MDS: A Fine-Grained Dataset for Multi-Modal Dialogue Summarization

Zhipeng Liu[1], **Xiaoming Zhang**[1,2]*, **Litian Zhang**[1], **Zelong Yu**[1]
[1] School of Cyber Science and Technology, Beihang University, Beijing, China
[2] State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
{lzpeng, yolixs, litianzhang, azyu11}@buaa.edu.cn

Abstract

Due to the explosion of various dialogue scenes, summarizing the dialogue into a short message has drawn much attention recently. In the multi-modal dialogue scene, people tend to use tone and body language to illustrate their intentions. While traditional dialogue summarization has predominantly focused on textual content, this approach may overlook vital visual and audio information essential for understanding multi-modal interactions. Recognizing the established field of multi-modal dialogue summarization, we develop a new multi-modal dialogue summarization dataset (MDS), which aims to enhance the variety and scope of data available for this research area. MDS provides a demanding testbed for multi-modal dialogue summarization. Subsequently, we conducted a comparative analysis of various summarization techniques on MDS and found that the existing methods tend to produce redundant and incoherent summaries. All of the models generate unfaithful facts to some degree, suggesting future research directions. MDS is available at https://github.com/R00kkie/MDS.

Keywords: Multi-Modal Learning, Dialogue Summarization, Language Resources

1. Introduction

Benefiting from the development of communication technology, people contact each other at any time. Due to the explosion of various conversation scenes, grasping critical information from redundant and complex conversion content is essential. Therefore, some works focus on summarizing dialogue from various domains, such as meeting (Janin et al., 2003; Carletta et al., 2006; Zhong et al., 2021; Li et al., 2019), daily chat (Gliwa et al., 2019; Chen et al., 2021; Zhang et al., 2023), film (Malykh et al., 2020; Zhu et al., 2021; Chen et al., 2022), customer service (Zhao et al., 2021; Lin et al., 2021; Zou et al., 2021), and medical conversation (Song et al., 2020; Joshi et al., 2020; Zhang et al., 2021a). In most realistic cases, dialogues occur in a multi-modal scene, in which the data contains the dialogue text and the audial-visual accompaniment of the dialogue background. However, previous dialogue summarization datasets only focus on the raw text content, which cannot learn the vital information from the multi-modal content in the multi-modal dialogue scenes.

When we talk to others, we tend to use tone and body language to illustrate our intentions, which can not be directly captured by the text content of dialogues. Visual and audial information in the entire conversion scene also provides crucial information. For example, some postures and expressions indicate the attitude of a person and the critical content of a talk, and the intonation and pauses in a speech can also indicate the importance of the content. Such visual and audial information is crucial to the whole dialogue conversion. Relying solely on textual information may result in the omission of crucial details originating from the visual and audial modalities, rendering it ineffective when generating summaries for multi-modal dialogue scenes. Consequently, it becomes imperative to incorporate multi-modal information into summarizing dialogue.

However, there is still a significant challenge in multi-modal dialogue summarization. First, few datasets are available for multi-modal dialogue summarization. It is time-consuming to annotate the multi-modal dialogue summarization (AMI only includes 137 pieces of data). Most previous dialogue summarization datasets focus on studying various domains but not various modalities. On the other hand, there are some multi-modal summarization datasets. However, the different content modalities are generally asynchronous. Synchronous multi-modal information can realize multi-modal data fusion better. The temporal relationship and correlation between them can be maintained by processing data of different modals simultaneously.

To tackle the challenge, we construct a novel multi-modal dialogue summarization dataset, MDS. We compare MDS with other summarization datasets in Table 1. MDS differs in two aspects. On the one hand, compared to previous dialogue summarization datasets, MDS contains multi-modal content, including over 16,000 minutes of video clips with images and audio. On the other hand, compared to conventional multi-modal summarization datasets, MDS provides synchronous audio and video data from the clips. To generate fine-grained information, a video scene cutter based on three-modality voting is proposed to split the

* Corresponding authors

LREC-COLING 2024, pages 11123–11137

	Data Size	Lang.	Input Tokens(Avg)	Speakers(Avg)	Image	Audio	Video	Syn.
Dialogue Summarization Datasets								
AMI Carletta et al. (2006)	137	EN	4757.0	4.0	Yes	Yes	Yes	Yes
ICSI Janin et al. (2003)	59	EN	10189.0	6.2	No	No	No	-
SAMSum Gliwa et al. (2019)	16.4k	EN	83.9	2.2	No	No	No	-
QMSum Zhong et al. (2021)	1.8k	EN	9069.8	9.2	No	No	No	-
SumScreen Chen et al. (2022)	26.9k	EN	6612.5	28.3	No	No	No	-
CSDS Lin et al. (2021)	1.1k	ZH	213.86	2.0	No	No	No	-
Multi-Modal Summarization Datasets								
MSMO Zhu et al. (2018)	314.6k	ZH	722.68	1.0	Yes	No	No	No
Hierarchical Zhang et al. (2022)	62.9k	ZH	955.26	1.0	Yes	No	No	No
Our Datasets								
MDS	11.3k	ZH/EN	186.77	3.4	Yes	Yes	Yes	Yes

Table 1: The comparison of different dialogue summarization datasets, multi-modal summarization datasets, and MDS.

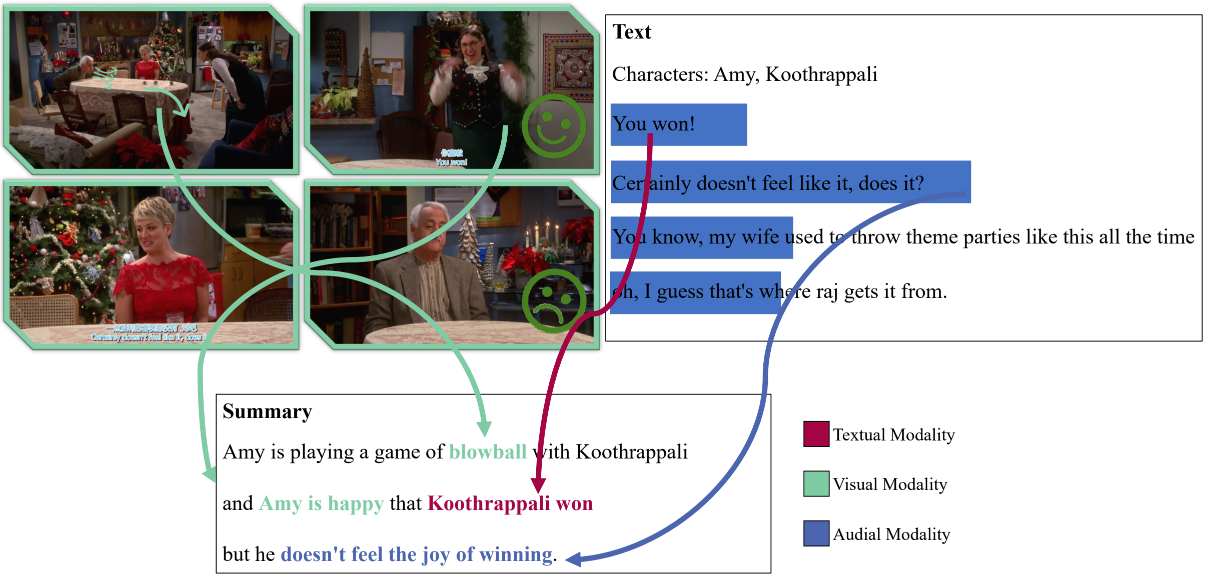

Figure 1: An example from MDS. A case consists of dialogue utterances, a video clip, and a human-written summary.

videos into fine-grained video clips. The annotators are asked to watch the clips and write a target summarization for the multi-modal dialogue. Then, several methods are empirically evaluated on MDS, including the conventional extractive and abstractive summarization models. Analytical experiments show that MDS is a highly abstractive summarization dataset, benefiting from multi-modal information. The poor performance on conventional extractive summarization models indicates that other modal information fuses in MDS. For example, in Figure 1, red is for text modality, blue is for audial modality, and green is for visual modality. In the summary text, the visual modality supplies extra information, "*blowball*", which cannot be generated from the textual modality. The mood of the characters is also captured by the visual modality, "*Amy is happy.*" The visual modality provides critical facts that do not appear in the textual modality. When we try to emphasize crucial points and draw attention, we usually raise our voices involuntarily. In the dialogue textbox, we denote the volume of each sentence by a histogram, and the second sentence is the most boisterous one. In the summary text, the noisiest sentence helps to find the key point "*doesn't feel*".

There are two contributions of this paper: (1) We introduce a multi-modal dialogue summarization dataset that expands the existing body of resources with its unique features and scope; (2) we build an annotation framework for multi-modal dialogue summarization, including a video scene-cutting model and a set of standards.

2. Related Work

As a data-driven task, several datasets have been proposed to promote dialogue summarization.

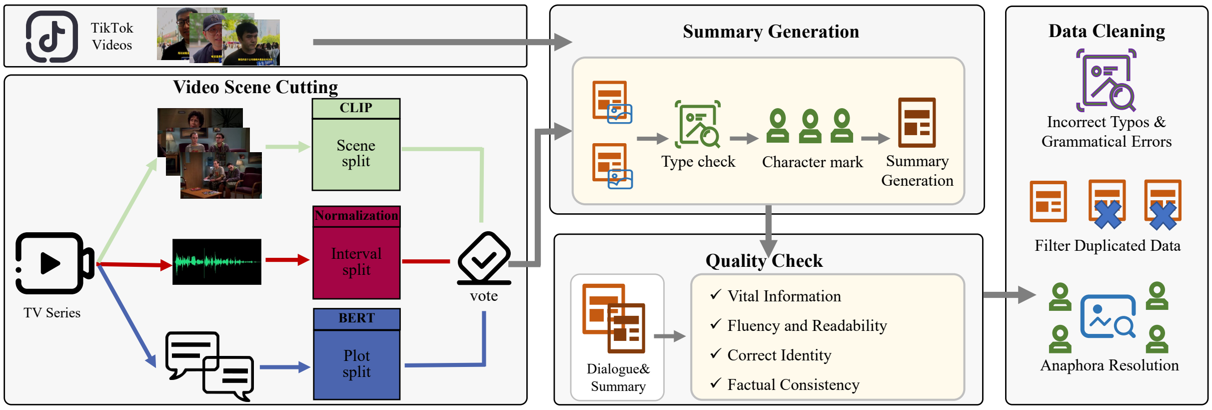

Figure 2: The overview of dataset construction.

SAMSum Gliwa et al. (2019) is the first large-scale dataset on dialogue summarization, which belongs to daily chat. **DialogSum** Chen et al. (2021) is also a daily chat dataset. SAMSum is written by one human, and DialogSum is derived from the existing dialogue dataset. **GupShup** Mehnaz et al. (2021) is a multi-lingual version of SAMSum, focusing on the code-switch problem in Hindi-English.

AMI Carletta et al. (2006) and **ICSI** Janin et al. (2003) are meeting datasets related to working and research scenarios. AMI includes microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard but small in scale. Both datasets are included in **QMSum** Zhong et al. (2021) and further broadened for query-summary pairs, particularly for lengthy and complex meetings.

SumTitles Malykh et al. (2020) and **SummScreen** Chen et al. (2022) crawled film data on the Internet, corresponding transcripts for dialogues and recaps for summaries, which are generally coarse-grained. **MediaSum** Zhu et al. (2021) comprises a rich collection of interviews extracted from prominent TV news programs.

TDS Zou et al. (2021), **TODSum** Zhao et al. (2021), and **CSDS** Lin et al. (2021) are for customer service. The summaries of TDS are from an agent perspective. TODSum contains more complex multi-domain conditions. CSDS summarizes JDDC Chen et al. (2020) and supplies a fine-grained annotation, including both agent and user perspectives. Such detailed annotations empower researchers to delve deeper into the nuances of customer service interactions.

Chunyu Song et al. (2020) is based on online health platforms. **Dr. Summarize** Joshi et al. (2020) is collected from a telemedicine platform. **DP** Zhang et al. (2021a) converts doctor-patient conversation audio to text contents and treats it as a text-only task. Regrettably, DP does not further develop for audial modality after transcription.

Existing dialogue summarization datasets are far from the actual conversation scene. The lack of multi-modal data fusion can not capture non-verbal information such as emotion, intonation, expression, and action, which do not help grasp the context and semantics of the dialogue and produce a more accurate and coherent dialogue summary.

3. Dataset Construction

MDS is comprised of 11305 dialogs divided from TV series ("The Big Bang Theory" sitcom) and TikTok video clips and human-written summaries under the reference of the corresponding recaps. Figure 1 shows an example in MDS. The summaries in MDS are meticulously created through four steps, ensuring the inclusion of all essential information while maintaining a coherent and informative narrative structure.

- **Video Scene Cutting**. We propose a multi-modal video scene cutter to split one episode from a TV series with several topics into fine-grained video clips.
- **Summary Generation**. We provided annotators with detailed and nuanced guidance and relevant recaps to ensure that the resulting summaries were meticulously crafted.
- **Quality Check**. There are cross-inspections between different annotators and random checks after annotation to identify potential discrepancies or inconsistencies.
- **Data Cleaning**. Dialogs with low information are removed. For the remaining dialogs, a linguistic data cleaning is performed. Then we clean the data and split MDS into three sets.

3.1. Video Scene Cutting

Conventional dialogue summarization datasets in TV series generally regard one episode as a piece of data, roughly with brief recaps crawled from the Internet for summary texts. It often leads to topic confusion since the length of the dialogue is overlong, and the crawled recaps are simplistic. Before annotators are asked to generate summaries under the existing recaps for a single dialog, we build a video scene cutter to split a video into a trail of clips. A coherent video clip is often accompanied by internal semantic, voice, and place coherence. We build a multi-modal method of video scene cutting according to this. There are three modal data of a dialog in the video: visual, audial, and text modality. We define a break-point for each modality. A break-point is the second that splits two clips with the complete semantics of a long video. For the visual modality, **CLIP** (Radford et al., 2021) is applied to find the break-point. For images drawn from each second:

$$I = \{i_1, i_2, \dots, i_n\} \tag{1}$$

Feed I and twenty-nine location labels appearing in the playscript with stage directions into the Vit-B/32 for location classification as $L = \{\ell_1, \ell_2, \dots, \ell_n\}$. ℓ_i represents the location label of i-th image. If ℓ_{i-1} is not the same as ℓ_i, the i-th second of this video is considered as a visual break-point. For example, if ℓ_{i-1} is *bedroom* and ℓ_i is *kitchen*, the i-th second is a visual break-point. The occurrence of topic changes is often accompanied by transitions in physical locations. It is worth noting that when there is a shift in scenes or settings, there is a heightened probability of encountering a distinct break-point in the ongoing discussion or conversation.

BERT (Devlin et al., 2019) is applied to find the break-point in the text modality. Specifically, paring off the subtitle text $T = \{t_1, t_2, \dots, t_n\}$ into pairs:

$$SP = \{(t_1, t_2), (t_2, t_3), \dots, (t_{n-1}, t_n)\} \tag{2}$$

and feed SP into the BERT for the next sentence prediction (NSP), reasoning whether t_i is the next sentence of t_{i-1}. The output is generated from $\{1, 0\}$. Zero represents that t_i is the next sentence of the t_{i-1}, so (t_{i-1}, t_i) is not a text break-point. One represents that the second corresponding to (t_{i-1}, t_i) is a text break-point. The instinct of textual break-point is that the sentences in the same conversation share semantics consistency. Sentences in different conversations are supposed to lose semantic coherence detected by NSP.

For the audio, the volume was normalized.

$$A = \{a_1, a_2, \dots, a_n\} \tag{3}$$

If the voice value of a second a_i is less than the threshold value, this is an audial break-point. Every new act should begin with a brief pause, distinguishing from the previous scene.

After finding the break-point from three modalities, we apply a voting mechanism to determine where to cut. For a given second, more than two of the three modalities vote, we consider it a 'real' break-point and cut the video here. Only videos from TV series need to be cut; videos from TikTok can skip this step because most of them are short enough to have a concentrated topic.

3.2. Summary Generation

The annotators are asked to write summaries for clips divided above, under the reference of the corresponding video clips, textual transcripts, and recaps. The annotation adheres to three criteria: (1) Type check. If the video content is not about dialogue, skip it. (2) Character mark. The annotator was instructed to complete the anaphora resolution and mark the speaker identity appearing in clips. (3) Summary generation. The annotator summarizes dialogues and derives core information through three modalities. Follow the above annotation guidelines to ensure that annotators follow consistent annotations. Communicate with annotators regularly, answer questions, and provide feedback to ensure they understand and perform tasks correctly. Annotators should identify the topic of the conversation and determine what is essential to it. The summary content should be logically structured and organized in chronological order, topic order, or importance order. The summary should concisely summarize the key content, avoid verbosity and unnecessary details, and use clear and unambiguous language to make it easier to understand. The subsequent quality control module ensures the quality of the work of annotators.

3.3. Quality Check

To ensure quality, cross-inspection between different annotators is performed after annotation. The annotator is paid to find incapable samples, and the annotators whose annotation is found with mistakes are punished while inspecting. The cross-inspection adheres to four criteria: (1) Summary contains all vital information in the dialogue. (2) The Summary is fluent in presentation and easy to follow. (3) There are no vague problems in pronoun reference, and the speaker identity is labeled precisely. (4) There are no Factual inconsistencies in the summary. After the second cross-validation, 15% cases are manually checked by us. If errors are found in one bunch, corresponding annotators are asked to re-annotate the whole bunch and repeat the process of inspection and sampling.

3.4. Data Cleaning

We delete incorrect typos and grammatical errors and filter out duplicated data based on text similarity. First, we delete clips with low information and repetition, like the beginning and end of each episode, and too short clips, even without one word. The next step is anaphora resolution. When we find personal pronouns, we convert pronouns to corresponding character names. Then, delete the meaningless function word from the text, like "*then*", "*later*", "*moreover*", "*furthermore*" and "*next*".

4. Dataset Analysis

MDS encompasses over two hundred distinct topics, making it an invaluable resource for research and analysis of multi-modal dialogue summarization. MDS provides synchronous audio and video data from the clips, and an insightful experiment conducted on MDS reveals its multi-modal nature, showcasing the incorporation of novel words and expressions that go beyond the textual domain. The dialogues within MDS maintain a succinct nature, and the short ones make up about 80% of the total. Furthermore, when evaluating the performance of extractive models on MDS, the effectiveness and advantages of leveraging multi-modalities in dialogue summarization become evident.

4.1. Split Coverage

We designed experiments to successfully verify the effectiveness of the video-cutting model. Inspired by **Intersection over Union (IoU)** (Yu et al., 2016) and **ROUGE-L** (Lin, 2004), which is based on the longest common subsequence, we proposed new evaluative criteria to measure the degree of correlation between the short video clips generated by the model and the label ones. The intersection region and union region between two video clips are calculated. These regions can be represented by timestamps or frame indices. We define **Video-IoU (VIou)**:

$$VIoU = \frac{Intersection\_Duration}{Union\_Duration} \tag{4}$$

In this formulation, "Intersection Duration" represents the temporal intersection of the short video segment generated by the model and the original video time period, and "Union Duration" represents their temporal union.

$$Split\_Coverage = \frac{1}{M}\sum_{i=1}^{M}\max(VIoU(c_i, l_j)) \tag{5}$$

c_i represents a short video clip generated by models. l_i represents a short video clip generated by human annotators. For the video clips generated by the cutting model, VIoU is calculated with all labeled videos, and the maximum value is taken. The average VIoU value of all short video clips is calculated to obtain the performance index of the whole video cutting model, **Split Coverage**. For comparison, we pick 100 minutes from raw videos and label them.

Model	Split Coverage
Audio	0.1859
Image	0.0994
Text	0.2940
Audio&Image	0.3680
Audio&Text	0.4210
Image&Text	0.2455
Proposed Model	**0.6564**

Table 2: Comparison of video cutting performance in different modality models.

Count	Name
	AMI
102	remote_group_buttons_design
23	project_manager_remote_team
12	group_project_design_research
	SumScreen
23377	time_home_baby_room
183	xander_baby_something_truth
66	brody_baby_father_son
28	slater_president_blessing_mess
	MDS
2728	ndustry_girl_future_student
991	okay_tops_cool_right
137	actor_actors_profession_star
118	door_noise_sound_knock
109	house_angry_home_air
107	apartment_tenant_weeks_guys
106	film_movie_theater_movies

Table 3: Top topics in MDS, AMI, and SumScreen detected by BERTopic

Our multi-modal model demonstrates a significant advantage over single-modal models in evaluating short video clip cutting. By seamlessly integrating audio, image, and text modalities, our model achieves a correlation score of 0.6564, surpassing all the other models. This resounding success crystallizes the central thesis of this study – the undeniable advantage of a multi-modal approach in video cutting.

4.2. Topic Analysis

We use **BERTopic** (Grootendorst, 2022), a topic modeling technique, to analyze the summary topics in MDS. We depict the top dataset topics in Table 3. There are 261 topics in total, and the amount of topics varies from 2728 to 10. This wide-ranging coverage underscores the breadth and depth of subjects MDS covers, especially compared to other datasets such as AMI, which contains a mere three

	uni-gram	bi-gram	tri-gram	four-gram
MDS	2.16	**54.14**	**91.79**	**98.53**
CSDS	3.06	31.14	68.96	85.44
AMI	**3.67**	43.80	68.32	73.31
SumScreen	2.36	29.27	59.59	82.63

Table 4: Fraction (%) of n-grams in the output summaries that do not appear in the inputs

topics, or SumScreen, which encompasses six topics. The multi-domain nature of the data within MDS plays a pivotal role in model training to adapt to a diverse range of scenarios and tasks effectively. By training on data sourced from different domains, models will have the opportunity to learn a broader spectrum of feature representations and patterns. Such data features in MDS instilled within the training process empower models to transfer knowledge and apply learned insights from one domain to another, thus fostering a more robust and adaptable framework for tackling the complexities in a multitude of scenarios and tasks. The outcome of the topic experiment demonstrates that MDS presents an arduous and intricate testbed for multi-modal dialogue summarization, specifically designed to evaluate model generalization.

4.3. Novel Words

We empirically compare MDS with existing dialogue summarization datasets. CSDS is a Chinese dataset for customer service. AMI and SumScreen are both English datasets. Table 4 compares the percentages of novel n-grams in the reference summary against the source document/dialogue. The result intuitively reflects the level of abstraction of annotated summaries. MDS outperforms in most metrics well. MDS is absolutely 10.34%, 22.93%, and 13.09% higher than other datasets on bi-gram, tri-gram, and four-gram. A Chinese dialogue dataset, CSDS, is also employed to remove the impact of language characteristics. Unlike the dialogue summarization datasets proposed before, MDS is a multi-modal dataset. The result of the novel word experiment verifies our original intention of presenting MDS, that information from other modalities complements the text.

4.4. Data distribution

We split instances by the number of words in reference. The statistics of the splits are shown in Table 5. The short summary makes up about 80% of the total. The section of video scene cutting splits the videos into segmented clips with the smallest complete semantic fragment. The statistics indicate the effectiveness of video scene cutting. Each episode is usually 22 minutes long, so cutting it into 10-20 video clips is appropriate. Each segment is about 1-2 minutes long.

	Train	Dev	Test
Short (summary<50 words)	7811	974	974
Medium (50 words≤summary<100)	924	113	113
Long (100 words≤Summary)	316	40	40
SUM	9051	1127	1127

Table 5: Statistics of train/dev/test splits and short/medium/long splits for MDS, short/medium/long split by the number of words in reference

4.5. Improvement from Multi-modal

Generally, compared to mono-modal summarization, multi-modal summarization is expected to bring extra information to the generated summary (Zhang et al., 2024). To measure the improvement brought by multi-modal information, we employ three extractive summarization models, TextRank (Mihalcea and Tarau, 2004), BertSum (Liu, 2019), and CentroidSum (Rossiello et al., 2017), to evaluate the performance of MDS and mono-modal datasets. These models extract summaries from existing texts and lack supplements from other modals. If a dataset yields high scores for extractive models, it suggests that it predominantly relies on textual information. We hypothesize that given the inferior performance of three extractive methods on MDS compared to other dialogue summarization datasets, incorporating multi-modal information will enhance summary quality.

The results are shown in Table 6. We use the ROUGE scores here. MDS sees the lowest ROUGE scores in all terms of models. None of the ROUGE scores exceeded 27. The experiment indicates the improvement brought by multi-modal information compared with text-only dialogue summarization datasets. When human annotators summarize the dialogue, they receive information from multiple modalities. As shown in Figure 1, the phrase "*playing a game of blowball*" never appears in the dialogue text. Annotators see the "*blowball*" and write the word in the final summary. Moreover, extractive models cannot handle this information, which has never appeared in the text before. However, English and Chinese belong to different language families, with significant disparities. To mitigate the influence of linguistic characteristics, we utilized two English datasets alongside a Chinese dataset (CSDS). Although homologous languages exhibit more minor differences than those between distinct languages, there remains an evident contrast between MDS and CSDS.

Furthermore, aimed to enhance the understanding of multi-modal improvement, we developed an annotation website that extracts nouns and pro-

	TextRank			BertSum			CentroidSum		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
MDS	24.01	11.92	18.37	24.39	11.34	17.76	23.40	8.95	12.88
CSDS	35.67	17.56	27.06	37.10	17.08	33.29	42.06	19.81	33.97
AMI	77.73	58.87	74.26	78.09	62.55	73.26	77.11	51.65	72.15
SumScreen	75.54	46.89	73.54	76.22	50.09	73.22	79.19	49.03	72.87

Table 6: The ROUGE scores of three extractive summarization models, TextRank, BertSum, and CentroidSum.

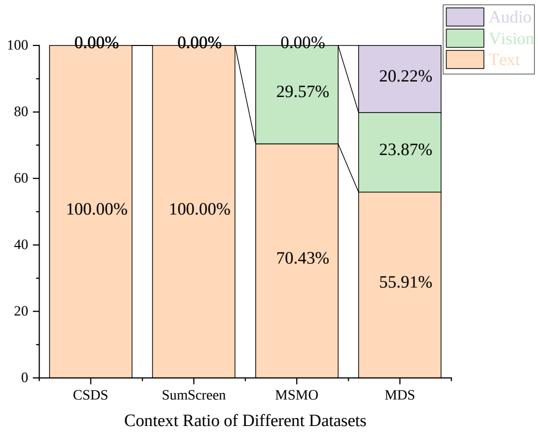

Figure 3: Comparison of context sources for responses in different dialogue datasets.

nouns from three summarization datasets. Annotators are asked to determine whether these words and whole-sentence responses refer to 1) audio context, 2) vision context, 3) text context, and 4) others. We randomly pick 100 samples from each dataset. MDS contains a combination of text (55.91%), vision (23.87%), and audio (20.22%). The lowest text ratio (**55.91%**) indicates that MDS is a multi-modal dataset, meaning it includes data from multiple sources or modalities, while CSDS and SumScreen are entirely composed of text data (100% in the Text column). MSMO, while also offering multi-modal data, emphasizes the vision modality more strongly, in contrast to the more balanced distribution of MDS.

The outcome of the experiments confirms our initial objective of presenting MDS, and information from other modalities supplements the textual content. Relatively high levels of novel n-grams, the lowest ROUGE scores of extractive models, and the lowest text ratio also prove it. These findings emphasize the complex and intertwined nature of our dataset, highlighting the improvement of considering various modalities in our dataset. MDS promises to contribute to being a challenging testbed for multi-modal dialogue summarization

5. Experiment

5.1. Experiment Setup

We compare MDS in three categories of baselines: text summarization, dialogue summarization, and multi-modal summarization, a total of eight models. **S2S** (Luong et al., 2015) is a standard text summarization model with sequence-to-sequence architecture using an RNN encoder-decoder and a global attention mechanism. **PGN** (See et al., 2017) is a text summarization model with an attention mechanism and pointer network. **Transformer** (Vaswani et al., 2017) is a classic text summarization model, which is a non-pre-trained baseline. **T5** (Raffel et al., 2020) is a universal pre-trained abstractive text summarization model on dozens of languages. **MDialBART** (Wang et al., 2022) presents a pre-trained dialogue summarization model. **ConDigSum** (Liu et al., 2021) proposes a dialogue summarization model of topic-aware contrastive learning. **HOW2** (Palaskar et al., 2019) is the first multi-modal summarization model proposed to summarize the video content. **VMSMO** (Li et al., 2020) proposes a dual-interaction multi-modal summarizer to generate multi-modal output. ROUGE-based methods (Lin, 2004) and BLEU-based methods (Papineni et al., 2002) are widely used metrics by measuring the overlap of n-grams between two texts. Here we choose R-1, R-2, R-L, B-1, B-2, B-3, and B-4 for comparison.

5.2. Results and Discussion

Table 7 presents the experimental results. HOW2 achieves the best R-1 (15.94) and R-L (14.50), while R-2 (2.07), B-1 (49.76), B-2 (37.20), B-3 (27.80), and B-4 (21.08) are achieved by T5. The dialogue summarization models cannot perform excellently in ROUGE scores more than other baselines. This may be because traditional dialogue summarization models only focus on specific domains, such as interviews or media, and cannot deal with datasets such as MDS, which contain data from multi-domains. Furthermore, compared to traditional textual models, HOW2 outperforms T5, a pre-trained model, in R-1 and R-L. Multi-modal models are able to capture multi-source informa-

	R-1	R-2	R-L	B-1	B-2	B-3	B-4
Traditional Textual Model							
S2S (Luong et al., 2015)	4.75	0.02	4.59	9.50	5.65	2.88	1.86
PGN (See et al., 2017)	13.26	1.12	11.83	18.58	13.28	10.02	7.81
Transformer (Vaswani et al., 2017)	14.26	1.48	13.16	31.44	22.48	15.98	11.31
T5 (Raffel et al., 2020)	13.59	**2.07**	11.99	**49.76**	**37.20**	**27.80**	**21.08**
Dialogue Summarization Model							
MDialBART (Wang et al., 2022)	9.94	0.56	8.83	23.13	17.71	12.66	8.74
ConDigSum (Liu et al., 2021)	9.62	0.75	9.18	19.76	15.08	11.40	8.56
Multi-Modal Summarization Model							
HOW2 (Palaskar et al., 2019)	**15.94**	1.93	**14.50**	19.58	14.80	11.78	9.52
VMSMO (Li et al., 2020)	11.67	1.53	11.26	15.75	10.75	8.26	6.58

Table 7: ROUGE score and BLEU scores of summarization baselines on MDS.

	R-1	R-2	R-L	B-1	B-2	B-3	B-4
HOW2	**15.94**	**1.93**	**14.50**	**19.58**	14.80	**11.78**	**9.52**
w/o vision	13.73	1.74	12.61	17.64	**15.01**	11.45	8.90
VMSMO	**11.67**	**1.53**	**11.26**	**15.75**	10.75	**8.26**	**6.58**
w/o vision	9.26	0.49	8.92	13.45	**10.88**	7.85	5.91

Table 8: Ablation study of multi-modal summarization model.

tion, and more comprehensive and rich summary content can be obtained. Instead of focusing on the content of the text, visual features can also be incorporated to generate more accurate and precise summaries. Multi-modal summarization models can take advantage of the complementarity and interaction between different data sources to improve the quality of summaries. In contrast, text-only summarization models may not capture the detailed information of the image and thus may be limited in the summarization quality. The result validates the effectiveness of multi-modal information. However, T5 performs more excellently in BLEU scores than conventional multi-modal models. It is possible for conventional multi-modal models to focus only on keyframes or part of the video clips while ignoring other important information. It may result in incomplete or inaccurate summary segments generated. The experiments indicate that existing models cannot handle the multi-modal dialogue summarization task, and MDS is a challenging testbed for it.

5.3. Ablation Study

An ablation study was conducted to show that MDS is a challenging testbed. The most apparent observation from the results is that both How2 and VMSMO, which incorporate visual information, outperform their counterparts that do not use visual information (How2 w/o vision and VMSMO w/o vision) across almost all evaluation metrics. Only the B-2 score doesn't see substantial improvements. Specifically, for How2, the inclusion of visual information results in substantial improvements. The R-1 score significantly increases from 13.73 (How2 w/o vision) to 15.94 (How2), indicating better alignment with the reference summaries. In the case of VMSMO, the impact of incorporating visual data is striking. The Rouge-1, Rouge-2, and Rouge-L scores all see substantial improvements, underlining the positive influence of visual information on content alignment and coherence.

The findings underline the positive impact of including visual data on the quality of generated dialogue summaries, emphasizing the cooperativity between text and visual modalities in enhancing the overall performance of dialogue summarization.

6. Conclusion and Future Work

This paper proposes a fine-grained bilingual dataset, MDS, for multi-modal dialogue summarization. We introduce a multi-modal dialogue summarization dataset that facilitates deeper understanding and improved analysis in multi-modal dialogue summarization. According to the experiments on MDS, multi-modal dialogue summarization is a unique and challenging task. To build up the dataset and solve problems existing before, we propose an annotation framework to produce a summary for multi-modal dialogue, including a video scene-cutting model. In general, we complement the gap in which current dialogue summarization research mainly focuses on textual utterance and ignores the multi-modal content.

Factual inconsistencies are still the central problem in dialogue summarization. In the future, we are devoted to solving the problem from the following perspectives: (1) Utilizing multi-modal information to constrain the generation of the summary. (2) Applying contrastive learning to multi-modal learning. (3) Extending evaluation methods of factual inconsistencies through the dialogue system.

7. Ethical Discussion

Data collection and privacy. MDS is a dataset of links obtained from Common Crawl that gathers content from TV episodes and publicly available Internet. It should be noted that the dataset may contain links to videos with personal information, such as photos of faces, location information, or other personal-related content. In addition, we offer a contact form on our website to facilitate the processing of requests for the removal or blacklisting of corresponding links from MDS in cases where problematic personal or copyrighted content is present. Bias against people of a specific gender or race. The series and interviews certainly perpetuate these antiquated beliefs about our society. Stereotypical depictions of both genders are a significant component of the sitcom. For instance, the character of the heroine is portrayed as a stereotypical "dumb blonde", a woman whose character features are at the forefront of both narrative and comedy. In the whole series, there is only one character of color, Raj, compared to the over five white actors and actresses. Aggressive and offensive content. Sitcoms can serve as a highly effective tool for addressing current issues in a non-threatening and approachable manner, facilitating productive dialogue and identification of concerns.

8. Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments. This work was supported in part by the National Natural Science Foundation of China (No. 62272025 and No. U22B2021), and in part by Fund of the State Key Laboratory of Software Development Environment.

9. Bibliographical References

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Ozan Caglayan, Pranava Swaroop Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2006. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 459–466.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. Summscreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. 2019. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2090–2096. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

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*, pages 3755–3763.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. Vmsmo: Learning to generate multimodal summary for video-based news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369.

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

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Csds: A fine-grained chinese dataset for customer service dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4436–4451.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Valentin Malykh, Konstantin Chernis, Ekaterina Artemova, and Irina Piontkovskaya. 2020. Sumtitles: a summarization dataset with low extractiveness. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5718–5730.

Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle G Lee, Anish Acharya, and Rajiv Shah. 2021. Gupshup: Summarizing open-domain code-switched conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6177–6192.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Shruti Palaskar, Jindřich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21, Valencia, Spain. Association for Computational Linguistics.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *arXiv preprint arXiv:2112.08713*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. Clidsum: A benchmark dataset for cross-lingual dialogue summarization.

Xuefeng Xi, Zhou Pi, and Guodong Zhou. 2020. Global encoding for long chinese text summarization. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(6):1–17.

Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. 2016. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520.

Litian Zhang, Xiaoming Zhang, Ziming Guo, and Zhipeng Liu. 2023. Cisum: Learning cross-modality interaction to enhance multimodal semantic coverage for multimodal summarization.

In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 370–378. SIAM.

Litian Zhang, Xiaoming Zhang, Linfeng Han, Zelong Yu, Yun Liu, and Zhoujun Li. 2024. Multi-task hierarchical heterogeneous fusion framework for multimodal summarization. *Information Processing Management*, 61(4):103693.

Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11676–11684.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. 2021a. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712.

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021b. Emailsum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6895–6909.

Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. Todsum: Task-oriented dialogue summarization with state tracking. *arXiv preprint arXiv:2110.12680*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934.

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*, pages 4154–4164.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14665–14673.

10. Language Resource References

Carletta, Jean and Ashby, Simone and Bourban, Sebastien and Flynn, Mike and Guillemot, Mael and Hain, Thomas and Kadlec, Jaroslav and Karaiskos, Vasilis and Kraaij, Wessel and Kronenthal, Melissa and others. 2006. *The AMI meeting corpus: A pre-announcement*. Springer.

Chen, Meng and Liu, Ruixue and Shen, Lei and Yuan, Shaozu and Zhou, Jingyan and Wu, Youzheng and He, Xiaodong and Zhou, Bowen. 2020. *The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service*.

Chen, Mingda and Chu, Zewei and Wiseman, Sam and Gimpel, Kevin. 2022. *SummScreen: A Dataset for Abstractive Screenplay Summarization*.

Chen, Yulong and Liu, Yang and Chen, Liang and Zhang, Yue. 2021. *DialogSum: A Real-Life Scenario Dialogue Summarization Dataset*.

Gliwa, Bogdan and Mochol, Iwona and Biesek, Maciej and Wawer, Aleksander. 2019. *SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization*.

Janin, Adam and Baron, Don and Edwards, Jane and Ellis, Dan and Gelbart, David and Morgan, Nelson and Peskin, Barbara and Pfau, Thilo and Shriberg, Elizabeth and Stolcke, Andreas and others. 2003. *The ICSI meeting corpus*. IEEE.

Joshi, Anirudh and Katariya, Namit and Amatriain, Xavier and Kannan, Anitha. 2020. *Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures.*

Lin, Haitao and Ma, Liqun and Zhu, Junnan and Xiang, Lu and Zhou, Yu and Zhang, Jiajun and Zong, Chengqing. 2021. *CSDS: A Fine-Grained*

Chinese Dataset for Customer Service Dialogue Summarization.

Malykh, Valentin and Chernis, Konstantin and Artemova, Ekaterina and Piontkovskaya, Irina. 2020. *SumTitles: a summarization dataset with low extractiveness*.

Mehnaz, Laiba and Mahata, Debanjan and Gosangi, Rakesh and Gunturi, Uma Sushmitha and Jain, Riya and Gupta, Gauri and Kumar, Amardeep and Lee, Isabelle G and Acharya, Anish and Shah, Rajiv. 2021. *GupShup: Summarizing open-domain code-switched conversations*.

Song, Yan and Tian, Yuanhe and Wang, Nan and Xia, Fei. 2020. *Summarizing medical conversations via identifying important utterances*.

Zhang, Litian and Zhang, Xiaoming and Pan, Junshu. 2022. *Hierarchical cross-modality semantic correlation learning model for multimodal summarization*.

Zhang, Longxiang and Negrinho, Renato and Ghosh, Arindam and Jagannathan, Vasudevan and Hassanzadeh, Hamid Reza and Schaaf, Thomas and Gormley, Matthew R. 2021. *Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations*.

Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. Todsum: Task-oriented dialogue summarization with state tracking. *arXiv preprint arXiv:2110.12680*.

Zhong, Ming and Yin, Da and Yu, Tao and Zaidi, Ahmad and Mutuma, Mutethia and Jha, Rahul and Hassan, Ahmed and Celikyilmaz, Asli and Liu, Yang and Qiu, Xipeng and others. 2021. *QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization*.

Zhu, Chenguang and Liu, Yang and Mei, Jie and Zeng, Michael. 2021. *MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization*.

Zhu, Junnan and Li, Haoran and Liu, Tianshang and Zhou, Yu and Zhang, Jiajun and Zong, Chengqing. 2018. *MSMO: Multimodal summarization with multimodal output*.

Zou, Yicheng and Zhao, Lujun and Kang, Yangyang and Lin, Jun and Peng, Minlong and Jiang, Zhuoren and Sun, Changlong and Zhang, Qi and Huang, Xuanjing and Liu, Xiaozhong. 2021. *Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling*.

Appendix A. Maintenance and Data Sources

License. MDS is distributed under the Creative Commons (CC) copyright licenses. It is important to note that the source documents used in the dataset are already in the public domain, thereby respecting copyright regulations. We have implemented a contact form on our website to address any concerns related to personal or copyrighted content within MDS. This form serves as a channel for users to submit requests for the removal or blacklisting of specific links or content that may infringe upon personal rights or copyrights. We are committed to promptly and diligently processing these requests to maintain the integrity and legality of the dataset. The authors bear all responsibility in case of violation of rights and confirm the dataset licenses.

Maintenance. The authors are committed to providing long-term support for the MDS dataset. Currently, MDS files are hosted on GitHub, allowing for easy access and collaboration. TikTok videos may be deleted by the publisher. To safeguard against the potential loss of TikTok videos, we have taken proactive measures by uploading all the necessary video content to OneDrive, an online storage platform. This backup ensures the availability and continuity of the dataset, even if the original TikTok videos become unavailable in the future. Additionally, the authors are committed to actively monitoring the usage of the dataset and addressing any issues that may arise. This includes promptly addressing bug fixes, resolving technical concerns, and providing necessary updates to ensure the dataset remains reliable and useful to the research community.

Data Sources. In MDS, our dataset comprises two primary sources: the sitcom "The Big Bang Theory" and TikTok videos. These sources were selected to create a rich multi-modal dialogue dataset with diverse content and unique characteristics. **"The Big Bang Theory" as a Data Source**. "The Big Bang Theory" sitcom serves as a valuable resource for multi-modal dialogue data due to its abundance and availability. Sitcoms, including "The Big Bang Theory," are widely recognized for their scripted nature and well-defined character interactions. The show's popularity and extensive episode collection make it an ideal choice for collecting dialogue data. By utilizing this source, we can tap into the humor, nuanced conversations, and dynamic exchanges that are characteristic of sitcoms. Moreover, the structured scenes within sitcoms provide a natural framework for understanding dialogue flow, facilitating the annotation and analysis process. **TikTok as a Data Source** Complementing the sitcom data, we incorporate TikTok videos as a contemporary and user-generated data source. TikTok has gained immense popularity as a social media platform known for its short-form videos, creative content, and diverse user base. To download videos from TikTok, we use an open-source project, TikTok Download [1]. We introduce a unique aspect of modern communication and expressions by including TikTok in our dataset. These videos capture a wide range of dialogues, encompassing various genres, trends, and cultural references. However, it is important to acknowledge the challenges associated with TikTok data, such as its dynamic nature, shorter video duration, and potential noise. We took careful steps to curate relevant and meaningful TikTok videos, ensuring they align with the objectives of our dataset.

Appendix B. Annotation Process

The annotation platform is built based on an open-source project, Label-Studio [2]. This platform allows annotators to generate summaries for individual dialogues, drawing references from various sources, including video clips, textual transcripts, and tip recaps. Textual transcripts are obtained from subtitle files to annotate the dialogues from sitcoms, ensuring accuracy and alignment with the corresponding scenes. Additionally, tip recaps for the sitcom dialogues are collected from TV drama websites[3], providing a concise summary of the episode or scene under consideration. These tip recaps offer contextual information and aid in capturing the key points and narrative highlights. For TikTok videos, the annotation process involves utilizing different sources. The textual transcripts for TikTok dialogues are obtained from Whisper[4]. These transcripts capture the spoken content within the TikTok videos, enabling a textual representation of the dialogues. Moreover, tip recaps for TikTok videos are derived from the titles accompanying the videos. These titles often provide a brief description or summary of the video content, aiding annotators in understanding the context and essence of the dialogues within the TikTok videos.

By leveraging these diverse sources, including subtitles, TV drama websites, Whisper transcripts, and video titles, the MDS annotation platform ensures that annotators have access to comprehensive references while writing the dialog summaries. This approach allows for a holistic and informed annotation process, promoting the creation of high-quality summaries that capture the essence of the dialogues across both sitcoms and TikTok videos.

Appendix C. Model Training Details

Text Translation. MDS is a bilingual dataset, and the annotations are conducted in Chinese for several reasons. The annotators responsible for generating the annotations are undergraduate students at Beihang University, whose mother tongue is Chinese. Leveraging their linguistic expertise and native fluency in Chinese allows for a meticulous and accurate capturing of the nuances and intricacies of dialogue in the Chinese language. However, recognizing the importance of promoting widespread accessibility and universality, we employ the Google Translate interface to translate the Chinese annotations into English. By leveraging machine translation technology, we aim to facilitate access to the MDS dataset for researchers and practitioners who may not be proficient in the Chinese language. The decision to conduct annotations in Chinese by native speakers and provide English translations through the Google Translate interface reflects our commitment to both capturing the richness of Chinese dialogue and promoting the usability of the dataset for a wider audience. This approach facilitates cross-lingual research, encourages collaboration, and fosters a more inclusive dialogue research community.

Model and Hyperparameter Choice. To carry out our experiments, we utilize the English version of the dataset. This decision enables us to focus on exploring and analyzing the characteristics and performance of the model in an English language context. The experiments are conducted on an NVIDIA Tesla V100 GPU. In the text embedding module of our research, we employ BERT bert-base-uncased as the pre-trained word embedding model. This choice allows us to initialize our embedding matrix, which has a size of 30,522 words, with BERT contextualized representations. The dimensions of the embedding matrix are set to 768, aligning with the output dimensions of BERT. To optimize the model during training, we utilize the Adam optimizer. To establish an effective learning rate schedule, we set the initial learning rate to 1e-3 and implement a decay strategy where the learning rate is multiplied by 0.9 every ten epochs. This approach facilitates stable and gradual learning throughout training, ensuring convergence to an optimal solution.

Appendix D. MDS Datasheet

[1] https://github.com/Evil0ctal/Douyin_TikTok_Download_API

[2] https://github.com/heartexlabs/label-studio

[3] https://www.tvmao.com/drama/

[4] https://github.com/openai/whisper

Dataset Facts

Dataset MDS
Instances Per Dataset 11,305

Composition

Sample or Complete Complete
Missing Data The dataset is entirely self-contained.

Collection

Ethical Review Bias against people of a specific gender or race in the sitcom "The Big Bang Theory". The series and interviews certainly perpetuate these antiquated beliefs about our society. Stereotypical depictions of both genders are a significant component of the sitcom.
Author Consent There is no confidential information in our dataset; all the source documents can be found on the Internet

Cleaning and Labeling

Cleaning Done Yes. We detail data cleaning in Section 3.4 of the paper
Labeling Done Yes. We detail summary writing guidelines in Section 3.2.

Uses and Distribution

Notable Uses MDS is a challenging testbed for multi-modal dialogue summarization.
Other Uses Probably None

Maintenance and Evolution

Corrections or Erratum The authors are committed to actively monitoring the usage of the dataset and addressing any issues that may arise.
Methods to Extend Maybe adding more data.

Breakdown	0% of Example*
Short 9,759 items	86.3%
Medium 1,150 items	10.2%
Long 396 items	3.5%

Figure 4: We develop the dataset sheet based on the template from Gebru et al.