Yes, this is what I was looking for! Towards Multi-modal Medical Consultation Concern Summary Generation

No Author Given

No Institute Given

Abstract. Over the past few years, the use of the Internet for healthcare-related tasks has grown by leaps and bounds, posing a challenge in effectively managing and processing information to ensure its efficient utilization. During moments of emotional turmoil, we frequently turn to the internet as our initial source of support, choosing this over discussing our feelings with others due to the associated social stigma. In this paper, we propose a new task of multi-modal medical concern summary (*MMCS*) generation, which provides a short and precise summary of patients' major concerns brought up during the consultation. Nonverbal cues, such as patients' gestures and facial expressions, aid in accurately identifying patients' concerns. Doctors also consider patients' personal information, such as age and gender, in order to describe the medical condition appropriately. Motivated by the potential efficacy of patients' personal context and visual gestures, we propose a transformer-based multi-task, multi-modal intent-recognition, and medical concern summary generation (*IR-MMCSG*) system. Furthermore, we propose a multitasking framework for intent recognition and medical concern summary generation for doctor-patient consultations. We construct the first multi-modal medical concern summary generation (*MM-MediConSummation*) corpus, which includes patient-doctor consultations annotated with medical concern summaries, intents, patient personal information, doctor's recommendations, and keywords. Our experiments and analysis demonstrate (a) the significant role of patients' expressions/gestures and their personal information in intent identification and medical concern summary generation, and (b) the strong correlation between intent recognition and patients' medical concern summary generation. The dataset and source code are available at `https://github.com/AnonymousRW/MMCSG`.

Keywords: Clinical Conversation · Concern Summary · Multi-modality · Modality Fusion · Multi-tasking · Summary Generation.

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 [18]. With the COVID-19 pandemic, internet utilization for healthcare activities has reached its peak and has become a new normal [29]. The outbreak has caused a striking 25% increase[1] in anxiety and depression, which are

[1] https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

severely straining the mental healthcare systems. Tele-health usage is being actively encouraged by healthcare providers, and patients are adopting it at the same pace. Consequently, a massive amount of medical data became available for the first time over the internet [3]. Thus, arranging this data efficiently is essential for proper referencing and facilitating their potential for reuse.

Fig. 1: Utility of multi-modal medical concern summary generation; generated medical concern summary for a video helps in selecting a proper video relevant to user (right side)

In moments of mental distress, individuals often turn to the internet to seek similar cases and recovery insights. However, they have to go through several lengthy videos before getting a relevant case. While many videos do provide summaries, these summaries combine information from both the patient and the doctor, rendering them ineffective as valuable references for finding pertinent cases that align closely with patients' main concerns (Fig. 1). Motivated by this, we propose a new task of generating Multimodal Medical Concern Summary (*MMCS*). *MMCS* is a synopsis of a patient's key concerns discussed in a patient-doctor interaction. Its potential benefits extend to both patients and clinicians, serving various purposes, including (a) aiding in the organization of consultations and enhancing search ranking for reference by other patients, (b) facilitating follow-up recommendations, and (c) contributing to resource allocation and planning.

When we consult with doctors, they also consider our facial expressions and gestures to analyze our concerns and plan treatment accordingly. Visual expression and audio tone are also affected by user personality, so one behavior may be triggering for one personality while being normal for another. If the patient's key concern is known, it is easy to identify the patient's intent and vice-versa. Thus, we hypothesize there is a significant correlation between intent and medical concern summary. Moreover, we anticipate that visual and audio features are strongly associated with demographic information such as age and gender (context), and they can significantly influence the understanding of users' behavior and concerns with such context-attended features. Hence, we propose contextualized M-modality fusion, a new modality fusion technique that incorporates an adapter-based module into traditional transformer architecture to effectively infuse different modalities and end-user demographic information.

Research Questions The paper aims to investigate the following research questions: (a) Does the visual appearance and expression of a patient aid in determining his/her key

medical concerns? (b) Is there a correlation between medical concern summary generation and user intent identification? (c) Can patients' personal characteristics, such as age and gender information, contribute to adequately understanding their medical issues and providing appropriate medical advice?

Key Contributions The key contributions of the work are four-fold, which are as follows:

- We propose a new task for multi-modal medical concern summary (*MMCS*) generation, which generates a precise summary of key medical concerns discussed during doctor-patient consultations, resulting in better content searchability and organization.
- We first curated a Medical Concern Summary annotated multi-modal medical dataset named (*MM-MediConSummation*), which consists of patient-doctor counseling sessions annotated with a precise medical concern summary (MCS), intent, patient's personal attributes, doctor's key points, and keywords.
- We present a multitask, multi-modal intent recognition, and multi-modal medical concern summary generation (*IR-MMCSG*) model incorporated with an adapter-based contextualized M-modality fusion mechanism that evaluates audio tone and visual expression in conjunction with user demographic context.
- The proposed contextualized M-modality fusion incorporated *IR-MMCSG* outperforms existing state-of-the-art multi-modal text generation model across all evaluation metrics, including human evaluation.

2 Related Works

The work is mainly relevant to the following two research areas: Medical dialogue understanding and Medical dialogue summarization. We have summarized the relevant works in the following paragraphs.

Medical Dialogue Understanding: Diagnosis and treatment of diseases begin with patient-doctor interaction. Therefore, understanding patients' concerns from their utterances is critical to diagnosis and treatment outcomes [7]. The existing works on medical dialogue understanding can broadly be grouped into two categories: (a) pre-trained transformer-based joint intent and entity detection model [28] and (b) multi-label entity classification [22]. The existing works on medical entity extraction [6,4] are limited to medical attributes like symptoms and medicine, which are functions of an utterance rather than a conversation. The work [13] proposed a pipelined machine learning system that identifies intent, symptoms, and disease from a patient-doctor conversation. In [25], the authors demonstrated that considering patients' education level, health literacy, and emotional state greatly improves the likelihood of recommending a relevant medical document.

Medical Dialogue Summarization: Joshi et al., [14] 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 aims to generate coherent

and important information from data having multiple modalities [30]. 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 [12], the authors have proposed a new task of user constraint-based summarization and proposed an attention mechanism to summarize the query-relevant content. Shang et al., [20] 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.

3 Dataset

We first extensively scrutinized the existing benchmark video summarization datasets, and the summary is presented in Table 1. The most relevant dataset for our proposed task is HOPE [17], which contains patient-doctor counseling sessions. The therapy sessions have been collected from open-source platforms such as YouTube, which are credible and authentic, recorded by psychiatrists and clinicians. The consultations cover therapy sessions for various mental distress like anxiety, depression, and post-traumatic stress disorder (PTSD). It contains only doctor-patient conversation transcripts and dialogue acts corresponding to each utterance. With the guidance of two psychiatrists and a doctor, we incorporate the following attributes into the existing HOPE dataset: medical concern summary, primary intent, secondary intent, doctor suggestion, focal point, and patient's personal context (gender and age group).

Table 1: Characteristics of some of the most relevant existing medical dialogue datasets and comparison with the curated dataset

Dataset	Description	Size	Video	Transcript	Intent	MMCS	Focal point	Keywords	Patient Personal Context
DAIC-WOZ [10]	Patient- Psychiatrist conversation	189	✓	✓	✓	×	×	×	×
Dr. Summarize [14]	Patient-Doctor conversations	1690	×	✓	×	×	✓	✓	×
GPT3-ENS SS [5]	Patient-Doctor conversations	210	×	✓	×	✓	×	✓	×
CoDEC [23]	Patient- Psychiatrist conversation	30	✓	✓	×	×	×	×	×
HOPE [17]	Patient- Psychiatrist conversation	212	✓	✓	×	×	×	×	×
MM-MediConSummation	Patient- Psychiatrist conversation	467	✓	✓	✓	✓	✓	✓	✓

3.1 Data Collection and Annotation

We, along with the three medical experts, first analyzed a few therapy sessions of the HOPE dataset. The clinicians viewed a small subset of the dataset, curated a sample dataset of 25 therapy sessions, and annotated it with a multi-modal medical concern summary, patients' intent, and other crucial information, namely the summary of the doctor's suggestion, focal point, and keywords. We employed two biology graduates and one medical student to scale up the dataset and annotation. We provided the sample dataset with a set of guidelines to annotate these tags. They first watched full videos and annotated the details based on a comprehensive understanding of all three modalities involved in a video, viz., text, audio, and visual, for identification. The process of annotating a video involves crafting a medical concern summary (MCS), identifying the intent, and delineating a specific segment within the video that effectively conveys essential details regarding the patient's concerns and the doctor's recommendations. They manually generated transcripts for specific therapy sessions in cases where video

captions were either unavailable or subject to restrictions. In order to ensure annotation agreement among the annotators, we calculated Fleiss' kappa [9]. It was found to be 0.71, indicating a significant uniform annotation. The *MM-MediConSummation* statistics and its word cloud are provided in Table 2 and Fig. 2, respectively.

Table 2: *MM-MediConSummation* dataset statistics

Entries	Value
# of doctor-patient sessions	467
avg duration of video	10 min 58 sec
# of utterances	74473
avg. conversation length (utterance)	159
# of unique words	13639
# of intents	7
avg length of MMCS (# of words)	21
# of age-groups	6
tags	MMCS, keywords, doctor suggestion, focus points, patient gender, patient age, primary intent, secondary intent, transcript

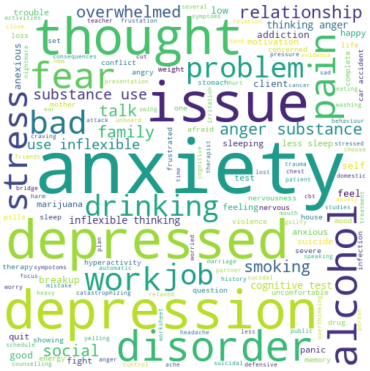


Fig. 2: Word cloud of the *MM-MediConSummation* corpus

3.2 Qualitative Aspects

To generate an adequate summary of patient concerns, we analyze different qualitative characteristics of therapy sessions and incorporate them accordingly. The different characteristics are analyzed and illustrated below.

Role of Medical Concern Summary Medical concern summary aims to aid online healthcare users in recognizing whether the content (particularly therapy session) contains the information they are looking for or not. For example, Fig. 1 illustrates that the patient easily finds out a relevant video when the therapy sessions are tagged with *MMCS*, intent, and keywords.

Role of Intent It is desirable to comprehend users' intent to serve them effectively. For example, a user's goal might be to get a suggestion for a medical condition. We can use it to effectively locate the relevant span in the concerned video. Sometimes, we observed patients having multiple intentions for consultation; thus, we tagged primary and secondary intentions (Fig. 3 and Fig. 4) to each session. We have used only primary intent for multi-tasking intent identification and MMCS generation.

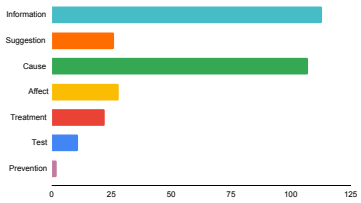


Fig. 3: Primary intent distribution

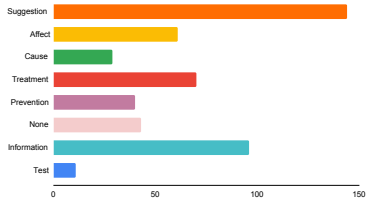


Fig. 4: Secondary intent distribution

Role of Patient's Personal Information In real life, doctors often consider patients' personal context for determining a medical condition and plan treatment accordingly.

Thus, we also annotated patients' personal information for each counseling session. Figures 5 and 6 demonstrate the distribution of gender and age among the patients in the dataset.

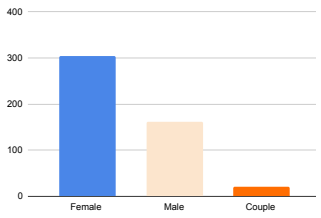


Fig. 5: Gender distribution

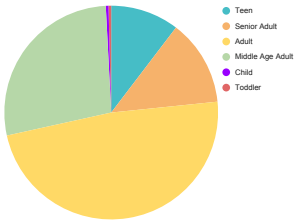


Fig. 6: Age group distribution

4 Methodology

We anticipate that *MMCSG* is affected by (a) patient's visual expression, (b) patient's intent of communication, and (c) patient's personal information. Thus, we propose a multi-tasking, multi-modal intent recognition, and multi-modal medical concern summary generation (*IR-MMCSG*) framework. The proposed architecture is illustrated in Fig. 7. There are three key stages: Multi-modal feature extraction, Contextualized M-modality fusion, and Medical concern summary generation. The explanation and illustration of each stage and the flow are described below.

4.1 Multi-modal Feature Extraction

In order to encode a counseling session, we have considered all three modalities of video, i.e., audio, text (transcript), and image (video frame). We extract the features of different modalities as follows:

Textual Features The textual features of a therapy session represent the text embedding of its transcript. We use existing T5 [19] and BART [16] tokenizers for transcript embedding.

Audio Features We extracted the audio feature from one of the most popular audio processing platforms named openSMILE [8]. The feature representation considers maxima dispersion quotients, glottal source parameters, low-level descriptors (LLD), voice quality, MFCC, pitch, and their statistics.

Video Features We extracted frames from each counseling session at ten frames per second (fps). The frames extracted from the video are analyzed using Katna's approach [26], aiming to identify frames with distinctive features. This process yielded a set of ten highly pertinent frames. Subsequently, these frames were passed to the pre-trained ResNet 152 model [11] to acquire embeddings of these frames. Finally, we computed the average of these embeddings to form a representation vector of the video.

4.2 Contextualized M-modality Fusion

We propose a novel multi-modal adapter-based infusion mechanism called contextualized M-modality fusion. It generates context- and modality-conditioned key and value

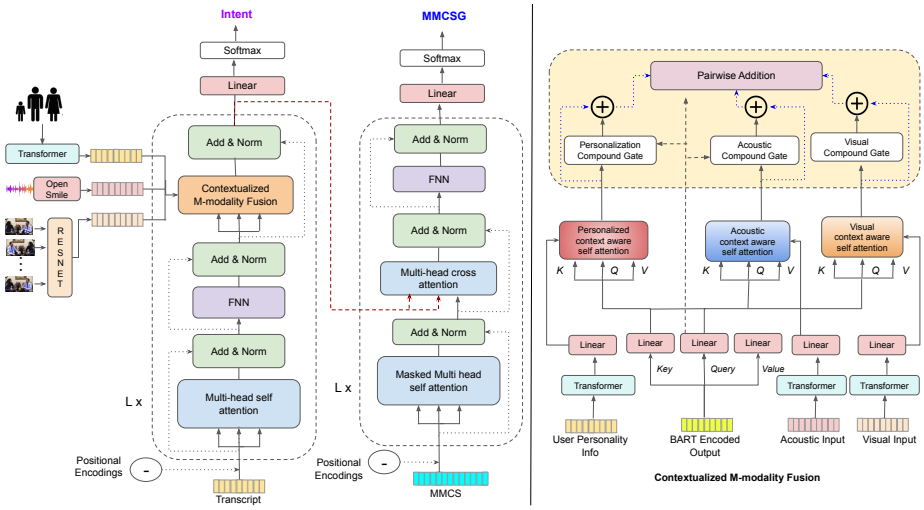


Fig. 7: Architecture of the proposed Multitasking multi-modal intent recognition and medical concern summary generation (*IR-MMCSG*) framework.

vectors and produces a scaled dot product attention vector. The contextualized modality attention vector is utilized to calculate the global information attended over audio, visual, and personal context, which is utilized for intent identification and medical concern summary generation. It takes the hidden state (H) and calculates the contextualized modality attention as follows:

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

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), respectively. The term W_Q, W_K, and W_v are the learnable parameters corresponding to the query, key, and value vectors, with the dimension of $\mathbb{R}^{dxd}$.

To understand the medical consultation effectively, we generate different modalities and patients' personal context-conditioned key ($\hat{K}$) and value ($\hat{V}$) vectors. The attention vectors transpose the query vector (video transcript) to generate a contextualized, multimodal, coherent information vector. The key and value pairs are calculated as follows:

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

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, M denotes modality, which could be audio, video, and personal information. U_k and U_v are the learnable parameters. The modality controlling parameters (λ) are calculated using the gating mechanism as follows:

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

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

$$\begin{aligned} H_a &= Softmax(\frac{Q\hat{K}_a^T}{\sqrt{d_k}})\hat{V}_a \\ H_v &= Softmax(\frac{Q\hat{K}_v^T}{\sqrt{d_k}})\hat{V}_v \\ H_p &= Softmax(\frac{Q\hat{K}_p^T}{\sqrt{d_k}})\hat{V}_p \end{aligned} \tag{4}$$

Fusion In order to infuse and control the amount of information transmitted from the different modalities (user personality information, audio, and visual), we build three compound gates: personalization (g_p), acoustic (g_a), and visual (g_v). The context information is transmitted via the gates as follows:

$$\begin{aligned} g_a &= [H \oplus H_a]W_a + b_a \\ g_v &= [H \oplus H_v]W_v + b_v \\ g_p &= [H \oplus H_p]W_p + b_p \end{aligned} \tag{5}$$

where $\oplus$ denotes a concatenation operation. W_a, W_v, W_p ($\in \mathbb{R}^{2dXd}$) and b_a, b_v, b_p ($\in R^{dX1}$) are parameters. The final contextualized attended vector ($\hat{H}$) is computed as follows:

$$\hat{H} = H + g_p \odot H_p + g_a \odot H_a + g_v \odot H_v \tag{6}$$

4.3 MMCS Generation

MMCS is our primary task, which is being comprehended with the other task, intent recognition. We take the attended multi-modal encoder representation vector ($\hat{H}$) and pass it to a linear layer for intent identification. The vector ($\hat{H}$) is fed to the decoder's multi-head attention layer as key and value, with the key as the hidden representation of the medical concern summary. The infused information is processed with the traditional transformer's layers and computes the vocabulary's probability distribution. We have utilized a joint categorical cross-entropy loss function, which is the sum of loss functions of classification (CL) and generation (GL) tasks, i.e., $L = \alpha_1 * CL + \alpha_2 * GL$ and $\alpha_1 (= 0.2) + \alpha_2 (= 0.8) = 1$.

4.4 Experimental Details

We have utilized the PyTorch framework for implementing the proposed model[2]. The generation models have been trained, validated, and evaluated with 80%, 6%, and 14% samples of the *MM MediConSummation* dataset, respectively. The hyperparameter values, which are selected empirically, are as follows: sequence length (480), output max len (50), learning rate (3e-05), batch size (16), and activation function (ReLU). The different baselines and state-of-the-art models are listed as follows:

[2] The dataset and source code are available at `https://github.com/AnonymousRW/MMCSG`

- **Seq2Seq-Transformer** It is a transformer-based sequence-to-sequence model [21], which takes a combined representation of transcript, audio, and video features as input and generates a medical concern summary.
- **BART** It is a denoising autoencoder model that is trained to reconstruct corrupted sentences [16].
- **T5** It [19] 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 [15] is a fusion model that incorporates an additional adapter-based layer in the encoder of BART to infuse information from different modalities.
- **MMCSG** is the proposed model with the proposed contextualized M-modality fusion mechanism only (without the multi-task intent recognition and MMCG generation setting).
- ***IR-MMCSG*** is the proposed multi-tasking, multi-modal intent identification, and medical concern summary generation model, incorporated with contextualized M-modality fusion mechanism.

5 Result and Discussion

The purpose of the proposed multi-task framework is to enhance the performance of the primary task, MCSG, by utilizing the additional task of intent recognition. Thus, the results and analysis of MCSG are emphasized as the main focus in all task combinations.

5.1 Experimental Results

The obtained performance by different multi-modal medical concern summary generation models are reported in Table 3. Furthermore, we also investigated these models' efficacies for doctor summary generation, and results are presented in Table 4. We ran experiments for ten iterations with different random seeds and reported the average values. The reported values in the following tables are statistically significant as the p-values obtained from Welch's t-test [27] at 5% significance level are less than 0.05.

Ablation Study In order to understand the effectiveness of various components within the proposed model, we carried out an ablation study involving different combinations of these components. The obtained result has been reported in Table 5. The results show that the model's performance is improving as the various different modalities are infused together to represent the context.

Human Evaluation We have also conducted a human evaluation of all test samples. In this assessment, two medical domain experts and one researcher (other than the authors) were employed to evaluate the generated medical concern summaries (twenty samples for each model) without revealing the models' names. The samples are assessed based on the following five metrics: *domain relevance (DR), adequacy, fluency, informativeness (Info), and patient's personal context coherence (PC)* on a scale of 0 to 5. The obtained scores are presented in Table 6.

Key Observations The main observations and insights are as follows: **(i)** The proposed medical concern summary generation model outperforms traditional sequence-to-sequence and transfer-learning-based generation models by a large margin, high-

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

Model	BLEU-1	BLEU-2	BELU-3	BELU-4	BLEU	ROUGE - 1	ROUGE - 2	ROUGE- L	METEOR
Seq2Seq-Transformer [21]	18.65	10.93	3.85	1.23	8.67	24.03	8.62	22.16	21.36
T5 [19]	19.67	11.73	5.92	1.93	9.81	28.23	11.01	26.48	23.44
BART [16]	19.46	12.47	6.90	2.62	10.36	26.85	12.92	26.01	32.46
MAF [15]	21.89	13.05	6.05	3.26	11.06	30.04	12.43	26.93	30.10
MMCSG w/o visual	21.63	13.57	7.33	2.97	11.37	30.60	14.27	27.56	34.44
MMCSG	23.26	14.13	7.80	**3.81**†	12.25	31.47	14.16	28.51	33.60
IR-MMCSG	**23.72**†	**14.68**†	**7.88**†	2.98	**12.31**†	**32.16**†	**14.41**†	**29.61**†	**35.87**†

Table 4: Performances of different models for doctors' impression summary generation. Here, † indicates statistical significant findings ($p < 0.05$ at 5% significance level).

Model	BLEU-1	BLEU-2	BELU-3	BELU-4	BLEU	ROUGE - 1	ROUGE - 2	ROUGE- L	METEOR
T5 [19]	17.37	9.18	2.51	0.088	7.28	26.70	7.83	23.60	18.44
BART [16]	20.87	9.96	1.22	0.417	8.12	26.86	8.04	22.84	23.35
MAF [15]	20.88	11.53	3.69	2.69	9.70	26.99	9.83	24.03	23.89
MMCSG w/o visual	21.10	11.66	2.52	0.934	9.05	28.59	10.17	25.17	26.88
MMCSG	23.28	12.38	3.71	**3.13**†	10.63	30.46	10.41	26.35	27.52
IR-MMCSG	**23.78**†	**12.67**†	**4.01**†	2.57	**10.76**†	**31.23**†	**10.86**†	**26.43**†	**29.17**†

Table 5: Performance of the proposed model with different modalities. Here, T, P, A, and V indicate transcript, personality information, audio, and video features, respectively.

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE- L	METEOR
T	11.00	28.88	14.05	26.88	33.71
T + A	11.37	30.60	14.27	27.56	35.71
T + P	11.95	30.78	14.64	27.80	34.91
T + V	11.73	30.89	14.44	28.18	35.80
T + A + V	11.56	30.46	**14.84**	28.48	35.82
T + A + V + P	**12.31**	**32.16**	14.41	**29.61**	**35.87**

Table 6: Human evaluation for medical concern summary generation models

Model	DR	Adequecy	Fluency	PC	Info	Avg.
Seq2Seq [21]	2.86	2.72	3.88	2.40	3.61	3.09
BART [16]	3.10	3.16	4.22	2.65	3.80	3.37
MAF [15]	3.28	3.56	4.38	2.80	3.94	3.59
MMCSG	3.44	3.81	4.55	**3.14**	3.84	3.76
IR-MMCSG	**3.82**	**3.94**	**4.74**	3.08	**4.06**	**3.93**

lighting the importance of (a) task-specific conditioning and (b) the incorporated contextualized M-modality fusion (BART with early fusion- a simple concatenation of modalities vectors). **(ii)** For MCSG, visual modality (movements and expressions) infusion with text was more important than audio and demographic information (Table 5). **(iii)** In human evaluation, we observed that the multi-task model incorporating intent representation generates a significantly more contextualized and informative summary of medical concerns (Table 6 and Fig. 8).

5.2 Findings to Research Questions

Based on the experiments, we report the following answers (with evidence) to our investigated research questions (RQs).

RQ 1: *Do visuals and patient expression help in identifying patient key concern and generating adequate medical advice for the same?* The proposed MMCSG model significantly outperforms the medical concern summary generation without visual information across all evaluation metrics (Table 3 and Table 5 - T vs. $T + V$ & $T + A$ vs. $T + A + V$). Furthermore, a similar trend has been observed for doctor suggestion summary generation (Table 4). The enhanced findings strongly establish the effectiveness of utilizing visual cues, such as patients' facial expressions and body movements, in accurately identifying their medical conditions and responding accordingly.

RQ 2:*Can the patient's demographic context aid in generating an appropriate and relevant medical concern summary /suggestion?* Some behaviors, namely facial expressions/body movement and medical conditions, are heavily influenced by demographic information such as age and gender. Therefore, the proposed contextualized M modality fusion aided *IR-MMCSG* (Table 3 - MAF vs. MMCSG, Table 4 - MAF vs. MMCSG and Table 5 - T vs. $T + P$ & $T + A + V$ vs. $T + A + V + P$) that exploits the information to analyze and constrain different modalities performed significantly better than the non-context aware models.

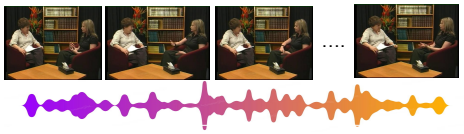
RQ 3: *Does any correlation exist between medical concern summary generation and user intent identification?* If a psychiatrist is aware of the reason for a patient's visit, it becomes more straightforward to identify the patient's primary mental issue and suggest him/her accordingly. The observed results firmly support the hypothesis, revealing that the multi-tasking *IR-MMCSG* clearly outperforms the single-task model, MMCSG (Table 3 and Table 4 - MMCSG vs. *IR-MMCSG*). Furthermore, we also observed a significant improvement by *IR-MMCSG* model in human evaluation (Table 6) across different metrics.

6 Analysis

We have analyzed the proposed model's generated medical concern summaries and different models' behavior on some test cases, which are presented in Fig. 8. The comprehensive analyses lead to the following key observations: (i) The proposed *IR-MMCSG* generates medical concern summaries (Fig. 8) that include (a) a comprehensive and contextualized understanding of the patient's concerns and (b) a sense of the discussion that will be undertaken during the session. **(ii)** During the human evaluation, we found a significant number of cases where our model has also generated a cause of abnormality in the MCS. **(iii)** We observed that the models added additional words in the medical concern summary, leading to low evaluation scores despite being relevant and informative.

7 Conclusion

In this paper, we introduce a new task of generating a succinct summary of the primary concerns and expectations expressed by the initiator of a conversation. We curated



Transcript
<1> Amanda thanks for coming in today you ..... <18>I'm just not feeling how I think I meant to feel um it's difficult sometimes ..... <87> like I'm a bad mother already but uh ..... <96> length they've been a really rough six *<97>* weeks appears to me and I can really ..... *<111> Thank you*

Gold Medical Concern Summary: The patient is a new mother who is having difficulties to care for her child and accept her role as a mother. She believes that she is not receiving enough support from family.

Generated Multimodal Medical Concern Summaries (MMCS)

BART	MAF	MMCSG	IR-MMCSG
The patient finds difficulty in accepting her as a mother.	The patient is a new mother and finds herself in a complex situation. She needs suggestions to be a good mother.	The patient is a new mother who is having difficulty caring for her child. She believes that no one is supporting her.	The patient is a new mother and she finds it hard to take care of the child. She did not consider herself as a good mother. She is not getting support from her family.

Fig. 8: A case study- performance of different baselines and the proposed *IR-MMCSG* model for a common test case

the first multi-modal medical concern summary generation (*MM-MediConSummation*) corpus annotated with medical concern summary, user demographic information, user intent, and summary of doctors' suggestions. We proposed a multi-tasking multi-modal intent recognition and medical concern summary generation (*IR-MMCSG*) model incorporated with a novel adapter-based contextualized multi-modality fusion mechanism for analyzing acoustics and visual features with demographic and personality context. With the obtained results of various sets of experiments and human evaluation, we found firm evidence of the efficacy of the proposed *IR-MMCSG* model and the infused fusion mechanism over existing state-of-art methods. The obtained improvements establish the crucial role of facial expression/movement behavior and demographic context in identifying patient's medical concerns and generating an adequate summary for them. In the future, we would like to build an explainable multi-modal medical concern summary generation that generates medical concern summary generation along with evidence highlighting video spans.

8 Ethical Consideration

We strictly followed the medical research's ethical[3], and regulatory guidelines particular to psychiatrist research [2] during the dataset curation process. We have not added or removed any utterances/medical entity from the conversation. The curated dataset does not reveal users' identities, such as their names. The names have been replaced with some synthetic names. The annotation guidelines were provided by two psychiatrists, and the curated 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 use the dataset and conduct the research. Thus, we assert with confidence that the dataset, along with its comprehensive creation protocol, stands in full compliance with the ethical and clinical imperatives of our discipline.

[3] https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/

References

1. Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Video summarization using deep neural networks: A survey. Proceedings of the IEEE **109**(11), 1838–1863 (2021)
2. Avasthi, A., Ghosh, A., Sarkar, S., Grover, S.: Ethics in medical research: General principles with special reference to psychiatry research. Indian Journal of Psychiatry **55**(1), 86 (2013)
3. Barnes, R.K.: Conversation analysis of communication in medical care: description and beyond. Research on Language and Social Interaction **52**(3), 300–315 (2019)
4. Bay, M., Bruneß, D., Herold, M., Schulze, C., Guckert, M., Minor, M.: Term extraction from medical documents using word embeddings. In: 2020 6th IEEE Congress on Information Science and Technology (CiSt). pp. 328–333. IEEE (2021)
5. Chintagunta, B., Katariya, N., Amatriain, X., Kannan, A.: Medically aware gpt-3 as a data generator for medical dialogue summarization. In: Machine Learning for Healthcare Conference. pp. 354–372. PMLR (2021)
6. Dreisbach, C., Koleck, T.A., Bourne, P.E., Bakken, S.: A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. International journal of medical informatics **125**, 37–46 (2019)
7. Enarvi, S., Amoia, M., Teba, M.D.A., Delaney, B., Diehl, F., Hahn, S., Harris, K., McGrath, L., Pan, Y., Pinto, J., et al.: Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In: Proceedings of the first workshop on natural language processing for medical conversations. pp. 22–30 (2020)
8. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1459–1462 (2010)
9. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
10. Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al.: The distress analysis interview corpus of human and computer interviews. Tech. rep., UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Huang, J.H., Murn, L., Mrak, M., Worring, M.: Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. pp. 580–589 (2021)
13. Jeblee, S., Khattak, F.K., Crampton, N., Mamdani, M., Rudzicz, F.: Extracting relevant information from physician-patient dialogues for automated clinical note taking. In: Proceedings of the tenth international workshop on health text mining and information analysis (LOUHI 2019). pp. 65–74 (2019)
14. Joshi, A., Katariya, N., Amatriain, X., Kannan, A.: Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 3755–3763 (2020)
15. Kumar, S., Kulkarni, A., Akhtar, M.S., Chakraborty, T.: 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). pp. 5956–5968 (2022)
16. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)

17. Malhotra, G., Waheed, A., Srivastava, A., Akhtar, M.S., Chakraborty, T.: Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. pp. 735–745 (2022)
18. Nittari, G., Khuman, R., Baldoni, S., Pallotta, G., Battineni, G., Sirignano, A., Amenta, F., Ricci, G.: Telemedicine practice: review of the current ethical and legal challenges. Telemedicine and e-Health **26**(12), 1427–1437 (2020)
19. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
20. Shang, X., Yuan, Z., Wang, A., Wang, C.: Multimodal video summarization via time-aware transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1756–1765 (2021)
21. Shi, T., Keneshloo, Y., Ramakrishnan, N., Reddy, C.K.: Neural abstractive text summarization with sequence-to-sequence models. ACM Transactions on Data Science **2**(1), 1–37 (2021)
22. Shi, X., Hu, H., Che, W., Sun, Z., Liu, T., Huang, J.: Understanding medical conversations with scattered keyword attention and weak supervision from responses. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 8838–8845 (2020)
23. Singh, G.V., Ghosh, S., Ekbal, A., Bhattacharyya, P.: Decode: Detection of cognitive distortion and emotion cause extraction in clinical conversations. In: European Conference on Information Retrieval. pp. 156–171. Springer (2023)
24. Song, Y., Tian, Y., Wang, N., Xia, F.: Summarizing medical conversations via identifying important utterances. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 717–729 (2020)
25. Stratigi, M., Kondylakis, H., Stefanidis, K.: Multidimensional group recommendations in the health domain. Algorithms **13**(3), 54 (2020)
26. Wang, Y., Liang, W., Huang, H., Zhang, Y., Li, D., Yu, L.F.: Toward automatic audio description generation for accessible videos. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2021)
27. Welch, B.L.: The generalization ofstudent's' problem when several different population variances are involved. Biometrika **34**(1/2), 28–35 (1947)
28. Weld, H., Huang, X., Long, S., Poon, J., Han, S.C.: A survey of joint intent detection and slot filling models in natural language understanding. ACM Computing Surveys (CSUR) (2021)
29. Wosik, J., Fudim, M., Cameron, B., Gellad, Z.F., Cho, A., Phinney, D., Curtis, S., Roman, M., Poon, E.G., Ferranti, J., et al.: Telehealth transformation: Covid-19 and the rise of virtual care. Journal of the American Medical Informatics Association **27**(6), 957–962 (2020)
30. Zhu, J., Li, H., Liu, T., Zhou, Y., Zhang, J., Zong, C.: Msmo: Multimodal summarization with multimodal output. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 4154–4164 (2018)