). The context information is transmitted via the gates as follows:

$$\begin{aligned} g_v &= [H \oplus H_v]W_v + b_v \\ g_{kn} &= [H \oplus H_{kn}]W_{kn} + b_{kn} \end{aligned} \tag{5}$$

where $\oplus$ denotes a concatenation operation. W_v & W_{kn} ($\in \mathbb{R}^{2dXd}$) and b_v, & b_{kn} ($\in R^{dX1}$) are parameters. The final contextualized attended vector ($\hat{H}$) is computed as Equation 6, which is being utilized for intent identification and encoder representation.

$$\hat{H} = H + g_v \odot H_v + g_{kn} \odot H_{kn} \tag{6}$$

4.3 Medical Department Identification and MM-CCS Generation

In the healthcare system, various medical departments exist, and specialists within each department are generally more adept at comprehending relevant cases. Motivated by this understanding, we aim to leverage the knowledge of the specific medical department to enhance the generation of precise summaries for clinical conversations. Thus, we build a multi-task department identification and summary generation framework (Figure 5) that utilizes the encoder representation to identify the medical department. The same encoded representation is then fed into the decoder to generate the summary. Note, clinical conversation summary generation is our primary task which is being comprehended with the other task of medical department identification.

Clinical Department The encoder takes clinical transcript and determines attention over visual and additional relevant knowledge using the proposed contextualized M-modality fusion. The attended multi-modal encoder representation vector, $\hat{H}$ (Equation

6), is passed to a fully connected neural network having a linear layer and nine nodes in the final layer representing different medical departments.

Multi-modal Clinical Conversation Summary (MM-CCS) The decoder block takes the attended multi-modal encoder representation vector ($\hat{H}$) and feeds it into the multi-head attention layer as key and value, with the query as the hidden representation of the clinical conversation text. The infused information is processed with the traditional transformer's layers of GPT-2 and computes the vocabulary's probability distribution.

Outcome Space and Loss Function The sizes of outcome space are 9 department classes and 5845 vocabulary tokens for classification and generation tasks, respectively. We have utilized a joint categorical cross-entropy loss function, which is the sum of classification (CL) and generation (GL) tasks, i.e., $L = \alpha_1 * CL + \alpha_2 * GL$ and $\alpha_1(= 0.2) + \alpha_2(= 0.8) = 1$.

5 EXPERIMENTAL SETUP

We have utilized the PyTorch framework for implementing the proposed model. The proposed *MM-CliConSummation* generation model was trained for 30 epochs on an RTX 2080 Ti GPU, which took around 30 minutes. The proposed model has been trained, validated, and evaluated with 80%, 5%, and 15% samples of the *ConSummation* dataset, respectively. The hyperparameter values for the model are as follows: sequence length/text (360), sequence length/visual (786), sequence length/knowledge graph (786), batch size (32), optimizer (Adam), activation function (ReLU), and learning rate (5e-05). Furthermore, the dataset and source code are available at https://github.com/NLP-RL/MM-CliConSummation.

Baselines We have utilized the following baselines to comprehend the efficacy and limitations of the proposed model:

- **GPT-2** Generative Pre-trained Transformer-2 (GPT-2) [4] is the state-of-the-art transformer-based language model trained on a humongous amount of English corpora in the self-supervised setting.
- **BART** BART [13] is a denoising autoencoder model that is trained to reconstruct corrupted sentences.
- **T5** T5 [22] is a versatile text-to-text model that combines encoder-decoder architecture with pre-training on a mixture of unsupervised and supervised tasks.
- **MAF** MAF [12] is a fusion model that incorporates an additional adapter-based layer in the encoder of BART to infuse information from different modalities.
- **K-CliConSummation** It is the proposed model with only knowledge infusion (w/o multimodal visual infusion).
- **M-CliConSummation** M-CliConSummation is the proposed model with multimodal visual (w/o knowledge infusion).
- **M-CliConSummation w/o fusion** It is the proposed model where different modalities, text, knowledge, and visuals are combined simply by concatenation.
- **KM-CliConSummation** KM-CliConSummation is the proposed framework with both multimodal visual and external knowledge infusion but without multi-tasking of department identification and summary generation.

6 RESULTS AND DISCUSSION

We employed the most popular automatic evaluation metrics for summarization/text generation, namely BLEU, Rouge, and METEOR [2, 15, 20], to evaluate the adequacy of summarization quality of the proposed model. The purpose of the proposed multi-task framework is to enhance the performance of the clinical dialogue summarization task by utilizing the additional task of medical department identification. Thus, the results and analysis mainly emphasized on summarization task. Based on the experiments, we report the following answers (with evidence) to our investigated research questions (RQs).

RQ1: How does the inclusion of visual cues, such as visual signs and patients' expressions, impact the clinical patient-doctor interaction summarization task? We experimented with different models with and without visual information for both overall summary generation and medical concern summary (MCS) generation. The obtained results are reported in Table 3 (overall summary) and Table 4 (MCS). The inclusion of visual description led to the following improvement in generation quality– *Overall summary*: BLEU (1.28 ↑), ROUGE-L (1.45 ↑), and METEOR (0.52 ↑), *MCS*: BLEU (1.35 ↑), ROUGE-L (1.06 ↑), and METEOR (1.75 ↑). It also improved other evaluation metrics. The improvements by M-ConSummation for both tasks across the evaluation metrics firmly demonstrate the effectiveness of visuals in clinical conversation summary generation.

RQ2 (a): Can the inclusion of external knowledge offer more relevant context, thereby enhancing the quality of generated medical dialogue summaries? Through the experiments, it became apparent that the infusion of knowledge played a pivotal role in enhancing the quality of generation, benefiting both the overall summary and medical concern summary (MCS). The knowledge infusion led to the following improvements– *Overall summary*: BLEU (0.68 ↑), ROUGE-L (0.40 ↑), and METEOR (1.14 ↑), *MCS*: BLEU (0.53 ↑), ROUGE-L (0.66 ↑), and METEOR (1.05 ↑). Furthermore, we also observed that knowledge infused with simple concatenation with text/visual performs very poorly compared to one with proposed contextualized M-modality fusion.

RQ2 (b): Does the amalgamation technique of various modalities, namely text, visuals, and knowledge, have any influence on the quality of generated summaries? To investigate the research question, we conducted experiments involving various techniques for integrating modalities at different model layers. The obtained results are reported in Table 3 and Table 5 (KM-CliConSummation w/o fusion and KM-CliConSummation w/ fusion). It shows that the model that incorporates modality order-driven multimodal infusion performs significantly superior to the model which simply concatenates different modalities. We anticipate that the order of modalities infusion and the distance between them is crucial to the effectiveness of combined information. Thus, we also experimented with different models having modalities infusion at different layers of the transformer (Table 4). The findings show that the most preferred position for modality infusion in the transformer is towards the last layers (we have a total of six layers in the encoder). Moreover, knowledge should be infused before visual as knowledge is also a kind of text and thus can be merged uniformly with text.

Model	B-1	B-2	B-3	B-4	BLEU	R- 1	R- 2	ROUGE- L	METEOR	Jaccard Sim	BERT Score	Accuracy	F1-Score
GPT-2 [4]	11.65	5.34	2.22	0.80	5.00	21.23	4.64	20.37	23.41	0.0717	0.6660	/	/
BART [13]	9.94	7.18	5.16	3.72	6.50	38.37	18.04	35.50	18.68	0.1833	0.8378	/	/
T5 [22]	42.27	32.00	24.60	18.58	29.36	54.16	31.74	51.24	43.22	0.2582	0.8841	/	/
MAF [12]	47.18	36.47	27.62	20.02	32.82	59.31	36.93	49.71	55.10	0.2699	0.9131	/	/
K-CliConSummation	47.87	36.57	27.89	21.67	33.50	59.45	37.21	50.11	56.24	0.2724	0.9096	/	/
M-CliConSummation	48.16	37.74	28.06	22.46	34.10	60.10	37.21	51.16	55.62	0.2766	0.9148	/	/
KM-CliConSummation (w/o fusion)	35.38	21.40	12.01	7.22	19.00	42.29	18.88	33.09	34.59	0.1536	0.7787	31.20	0.2201
KM-CliConSummation (w/ fusion)	48.77	37.37	28.44	22.16	34.18	60.27	37.90	50.87	56.70	0.2753	0.9127	/	/
MM-CliConSummation$	**49.26**	**37.88**	**29.03**	**22.78**	**34.74**	**60.47**	**38.13**	**51.77**	**57.15**	**0.2778**	**0.9184**	**60.68**	**0.5631**

Table 3: Performances of different models for multi-modal clinical conversation summary generation. Here, $ indicates statistical significant findings ($p < 0.05$ at 5% significance level)

Visual layer	Knowledge layer	B-1	B-2	B-3	B-4	BLEU	R-1	R-2	ROUGE- L	METEOR	Jaccard Sim	BERT Score	Accuracy	F1-Score
2	3	47.21	35.54	26.86	20.39	32.50	59.05	36.19	49.20	54.65	0.2701	0.9118	49.57	0.4578
2	4	47.26	35.23	26.17	19.69	32.09	58.96	35.67	48.60	53.32	0.2636	0.9104	48.71	0.4469
3	4	48.49	36.91	28.06	21.53	33.75	60.07	37.09	50.48	56.02	0.2724	0.9150	58.97	0.5448
3	2	48.68	37.06	28.18	21.62	33.88	60.01	37.47	50.06	56.23	0.2735	0.9146	52.13	0.5025
4	2	47.48	36.10	27.43	20.87	32.97	58.98	36.62	49.54	56.01	0.2683	0.9109	55.12	0.5325
4	3	**49.26**	**37.88**	**29.03**	**22.78**	**34.74**	**60.47**	**38.13**	**51.77**	**57.15**	**0.2778**	**0.9184**	**60.68**	**0.5631**

Table 4: Performance of the proposed multi-modal clinical summary generation model with different modality infusion orders. There are 6 layers in the encoder (Figure 4), and the higher layer number indicate a layer towards the end of the encoder

Model	BLEU	R-1	R-2	ROUGE- L	METEOR	Jaccard Sim	BERT Score
GPT 2 [4]	2.45	14.1	3.03	13.57	19.45	0.0512	0.6558
BART [13]	23.91	46.92	27.11	44.16	43.59	0.2964	0.8784
T5 [22]	26.37	48.02	27.69	44.10	49.28	0.2994	0.8770
MAF [12]	33.83	61.12	42.23	58.95	57.79	0.3584	0.8778
K-CliConSummation	34.36	61.59	42.70	59.61	58.84	0.3630	0.8810
M-CliConSummation	35.18	62.42	43.18	60.01	59.54	0.3737	0.8859
KM-CliConSummation (w/o fusion)	25.01	48.64	29.54	46.01	43.36	0.2746	0.8153
KM-CliConSummation (w/ fusion)	35.04	62.40	43.03	60.22	59.32	0.3720	0.8842
MM-CliConSummation$	**35.78**	**62.83**	**43.96**	**60.90**	**59.51**	**0.3772**	**0.8862**

Table 5: Performance of the different models for medical concern summary generation. Here, $ indicates statistical significant findings ($p < 0.05$ at 5% significance level)

Model	Adequacy	Fluency	DR	Consistency	Info	Avg
T5 [22]	2.80	4.26	3.96	3.34	3.98	3.67
MAF [12]	3.24	4.28	4.01	3.64	4.15	3.86
K-ConSummation	3.46	4.32	4.38	3.86	4.22	4.05
M-ConSummation	3.56	4.31	4.40	3.82	4.28	4.07
KM-ConSummation	3.65	4.36	4.44	3.94	4.34	4.15
MM-ConSummation	**3.88**	**4.40**	**4.56**	**4.02**	**4.42**	**4.26**

Table 6: Human evaluation of different summary generation models

RQ3: Is there a correlation between the identification of medical departments and the summarization of medical dialogues? To investigate the research question, we experimented with a multi-task framework that identifies the medical department as well as generates a summary of a clinical conversation. The results are reported in Table 3 and Table 4. The proposed multi-tasking framework performs superior to all baselines across various evaluation metrics for both overall summary and medical summary generation. Similar behavior is also obtained in human evaluation; the model outperformed all others, even with human perception. Note that the human evaluation conducted was performed in a blind review manner, ensuring that no information regarding the model names was provided alongside the summaries.

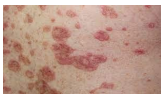
Human Evaluation We also conducted human evaluation of 100 test samples. In this assessment, two medical domain experts and one researcher (other than the authors) were employed to evaluate the generated summaries by different models (without revealing the models' names). The samples are assessed based on the following five metrics: *adequacy, fluency domain relevance (DR), consistency, and informativeness (info)* on a scale of 1 (extremely poor) to 5 (idle). The obtained scores are presented in Table 6.

Key Observations The key observations and insights are as follows: **(i)** Gaining an understanding of one task by leveraging knowledge from a related task is consistently advantageous. The proposed model exhibits similar behavior, both in terms of modality infusion (where knowledge is infused with text first, followed by visual infusion with text, and finally combining text, attended knowledge, and visual vectors) and multi-tasking (medical department identification and summary generation). **(ii)** The model that incorporates knowledge/visual features at the initial layers of the encoder (as shown in Table 4) exhibits subpar performance. This can be attributed to the prominent importance of the text modality in summarization, requiring some processing before merging with supplementary information from heterogeneous sources. **(iii)** We observed that the models infused with knowledge perform significantly superior for medical department and disease identification (department/disease is also included in some summaries).

7 ANALYSIS

We conducted a thorough qualitative analysis of the summaries generated by different baselines and our proposed models. We also performed some case studies; one such instance is illustrated in Figure 6. The analysis leads to the following: (a) The state-of-art models (T5 and MAF–unimodal and w/o external knowledge) and baselines quite often either do not include disease or infer an incorrect disease. In the majority of cases, the predicted diseases tend to belong to the same group that encompasses the disease afflicting the patient. The behavior can be attributed to common symptoms across diseases of the same medical department. (b) In some cases, the baseline models include some most frequently occurring symptoms in the dataset despite having different contexts, such as fever and pain. We observed that the inclusion of knowledge in the proposed model has led to more factual consistency. (c) There are numerous ways to craft a summary that effectively captures the essence of a conversation. During our analysis, we encountered several instances where the BLEU score, a word-based

Transcript **-** <Patient> Doctor, I have been feeling intense itching of skin. Please help. <Doctor> Don't panic, Please describe if you have any skin lesion or any other problem that makes me to understand better. <Patient> Right, ... <Doctor> Do you have any skin rash. <Patient> Not sure but please see this $S31_1_38.jpg$ <Doctor >:Ok, it looks like you have <MASK> disease.
Overall Summary: The patient describes intense skin itching with skin lesions and moles, and slight pain in the moles. The doctor asks for further details and determines the patient also has a skin rash. The diagnosis is contact dermatitis.
Medical Concern Summary : Patient is experiencing intense itching, has skin lesions and moles, and slight pain in the moles.
Concerned Department: Dermatology


S31_1_38.jpg

Model	Summary	Medical Concern Summary
T5	The patient is experiencing intense itching of the skin and seeks the doctor's help. The doctor asks about skin moles. <no disease>	The patient is experiencing intense itching of the skin and has a skin lesion, skin moles, and pain.
MAF	The conversation is between a patient and a doctor. The patient complains of intense itching of the skin and seeks the doctor's help. .... the doctor diagnoses the patient with dermatitis due to sun exposure.	Patient is experiencing intense itching of the skin with lesions and pain.
K-CliConSummation	The patient is experiencing intense itching of the skin and seeks medical help from the doctor. The doctor asks for more details about the patient's symptoms and the presence of skin lesions and moles. Based on the symptoms, the doctor diagnoses the patient with Contact dermatitis.	Patient is experiencing intense itching of skin and also has skin lesion issue.
M-CliConsummation	The patient complains of intense itching of skin and seeks the doctor's help. The doctor asks about any skin lesions or other problems that makes the patient understand better, and the patient confirms having skin moles, slight pain, and **skin rash**. Based on the patient's symptoms, the doctor diagnoses them with Eczema.	Patient is experiencing intense itching of skin with lesions, moles and skin rash.
MM-CliConSummation	The conversation is between a patient and a doctor regarding the patient's intense itching of the skin. The doctor asks if the patient has any skin lesions or any other skin problems. The doctor observes skin lesion, skin moles, and **skin rash**. Based on the symptoms, the doctor diagnoses the patient with **Contact dermatitis**.. **Medical Department:** Dermatology	Patient is experiencing intense itching of skin with skin lesions, moles and skin rash.

Figure 6: Performance of the different baselines and proposed models for a common test case

matching metric, was relatively low. However, the generated summaries were highly relevant and had higher BERT scores, which evaluate similarity based on contextual embeddings. It is important to note that while the improvement in the BERT score margin may be less, it holds significant value since the BERT score scale differs from other metrics.

Limitations Despite the significant improvement demonstrated by the proposed knowledge-infused multi-modal dialogue summary model, we observed some weaknesses and limitations. (i) In a few cases, the model generates incomplete names of some long symptoms (spots for spots in vision). Also, it mises medical condition in summary in a few cases. Nevertheless the number was quite less (16/100 in human evaluation). (ii) The avg. length (in words) of gold summaries and generated summaries for test samples were 46 and 42, respectively. The proposed model tends to generate relatively small summaries, particularly for dialogues having a large number of utterances. (iii) During the process of summarizing a case for a senior doctor, junior doctors often provide comprehensive descriptions of visual symptoms, including details such as severity and the affected area. However, the *MM-CliConSummation* model does not delve into these specific details. Instead, it identifies symptoms from images and includes their names in the summaries. This limitation is primarily attributed to the lack of meticulousness in annotating visual symptom images.

8 CONCLUSION

In this work, we proposed the task of multi-modal clinical conversation summarization and medical concern summary generation. We curated a multimodal clinical conversation summary generation dataset, *MM-CliConSumm*, and annotated each conversation with two additional executive summaries, a medical concern summary, and a doctor impression. When we summarize a document, we tend to focus on some crucial evidence and part of the document relevant to the referenced document. Motivated by the observation, we propose a multi-tasking, knowledge-infused, multimodal clinical conversation summary generation, *MM-CliConSummation* framework. It takes clinical conversation (having both text and visual) as input and generates a precise summary, and identifies the concerned clinical department. The proposed *MM-CliConSummation* model extracts relevant knowledge graphs depending on dialogue context, constructs visual representation for visual reporting, and infuses them with the modality-driven contextualized fusion technique. The model identifies the concerned medical department with encoder representation, and the decoder generates a summary. With the extensive set of experiments, including human evaluation, the proposed *MM-CliConSummation* model demonstrated significant improvement over baselines and state-of-the-art models across all evaluation metrics for both summary and MCS generation. Summaries can be crafted using different word sequences or synonymous terms compared to the gold standard summary. Thus, relying solely on word-based matching is inadequate; semantic comprehension should also be taken into account. In the future, we aim to develop a novel loss function for summary generation that optimizes both semantic and lexical aspects.

9 ACKNOWLEDGEMENT

Abhisek Tiwari expresses sincere appreciation for receiving the Prime Minister Research Fellowship (PMRF) Award from the Government of India, which provided support for conducting this research. Dr. Sriparna Saha extends heartfelt gratitude for the Young Faculty Research Fellowship (YFRF) Award, supported by the Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, and implemented by Digital India Corporation (formerly Media Lab Asia), which has been indispensable in facilitating this research.

REFERENCES

[1] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *Proc. IEEE* 109, 11 (2021), 1838–1863.

[2] 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*. 65–72.

[3] Mohammed Salem Binwahlan, Naomie Salim, and Ladda Suanmali. 2009. Swarm based text summarization. In *2009 International Association of Computer Science and Information Technology-Spring Conference*. IEEE, 145–150.

[4] Paweł Budzianowski and Ivan Vulic. 2019. Hello, It's GPT-2-How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. *EMNLP-IJCNLP 2019* (2019), 15.

[5] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* 509 (2020), 257–289.

[6] 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*. PMLR, 354–372.

[7] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications* 165 (2021), 113679.

[8] Manas Gaur, Ugur Kursuncu, Amit Sheth, Ruwan Wickramarachchi, and Shweta Yadav. 2020. Knowledge-infused deep learning. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 309–310.

[9] 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*. 770–778.

[10] Jia-Hong Huang, Luka Murn, Marta Mrak, and Marcel Worring. 2021. Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 580–589.

[11] Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures.. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3755–3763.

[12] 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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5956–5968.

[13] 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* (2019).

[14] Kangenbei Liao, Qianlong Liu, Zhongyu Wei, Baolin Peng, Qin Chen, Weijian Sun, and Xuanjing Huang. 2020. Task-oriented Dialogue System for Automatic Disease Diagnosis via Hierarchical Reinforcement Learning. *arXiv preprint arXiv:2004.14254* (2020).

[15] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[16] Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global Encoding for Abstractive Summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 163–169.

[17] Ruibo Liu, Guoqing Zheng, Shashank Gupta, Radhika Gaonkar, Chongyang Gao, Soroush Vosoughi, Milad Shokouhi, and Ahmed Hassan Awadallah. 2022. Knowledge infused decoding. *arXiv preprint arXiv:2204.03084* (2022).

[18] Gaur Manas, Vamsi Aribandi, Ugur Kursuncu, Amanuel Alambo, Valerie L Shalin, Krishnaprasad Thirunarayan, Jonathan Beich, Meera Narasimhan, Amit Sheth, et al. 2021. Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study. *JMIR Mental Health* 8, 5 (2021), e20865.

[19] 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 (2020), 1427–1437.

[20] 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*. 311–318.

[21] Vladimir V Popov, Elena V Kudryavtseva, Nirmal Kumar Katiyar, Andrei Shishkin, Stepan I Stepanov, and Saurav Goel. 2022. Industry 4.0 and digitalisation in healthcare. *Materials* 15, 6 (2022), 2140.

[22] 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 (2020), 5485–5551.

[23] Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal Video Summarization via Time-Aware Transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1756–1765.

[24] 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*. 717–729.

[25] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

[26] Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet 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*. 1935–1944.

[27] Abhisek Tiwari, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A knowledge infused context driven dialogue agent for disease diagnosis using hierarchical reinforcement learning. *Knowledge-Based Systems* 242 (2022), 108292.

[28] Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6638–6660.

[29] Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving knowledge-aware dialogue generation via knowledge base question answering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 9169–9176.

[30] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 201–207.

[31] Jedrek Wosik, Marat Fudim, Blake Cameron, Ziad F Gellad, Alex Cho, Donna Phinney, Simon Curtis, Matthew Roman, Eric G Poon, Jeffrey Ferranti, et al. 2020. Telehealth transformation: COVID-19 and the rise of virtual care. *Journal of the American Medical Informatics Association* 27, 6 (2020), 957–962.

[32] Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7346–7353.

[33] Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhumin Chen, Zhaochun Ren, and Huasheng Liang. 2021. Mˆ 2-MedDialog: A Dataset and Benchmarks for Multi-domain Multi-service Medical Dialogues. *arXiv preprint arXiv:2109.00430* (2021).

[34] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. MedDialog: Large-scale Medical Dialogue Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[35] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11765–11773.

[36] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 4154–4164.