MISP-Meeting: A Real-World Dataset with Multimodal Cues for Long-form Meeting Transcription and Summarization

Hang Chen[1], **Chao-Han Huck Yang**[2], **Jia-Chen Gu**[3],
Sabato Marco Siniscalchi[4], **Jun Du**[1,5*]

[1]NERC-SLIP, University of Science and Technology of China, [2]NVIDIA Research,
[3]University of California, Los Angeles, [4]University of Palermo,
[5]MoE Key Lab of BIPC, University of Science and Technology of China
[*]**corresponding authors:** jundu@ustc.edu.cn

Abstract

We introduce MISP-Meeting, a new real-world, multimodal dataset that covers subject-oriented long-form content. MISP-Meeting integrates information from speech, vision, and text modalities to facilitate automatic meeting transcription and summarization (AMTS). Challenging conditions in human meetings, including far-field speech recognition, audio-visual understanding, and long-term summarization, have been carefully evaluated. We benchmark state-of-the-art automatic speech recognition (ASR) and large language models (LLMs) on this dataset, enhanced with multimodal cues. Experiments demonstrate that incorporating multimodal cues, such as lip movements and visual focus of attention, significantly enhances transcription accuracy, reducing the character error rate (CER) from 36.60% to 20.27% via guided source separation (GSS), fine-tuning, and audio-visual fusion. Furthermore, our summarization analysis reveals a direct correlation between ASR quality and summary coherence, underscoring the importance of robust multimodal modeling. Our dataset and codebase have been released as open source.[1]

1 Introduction

Meetings dominate professional and academic spheres as a cornerstone of information exchange, with millions held globally daily, consuming substantial time and organizational resources (Mroz et al., 2018). (Rogelberg et al., 2007) reported U.S. employees and managers dedicate 6 and 23 weekly hours to meetings, respectively. After the COVID-19 pandemic, the widespread adoption of videoconferencing has led to more prolonged and more frequent meetings (Kost, 2020), resulting in increased fatigue and less time to digest the information exchanged (Fauville et al., 2021). In this context, there is a growing demand for automatic meeting transcription and summarization (AMTS) systems (Gu et al., 2021, 2022, 2023), as shown in Figure 1, capable of recognizing spoken content, extracting key information, and producing concise summaries (Song et al., 2021; Elciyar, 2021).

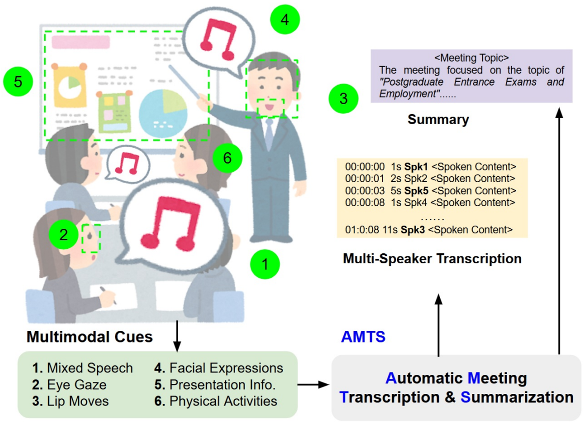

5
4
3
6
2
1
<Meeting Topic>
The meeting focused on the topic of "Postgraduate Entrance Exams and Employment"......
Summary
00:00:00 1s Spk1 <Spoken Content>
00:00:01 2s Spk2 <Spoken Content>
00:00:03 5s Spk5 <Spoken Content>
00:00:08 1s Spk4 <Spoken Content>
......
01:0:08 11s Spk3 <Spoken Content>
Multi-Speaker Transcription
Multimodal Cues
1. Mixed Speech
2. Eye Gaze
3. Lip Moves
4. Facial Expressions
5. Presentation Info.
6. Physical Activities
AMTS
Automatic Meeting Transcription & Summarization

Figure 1: A schematic overview of multimodal cues in the meeting and the automatic meeting transcription and summarization systems.

Recent research has categorized AMTS into two sequenceful sub-tasks: automatic meeting transcription (AMT) (Yoshioka et al., 2019; Von Neumann et al., 2024) and automatic meeting summarization (AMS) (Tan et al., 2023). AMT is dedicated to capturing "who said what and when" from lengthy and unstructured audio recordings of meetings (Raj et al., 2021). Then, AMS summarizes key insights in the transcribed text into well-structured and concise sentences (See et al., 2017; Lewis et al., 2020). Over the past two decades, the rapid development of deep neural networks (DNNs) (Liu et al., 2018; Zhang et al., 2020; Sklyar et al., 2022; Kanda et al., 2022) and the availability of large-scale datasets (Yu et al., 2022; Fu et al., 2021) have significantly improved the performance of AMT and AMS systems. However, these state-of-the-art (SOTA) audio-only technologies still encounter challenges in real-world scenarios. For example, the best recognition performances in the AliMeet-

[1] https://github.com/coalboss/MISP-Meeting

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15479–15492
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ing dataset (Ye et al., 2022) achieved a character error rate (CER) of approximately 20%. The generated meeting transcription is filled with noisy contexts, hindering the effective capture of relevant details for summarization (Rennard et al., 2023).

In addition to audio, various multimodal cues are present in meetings. Figure 1 illustrates examples such as eye gaze, lip movements, facial expressions, presentation materials, and physical activities, all of which play a crucial role in effective communication and understanding. The McGurk effect (McGurk and MacDonald, 1976) suggests a strong influence of visual cues on human auditory perception. Follow-up studies (Rosenblum, 2008; Massaro and Simpson, 2014) have shown that visual cues, such as lip movements, can aid speech perception, particularly in noisy environments. Recent studies (Chen et al., 2024; Dai et al., 2024; Hong et al., 2023) have also demonstrated that incorporating the visual modality can substantially improve recognition accuracy. Furthermore, the participant's head orientation and eye gaze provide the visual focus of attention (Li et al., 2019), which helps identify salient utterances. Specific visual motion activities can enhance the detection of summary-worthy events (Erol et al., 2003). Further support can be found in (Xie and Liu, 2010; Nihei et al., 2016, 2018; Nihei and Nakano, 2019). These insights provide a strong foundation for exploring multimodal meeting transcription and summarization (MMTS) (Renals et al., 2008).

In this paper, we present the multimodal information-based speech processing in meetings (MISP-Meeting) dataset to advance MMTS research. Specifically, MISP-Meeting records and annotates 163 real Mandarin meetings, yielding 125.15 hours of multimodal data and labels: **(1) Raw audio-visual recordings**, including near-field mono speech for each speaker, far-field 8-channel audio, and 360-degree panoramic video, and **(2) Manual annotations,** including professionally generated sentence-level text transcriptions and two types of summaries (brief and detailed versions). Notably, the panoramic camera not only captures each participant's facial expressions and body movements but also records the entire panorama of the meeting room, from which various multimodal cues can be extracted. Furthermore, we benchmark SOTA automatic speech recognition (ASR) and large language models (LLMs) on MISP-Meeting and explore improvements with multimodal cues, such as lip movements. Experiments show that while the best ASR and LLM models still have significant room for improvement, multimodal cues significantly enhance transcription accuracy and summary coherence. In summary, our contributions are as follows:

1. A real-world dataset with multimodal cues towards meeting scenarios, namely MISP-Meeting. To our best knowledge, MISP-Meeting is the first Mandarin multimodal meeting corpus and comprises the largest collection of audio-visual-text data pairs related to meetings.

2. Benchmarking and improving SOTA models. We conduct extensive experiments on MISP-Meeting using SOTA ASR and LLMs, enhanced with multimodal cues, demonstrating the challenging nature and the potential for improvement.

3. Significant appeal and broad applications. Over 60 applications have sought access to MISP-Meeting for various research purposes, including not only MMTS but also audio-visual speaker diarization and speech enhancement, lipreading, object detection and tracking in panoramic video, etc.

2 Related Work

Producing meeting corpora and their associated summaries requires significant resources and raises privacy concerns, resulting in a scarcity of datasets for MMTS. The AMI dataset (Renals et al., 2008) includes 137 meetings with 100 hours of audio-visual recordings, text transcriptions, and partial summary labels, but it focuses mainly on industrial product design and was recorded in just three rooms. Similarly, the ICSI meeting corpus (Janin et al., 2003) consists of 75 academic meetings of research discussions at ICSI in Berkeley, totaling 72 hours of audio recordings and text labels, but lacks video recordings and is confined to 1 meeting room. Additionally, several text-only meeting datasets have also been developed, including the ELITR minuting dataset (Nedoluzhko et al., 2022), which features an impressive 179 project meetings, with 120 in English and 59 in Czech. The AMC dataset (Zhang et al., 2023) consists of 654 Mandarin meetings spanning various topics. Furthermore, the QMSum dataset (Zhong et al., 2021) has enhanced the AMI and ICSI meetings by re-annotating query-based summarization labels and incorporating 36 parliament committee meetings. However, all these datasets share a critical limitation: the absence of audio-visual recordings, which

Dataset	**Lang.**	**Meetings**	**Avg. Len. of Meet.**	**Avg. Len. of Sum.**	**Avg. Turns**	**Avg. Spks.**
ELITR (Czech)	ces	59	8534	373	1205	7.6
ELITR (English)	eng	120	7066	236	727	5.9
ICSI	eng	59	8567.7	488.5	819	6.3
AMI	eng	137	5570.4	296	535.6	4
AMC	cmn	654	10772.5	250	376.3	2.5
MISP-Meeting	cmn	163	**12680.65**	272/**1102**	445.43	5.55

Table 1: Comparison of Statistics among the ELITR, ICSI, AMI, AMC, and MISP-Meeting datasets. **Avg. Len. of Meet.** and **Avg. Len. of Sum.** represent the average character/word count of transcripts and summaries per meeting, respectively. **Avg. Turns**, and **Avg. Spks.** denote the average dialogue turns and speakers per meeting, respectively. **Lang.** refers to the dataset language. ces, eng, and cmn represent Czech, English, and Mandarin, respectively.

restricts the exploration of multimodal cues.

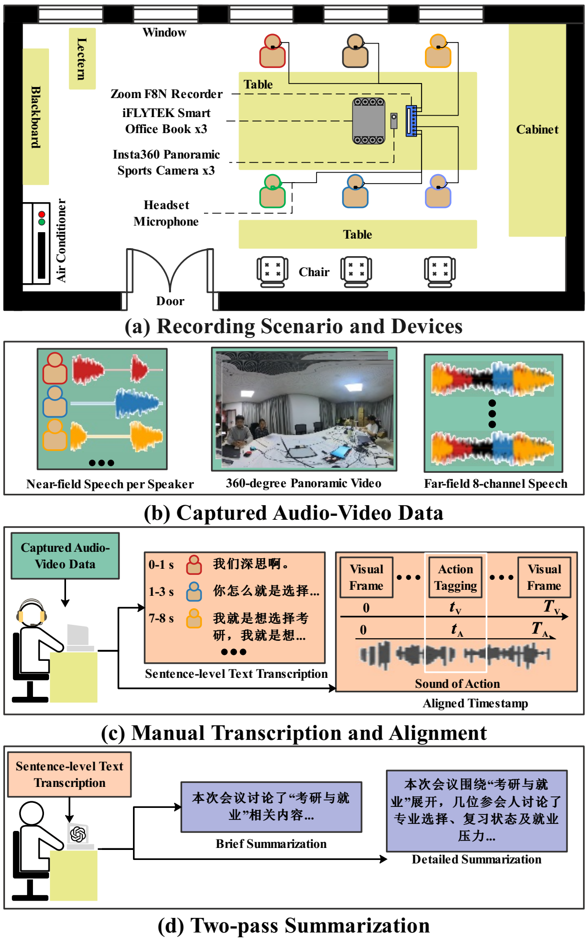
Table 1 presents a comparison of statistical information across various datasets and MISP-Meeting, highlighting crucial factors such as the languages , meeting count, the average character/word counts of transcripts and summaries, average dialogue turns, and average speakers per meeting. The standout features of MISP-Meeting is its exceptionally long meeting transcripts, which are nearly 20% longer than those of the second-longest dataset, AMC. This impressive length can be attributed to more participants and longer meeting durations, both of which contribute significantly to the complexity and richness of information captured in the long-form meetings. Accordingly, we have introduced a new detailed summarization track alongside the traditional brief summarization track, which requires in-depth summaries that highlight overarching insights and delve into intricate local details within the meeting recordings. Additionally, MISP-Meeting continuously performs above average across various statistical metrics.

Most publicly available meeting corpora are limited in scope and often lack summary annotations. Take the CHIL dataset (Mostefa et al., 2007) for example, and it includes just 20 English meetings with 80 speakers, totaling 72 hours of recorded content. Furthermore, some audio-only datasets like AliMeeting (Yu et al., 2022) and Aishell-4 (Fu et al., 2021) feature 500 and 60 Mandarin meetings respectively. Additionally, the simulated audio-only dataset LibriCSS (Chen et al., 2020) attempts to capture meeting dynamics by replaying utterances from the LibriSpeech (Panayotov et al., 2015) dataset through multiple high-fidelity loudspeakers in a meeting room. However, the dialogues produced in this setup lack the necessary continuity that characterizes genuine conversations.

Dataset	**Mod.**	**Lang.**	**Dur. (h)**	**Room**	**Spks.**
AliMeeting	A	cmn	118.75	13	500
Aishell-4	A	cmn	120	10	60
LibriCSS	A	eng	10	1	\
ICSI	A	eng	72	1	53
AMI	AV	eng	100	3	189
CHIL	AV	eng	20	5	80
MISP-Meeting	AV	**cmn**	**125.15**	**23**	274

Table 2: Comparison of statistical information among the AliMeeting, Aishell-4, LibriCSS, ICSI, AMI, CHIL, and MISP-Meeting datasets. **Mod.**, **Dur.**, **Room** and **Spks.** represent the modality, duration, meeting room count, and speaker count, respectively. A and AV denote audio-only and audio-visual, respectively.

Table 2 compares statistical information across these datasets and MISP-Meeting, with a focus on modality, language, duration, meeting room diversity, and speaker count. The MISP-Meeting dataset exhibits significant advantages across various dimensions. As the first Mandarin multimodal meeting corpus, it also features the largest collection of audio-visual recordings of natural meetings. The 125.15 hours of duration surpasses the second-largest AMI dataset by nearly 25%. Even among audio-only datasets, MISP-Meeting exceeds Aishell-4 by an additional 5 hours, firmly establishing itself as the largest meeting corpus available. Another striking aspect is the environment diversity, encompassing 23 distinct meeting rooms. This figure vastly outshines other audio-visual and audio-only datasets, such as AliMeeting (13 rooms), Aishell-4 (10 rooms) and CHIL (5 rooms). These meeting rooms provide a wide range of acoustic and visual environments, significantly enhancing the generalizability of models trained on the dataset. MISP-Meeting also stands out with 274 speakers, the highest among multimodal datasets


Window
Lectern
Blackboard
Table
Zoom F8N Recorder
iFLYTEK Smart Office Book x3
Insta360 Panoramic Sports Camera x3
Cabinet
Headset Microphone
Table
Air Conditioner
Chair
Door
Near-field Speech per Speaker
360-degree Panoramic Video
Far-field 8-channel Speech
Captured Audio-Video Data
0-1 s 我们深思啊。
1-3 s 你怎么就是选择...
7-8 s 我就是想选择考研，我就是想...
Sentence-level Text Transcription
Visual Frame
Action Tagging
Visual Frame
0
t_V
T_V
0
t_A
T_A
Sound of Action
Aligned Timestamp
Sentence-level Text Transcription
本次会议讨论了"考研与就业"相关内容...
Brief Summarization
本次会议围绕"考研与就业"展开，几位参会人讨论了专业选择、复习状态及就业压力...
Detailed Summarization

(a) Recording Scenario and Devices

(b) Captured Audio-Video Data

(c) Manual Transcription and Alignment

(d) Two-pass Summarization

Figure 2: A schematic overview of the misp-meeting data collection and processing, including (a) the recording scenario and devices, (b) the captured audio-video data, (c) the manual transcription and alignment, and (d) the two-pass summarization.

and 45% more than AMI features 189 speakers. These unique advantages position MISP-Meeting as an invaluable resource in the MMTS field.

3 MISP-Meeting Dataset

3.1 Recording Scenarios and Devices

As depicted in Figure 2 (a), 4–8 meeting attendees sit around an 8-microphone array and a panoramic camera, both placed adjacent to each other on the table in a standard meeting room, engaging in a natural conversation. Additionally, Each participant wore a headset microphone synchronized with a Zoom F8N recorder to share a common clock. This novel recording setup yields a wealth of audio-visual data, as illustrated in Figure 2 (b), including near-field mono speech for each speaker, far-field 8-channel speech, and 360-degree panoramic video. Significantly, the far-field 8-channel spech not only records each participant's spoken contributions but also captures the rich tapestry of background sounds, such as clicking, keyboard typing, door opening and closing, and fan sounds. In contrast, the near-field mono speech effectively reduces interference from unwanted sources while maintaining a remarkable signal-to-noise ratio (SNR) greater than 15 dB. The panoramic camera captures the entire meeting room, including each participant's facial expressions, body movements, and the visual focus of attention, providing a rich source of multimodal cues for analysis. More details about devices can be found in Appendix A.1.1

Moreover, meticulous attention is given to the metadata of the meeting rooms and attendees, including area, age, occupation, and field of study, which are carefully documented to support future research after anonymization. The topics of the meetings are thoughtfully selected based on the attendees' professional backgrounds and areas of expertise, creating a dynamic array of topics, including medical treatment, education, business, and industrial production. This deliberate strategy enhanced attendee engagement, resulting in a rich and valuable archive of meeting records.

3.2 Manual Annotations

As depicted in Figure 2 (c) and (d), the manual annotation pipeline includes three parts:

1. **Professional Transcription:** Skilled transcribers perform manual transcriptions via auditive and waveform analysis based on near-field speech, marking the start and end points of each sentence and the corresponding spoken content. Dual-stage verification enforces $<$ 100ms temporal precision and $>$ 99% character accuracy.

2. **Audio-Visual Synchronization:** The microphone, camera, and recorder clocks are synchronized through manual calibration using cup-strike reference events. Audio-visual reference timestamps (impact waveform peaks and frame-level contact moments) undergo dual validation, enforcing $<$ 100ms temporal alignment.

3. **Two-pass Summarization:** Structured transcripts (time-speaker-content tuples) are processed through ChatGPT-o1 (OpenAI, 2023) for 2 versions of summaries (brief and detailed), followed by expert editorial refinement. Tripartite consensus validation ensures logical coherence and information completeness across all summaries. See Appendix A.1.2 for more details.

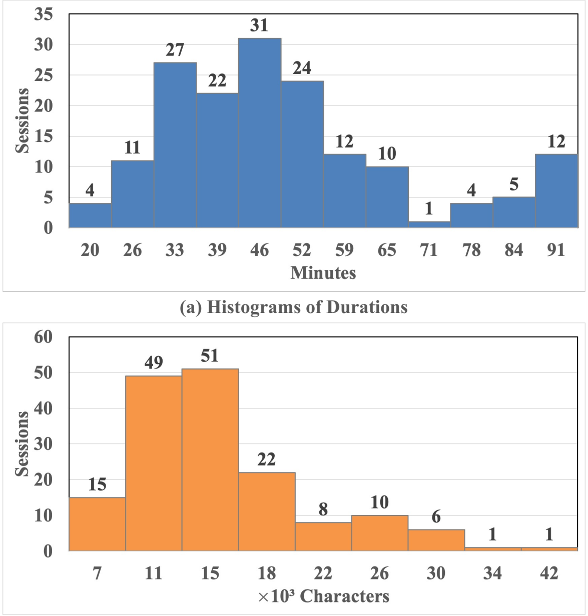
Set	Train	Dev	Eval	Total
Sessions	151	6	6	163
Dur. (h)	118.80	3.24	3.11	125.15
Room	15	4	4	23
Speaker	233	25	28	286
- Male	115	13	14	142
- Female	118	12	14	144
Avg. Dur. (min)	47.21	32.39	31.05	46.07
Avg. Len. (k)	13.09	7.56	7.62	12.68
Avg. Turns	463.33	118.83	321.67	445.43
Avg. Spks.	5.57	5.00	5.50	5.55

Table 3: The overall statistics the MISP-Meeting Dataset. **Avg. Dur.**, **Avg. Len.**, **Avg. Turns** and **Avg. Spks.** represent the average session duration, character count, turns and speakers per session, respectively.

3.3 Statistical Information

The overall statistics of MISP-Meeting are detailed in Table 3. Specifically, MISP-Meeting consists of 163 meetings, divided into 151 for training, 6 for validation, and 6 for evaluation. The training set features extensive and diverse discussions spanning various topics with durations ranging from 8 to 100 minutes, totaling an impressive 125.12 hours of audio-visual data. In contrast, the validation and evaluation sessions are designed to be more focused. Each session is centered around a specific topic and lasts between 20 and 30 minutes, contributing 3.24 hours and 3.11 hours of data, respectively. The average durations per session are 47.21, 32.39, and 31.05 minutes in the training, validation, and evaluation sets, respectively. The distribution histogram of durations is visualized in Figure 3 (a), revealing diversity patterns in the temporal structure of the MISP-Meeting dataset. Specifically, most meetings are between 30 and 40 minutes, though a few meetings last over 70 minutes. This finding highlights that MISP-Meeting preserves the temporal diversity of real meetings.

As for the speakers, the entire dataset comprises 286 speakers, with 233 allocated for training, 25 for validation, and 28 for evaluation, ensuring no overlap. The gender distribution is also well-balanced, with the proportion of male and female speakers being $1 : 1.03$, $1 : 0.92$ and $1 : 1$ in the training, validation and evaluation sets, respectively. Each meeting session includes 4-8 speakers, with an average of 5.57, 5.00 and 5.50 speakers per session in the training, validation and evaluation sets, respectively. The duration of meetings and the number of participants jointly determine the number of dialogue turns and the character count. The average number of dialogue turns per session is 463.33, 118.83, and 321.67 in the training, validation, and evaluation sets, respectively. Correspondingly, the average character count per session is 13.09, 7.56, and 7.62 thousand characters in these sets. Figure 3 (b) presents a histogram of character counts, illustrating that most meetings contain between 10 and 15 thousand characters, with a few sessions exceeding 34 thousand characters. These distributions highlights that MISP-Meeting captures a wide range of meeting complexities, providing a diverse and realistic environment for training and evaluation.



(a) Histograms of Durations

(b) Histograms of Character Counts

Figure 3: Distribution of session durations and character counts in the MISP-Meeting dataset: (a) the histogram of durations and (b) the histogram of character counts.

MISP-Meeting includes 23 meeting rooms, divided into 15 rooms for training, 4 for validation, and 4 for evaluation, covering a range of room sizes from small to large. This environmental diversity is crucial for developing models that generalize effectively to real-world scenarios. More details about the meeting rooms can be found in Appendix A.1.3.

4 Models and Experiments

4.1 Baseline Model

As illustrated in Figure 4, the baseline model is built on a sequential two-component framework: a recognition module followed by a summarization module. This process mirrors the annotation procedure presented in Figure. 2 (c) and (d). Initially, the

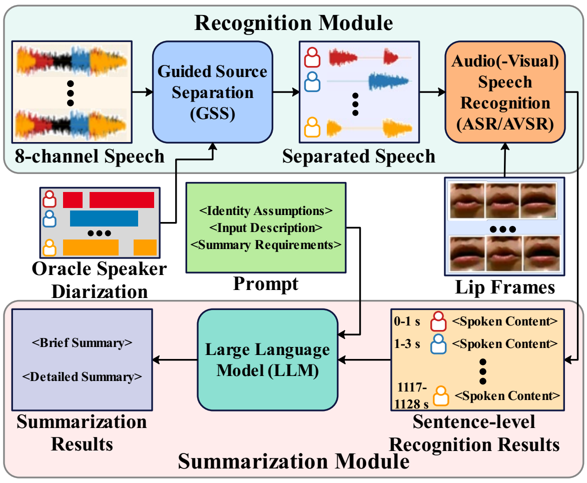


Figure 4: Illustration of the baseline model for MISP-Meeting. The model features two key components: the recognition and summarization modules. We use Whisper models as the foundation for the recognition module, exploring improvements such as GSS, fine-tuning, and AVSR. The summarization module evaluates multiple SOTA LLMs and investigates how recognition performance influences final summaries.

long-duration mixed speech is segmented based on the oracle speaker diarization label. Then, each segment is transcribed using the recognition module. Finally, the transcribed meeting record is passed to the summarization module, where the LLM is guided by the same prompts employed during annotation to produce brief and detailed summaries.

Regarding the recognition module, we first employ Whisper-v2-Large and Whisper-v3-Large models (Radford et al., 2023) to directly recognize far-field speech. Specifically, we extract a single channel from the far-field 8-channel speech to compute an 80-channel log-magnitude Mel spectrogram using 25-millisecond windows with a 10-millisecond stride. Followed by feature normalization, the input spectrogram is globally rescaled to lie between -1 and 1 with approximately 0 mean. The Transformer-based encoder processes this normalized representation and the sinusoidal position embedding via pre-activation residual blocks. The decoder, which has the same Transformer blocks as the encoder, uses learned position embeddings and tied input-output token representations to generate recognized characters autoregressively. Further, we proactively seek to enhance the recognition module through three key strategies:

1. Multi-Channel Speech Enhancement: Guided source separation (GSS) (Raj et al., 2023) is adopted to replace the single-channel extraction, performing dereverberation and source separation on the far-field 8-channel speech to mitigate the mismatch between training and testing caused by complex acoustic environments.

2. Fine-Tuning: We fine-tune the Whisper-v3-Large model with the enhanced speech of the MISP-Meeting training set by freezing the encoder and re-initializing an attention-based decoder. Additionally, a language model is trained on the training transcriptions and employed in decoding with a weight of 0.2. Appendix A.2.1 shows the specific model structure and training details.

3. Audio-Visual Speech Recognition: We also extend the fine-tuned audio-only model with a Transformer-based visual encoder and a cross-modal attention-based audio-visual fusion module, similar to those in (Dai et al., 2024), to leverage the robust nature of the visual modality against complex acoustic environments and more effectively extract the target audio components. More details can be found in Appendix A.2.1.

As for the summarization module, we evaluate the performance of various open-source LLMs, including Qwen 2.5 Max (Qwen, 2024), DeepSeek R1 (DeepSeek-AI, 2025), Moonshot v1 (AI, 2023), Gemini 2.0 Flash (DeepMind, 2024), Llama 3.1 (Meta, 2023), and Llama 3.2 (Meta, 2024). We adopt the same prompts used during annotation (details can be found in Appendix A.1.2) to guide these models in generating both brief and detailed meeting summaries.

4.2 Evaluation Metrics

For recognition performance, CER serves as metric and is calculated as follows:

$$\mathrm{CER} = \frac{\mathrm{N_s} + \mathrm{N_d} + \mathrm{N_i}}{\mathrm{N_c}} \tag{1}$$

where $\mathrm{N_c}$ is the total number of reference characters, and $\mathrm{N_s}$, $\mathrm{N_d}$ and $\mathrm{N_i}$ denote the number of substitution, deletion and insertion errors, respectively. Lower CER values indicate better performance.

Regarding summarization performance, we utilize F-scores of the ROUGE-1, ROUGE-2, and ROUGE-L (denoted as R-1, R-2, and R-L) as metrics. ROUGE measures the quality of generated summaries by comparing them against human-generated references, all R-1, R-2, and R-L scores range from 0 to 1, and higher values indicate better performance. We employs the `rouge_score` package from the NLTK library to compute ROUGE

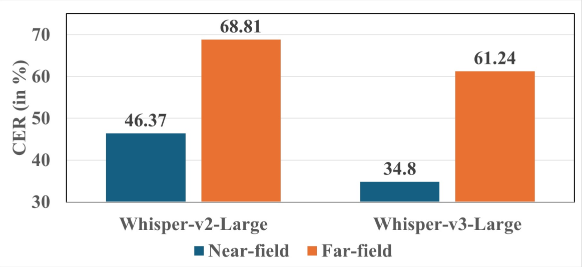


Figure 5: Comparison of average CER between Whisper-v2-Large and Whisper-v3-Large models using near-field mono and far-field 8-channel speech on the MISP-Meeting evaluation set. The far-field result represents the lowest CER among all 8 channels.

scores. Additionally, we remove extra spaces and duplicate punctuation before calculating ROUGE.

4.3 Analysis of Recognition Performance

We begin with evaluating the performance of two SOTA ASR models, namely Whisper-v2-Large and Whisper-v3-Large, on the MISP-Meeting evaluation set. Figure 5 compares the average CER between two models using near-field mono and far-field 8-channel speech. The far-field result is derived from the channel with the lowest CER among the 8 channels. Remarkably, both models encounter substantial performance declines when transcribing far-field speech compared to near-field speech. Specifically, Whisper-v2-Large and Whisper-v3-Large suffer increases in CER of 22.44 and 26.44, respectively. These findings underscore the formidable challenges inherent in transcribing real-world meetings, where adverse acoustic conditions create significant obstacles, such as far-field channel attenuation, pervasive background noise, and reverberation. Additionally, the complexities of multi-speaker interactions often lead to extensive speech overlap, further complicating transcription efforts. Importantly, Whisper-v3-Large demonstrates a clear superiority over its predecessor, achieving impressive reductions in CER of 20.27 for near-field mono speech and 23.17 for far-field 8-channel speech. Accordingly, Whisper-v3-Large has been adopted as the default recognition model in our subsequent experiments.

Next, we systematically investigate the impact of three enhanced strategies on far-field 8-channel speech recognition performance: GSS, fine-tuning, and AVSR. Table 4 presents the average CERs for Whisper-v3-Large models incorporating these strategies on the MISP-Meeting evaluation set.

Our findings reveal substantial error reduction through progressive strategy integration, where GSS achieves a CER of 36.60, demonstrating a 40.64% relative reduction from the baseline. This improvement is uniformly distributed across substitution, deletion, and insertion errors, confirming GSS's effectiveness in mitigating noise, reverberation, and speech overlap through spatial filtering.

Model	**Strategy**			**CER** (%)			
	GSS	**FT**	**AV**	**Sub.**	**Del.**	**Ins.**	**Tot.**
Whisper -Large -v3	×	×	×	20.40	19.75	21.09	61.24
	✓	×	×	13.50	11.63	11.48	36.60
	✓	✓	×	15.04	6.18	**1.95**	23.17
	✓	✓	✓	**13.63**	**4.63**	2.01	**20.27**

Table 4: Comparison of average CER for Whisper-v3-Large models employing various improvement strategies on the far-field 8-channel speech of the MISP-Meeting evaluation set. **FT** and **AV**: Fine-tuning and Audio-visual. **Sub.**, **Del.**, **Ins.** and **Tot.**: Substitution, Deletion, Insertion and Total errors.

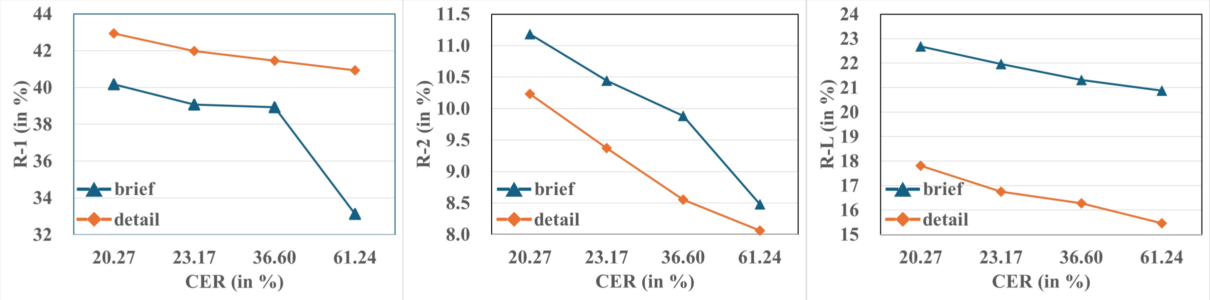
Fine-tuning further reduces CER to 23.17% (13.43% absolute reduction from GSS-only), primarily driven by decreased deletion and insertion errors. However, we observe an unexpected 1.53 increase in substitution errors, attributable to the model's over-adaptation to overlapping speech patterns in meeting scenarios where target/interferer speakers share similar acoustic characteristics.

AVSR delivers the most significant improvement (CER = 20.27%, 12.50% absolute reduction from audio-only fine-tuned), with error reduction concentrated in substitution and deletion categories. This highlights the visual modality's capability to resolve acoustic ambiguities by extracting articulatory features from lip movements, particularly effective in far-field overlapping speech scenarios (see Appendix A.2.2 for visualization examples).

These results establish a clear technological progression: Spatial processing → Acoustic adaptation → Multimodal disambiguation, ultimately achieving 40.97% total CER reduction from baseline. The findings underscore the critical importance of synergistic integration of multi-channel processing (GSS), domain-adaptive fine-tuning, and audio-visual fusion for robust automatic recognition in real-world meeting environments.

4.4 Analysis of Summarization Performance

Finally, we evaluate the summarization performance of Qwen 2.5 Max, DeepSeek R1, Moonshot v1, Gemini 2.0 Flash, Llama 3.1, and Llama 3.2, on the MISP-Meeting. Each model gener-



(a) R-1

(b) R-2

(c) R-L

Figure 6: Line plots of ROUGE scores as a function of CER levels: (a) R-1, (b) R-2, and (c) R-L. The Gemini 2.0 model generates all brief summaries, while all detailed summaries are generated by the Qwen-Max model. The models corresponding to the recognition results are presented in Table 4.

LLM	Brief (%)			Detailed (%)		
	R-1	R-2	R-L	R-1	R-2	R-L
Qwen 2.5 Max	34.1	6.88	16.8	**42.9**	**10.2**	**17.8**
DeepSeek R1	34.6	7.02	20.7	40.2	7.47	16.6
Moonshot v1	35.1	7.79	19.5	36.1	6.96	16.4
Gemini 2.0 Flash	**40.2**	**11.2**	**22.7**	38.6	8.33	16.4
Llama 3.1	23.4	3.05	12.3	30.5	3.57	11.9
Llama 3.2	22.2	2.73	12.2	20.1	2.50	10.9

Table 5: Comparison of average ROUGE scores among Qwen 2.5 Max, DeepSeek R1, Moonshot v1, Gemini 2.0 Flash, Llama 3.1, and Llama 3.2 on the MISP-Meeting evaluation set. All models generate both brief and detailed summaries based on the AVSR recognition results and identical annotation prompts.

ates brief and detailed summaries from AVSR outputs using standardized prompts (same as those in the annotation, details in Appendix A.1.2). As shown in Table 5, Gemini 2.0 Flash achieves superior performance for brief summaries (R-1=40.2, R-2=11.2, R-L=22.7), while Qwen 2.5 Max excels in detailed summaries (R-1=42.9, R-2=10.2, R-L=17.8). Conversely, Llama 3.2 underperforms across both tasks.

Manual analysis reveals distinct error summary patterns when encountering the same recognition errors. Specifically, Gemini 2.0 Flash exhibits conservative summarization, omitting key points, while Llama 3.2 generates hallucinated content unrelated to the source material. Appendix A.1.2 illustrate an example of high-scoring and low-scoring brief summaries from Gemini 2.0 Flash and Llama 3.2, respectively. However, all models significantly trail human performance, a gap strongly correlated with recognition error rates.

Figure 6 reveals inverse correlations between CER levels and ROUGE metrics (R-1/R-2/R-L). The non-linear degradation of summary quality with rising CER confirms the forward-looking error propagation pattern in cascaded systems. Two primary pathways can break this error cascade: error suppression, which involves optimizing the recognition module to minimize CER at the source, and error tolerance, which focuses on enhancing the robustness of LLMs.

5 Conclusion

This paper advances MMTS through three principal contributions. First, we introduce the MISP-Meeting dataset, the first large-scale Mandarin multimodal meeting dataset encompassing 163 real-world meetings covering various topics, 23 meeting rooms, 274 speakers with meta information, sentence-level manual transcription, and two types of summary labels. It not only fully captures multimodal cues using panoramic cameras but also authentically replicates critical challenges such as far-field channel attenuation, reverberation, background noise, and persistent speech overlaps. Second, our benchmark framework integrates GSS, fine-tuning, and cross-modal attention fusion, achieving a 67% CER reduction (from 61.24 to 20.27) against Whisper-v3-Large baselines. Quantitative analysis reveals strong recognition-summary interdependence, where this CER reduction directly correlates with 8.6% and 15% ROUGE-L gains in brief and detailed summarization. Third, the exposed performance gaps expose a critical issue: current SOTA models achieve merely about 40 for ROUGE-1, 10 for ROUGE-2, and 20 for ROUGE-L scores for both brief and detailed summaries. These figures fall drastically short of human performance, underscoring fundamental limitations in existing MMTS architectures for long-context multimodal reasoning.

Limitation

This paper has some limitations, both in terms of data and algorithms:

Data Limitations: MISP-Meting currently includes only Mandarin, limiting non-native researchers from conducting in-depth analyses. To address this, we are implementing two key initiatives: (1) We will provide English translations for transcriptions and summary labels, enhancing accessibility for a broader audience. (2) We will include some English meetings with Chinese participants with an English proficiency certificate (like TEM-8). These updates will increase the dataset's multilingual diversity and improve its relevance for international research.

Algorithmic Limitations: We have not utilized the MISP-Meeting dataset to fine-tune LLMs or develop an end-to-end summarization model. Our future work will explore these two important areas.

Ethical and Societal Considerations

The MISP-Meeting dataset was developed closely with ISO/IEC 27001 (Information Security) and 27701 (Privacy Information Management) certified data partners specializing in multimodal data collection and anonymization. The development partners implement enterprise-grade security and privacy safeguards encompassing encrypted data transmission, standardized de-identification processes, and granular access governance frameworks that satisfy international data protection regulations. All participant consent agreements integrate dynamic revocation mechanisms, allowing retrospective withdrawal until the final dataset publication.

MISP-Meeting is licensed under CC BY-NC-ND 4.0, which allows academic purpose usages and freely available upon authorization. By open-sourcing MISP-Meeting, we are committed to enhancing transparency and fostering collaboration in the research community. To strike an effective balance between accessibility and accountability, we have implemented a robust data usage agreement prohibiting commercial exploitation without our express permission. Furthermore, users must acknowledge and reflect on potential biases, including linguistic and cultural nuances in meeting dynamics.

Long meeting analysis technologies enabled by MISP-Meeting can revolutionize workplace productivity, enhance information communication, and improve accessibility with features such as real-time summaries for those with hearing impairments. However, it's crucial to be aware of the risks of over-reliance on automated systems that may misinterpret nuanced discussions. To address this, we advocate for human oversight in critical situations and encourage researchers to be transparent about their models' limitations. Our collaboration with domain experts during dataset curation ensures diverse representation across various meeting types and speaker demographics, actively reducing systemic biases.

The long-term stewardship of MISP-Meeting will be robustly managed by our dedicated academic team and trusted data partners. We are committed to a long-term governance model incorporating community feedback and ensuring continuous enhancement and relevance. A specialized committee will rigorously address any ethical concerns users raise and proactively update anonymization protocols to maintain the highest data privacy and integrity standards.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62171427. We thank anonymous reviewers for their valuable comments.

References

Moonshot AI. 2023. Moonshot v1. Software. Version 1.0.

Hang Chen, Qing Wang, Jun Du, Bao-Cai Yin, Jia Pan, and Chin-Hui Lee. 2024. Optimizing audio-visual speech enhancement using multi-level distortion measures for audio-visual speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2508–2521.

Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, Jian Wu, Xiong Xiao, and Jinyu Li. 2020. Continuous speech separation: Dataset and analysis. In *Proc. ICASSP 2020*, pages 7284–7288.

Yusheng Dai, Hang Chen, Jun Du, Ruoyu Wang, Shihao Chen, Haotian Wang, and Chin-Hui Lee. 2024. A study of dropout-induced modality bias on robustness to missing video frames for audio-visual speech recognition. In *Proc. CVPR 2024*, pages 27445–27455.

Google DeepMind. 2024. Gemini 2.0 flash: Efficient and scalable ai model. `https://deepmind.g` oogle/technologies/gemini/flash-lite/. Accessed: 2025-02-14.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Kemal Elciyar. 2021. Overloading in lockdown: Effects of social, information and communication overloads in covid-19 days. *İnönü Üniversitesi İletişim Fakültesi Elektronik Dergisi (İNİF E-Dergi)*, 6(1):329–342.

Berna Erol, D-S Lee, and Jonathan Hull. 2003. Multimodal summarization of meeting recordings. In *Proc. ICME 2003*, volume 3, pages III–25.

G. Fauville, M. Luo, A.C.M. Queiroz, J.N. Bailenson, and J. Hancock. 2021. Zoom exhaustion & fatigue scale. *Computers in Human Behavior Reports*, 4:100119.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen. 2021. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. In *Proc. Interspeech 2021*, pages 3665–3669.

Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, Cong Liu, and Guoping Hu. 2023. GIFT: Graph-induced finetuning for multi-party conversation understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11645–11658, Toronto, Canada. Association for Computational Linguistics.

Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022. HeterMPC: A heterogeneous graph neural network for response generation in multi-party conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5086–5097, Dublin, Ireland. Association for Computational Linguistics.

Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. 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 3682–3692, Online. Association for Computational Linguistics.

Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. 2023. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *Proc. CVPR 2023*, pages 18783–18794.

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 *Proc. ICASSP 2003*, volume 1, pages I–I.

Naoyuki Kanda, Jian Wu, Yu Wu, Xiong Xiao, Zhong Meng, Xiaofei Wang, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Takuya Yoshioka. 2022. Streaming multi-talker asr with token-level serialized output training. In *Interspeech 2022*, pages 3774–3778.

Danielle Kost. 2020. You're right! you are working longer and attending more meetings. *Harvard Business School Working Knowledge*.

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 *Proc. ACL 2020*, pages 7871–7880.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proc. ACL 2019*, pages 2190–2196.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *Proc. ICLR 2018*.

D. W. Massaro and J. A. Simpson. 2014. *Speech perception by ear and eye: A paradigm for psychological inquiry*. Psychology.

H. McGurk and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature*, pages 746–748.

Meta. 2023. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Djamel Mostefa, Nicolas Moreau, Khalid Choukri, Gerasimos Potamianos, Stephen M Chu, Ambrish Tyagi, Josep R Casas, Jordi Turmo, Luca Cristoforetti, Francesco Tobia, et al. 2007. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language resources and evaluation*, 41:389–407.

Joseph E. Mroz, Joseph A. Allen, Dana C. Verhoeven, and Marissa L. Shuffler. 2018. Do we really need another meeting? the science of workplace meetings. *Current Directions in Psychological Science*, 27(6):484–491.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182.

Fumio Nihei and Yukiko I Nakano. 2019. Exploring methods for predicting important utterances contributing to meeting summarization. *Multimodal Technologies and Interaction*, 3(3):50.

Fumio Nihei, Yukiko I Nakano, and Yutaka Takase. 2016. Meeting extracts for discussion summarization based on multimodal nonverbal information. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 185–192.

Fumio Nihei, Yukiko I Nakano, and Yutaka Takase. 2018. Fusing verbal and nonverbal information for extractive meeting summarization. In *Proceedings of the Group Interaction Frontiers in Technology*, pages 1–9.

OpenAI. 2023. Gpt-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *Proc. ICASSP 2015*, pages 5206–5210.

Qwen. 2024. Qwen2.5 technical report.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML 2023*, pages 28492–28518.

Desh Raj, Pavel Denisov, Zhuo Chen, Hakan Erdogan, Zili Huang, Maokui He, Shinji Watanabe, Jun Du, Takuya Yoshioka, Yi Luo, et al. 2021. Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. In *Proc. SLT 2021*, pages 897–904.

Desh Raj, Daniel Povey, and Sanjeev Khudanpur. 2023. Gpu-accelerated guided source separation for meeting transcription. In *Proc. Interspeech 2023*, pages 3507–3511.

Steve Renals, Thomas Hain, and Hervé Bourlard. 2008. Interpretation of multiparty meetings the ami and amida projects. In *Proc. HSCMA 2008*, pages 115–118.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023. Abstractive meeting summarization: A survey. *Transactions of the Association for Computational Linguistics*, 11:861–884.

Steven G Rogelberg, Cliff Scott, and John Kello. 2007. The science and fiction of meetings. *MIT Sloan management review*, 48(2):18–21.

L. D. Rosenblum. 2008. Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, pages 405–409.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. ACL 2017*, volume 1, pages 1073–1083.

Ilya Sklyar, Anna Piunova, Xianrui Zheng, and Yulan Liu. 2022. Multi-turn rnn-t for streaming recognition of multi-party speech. In *Proc. ICASSP 2022*, pages 8402–8406.

Yuanfeng Song, Di Jiang, Xuefang Zhao, Xiaoling Huang, Qian Xu, Raymond Chi-Wing Wong, and Qiang Yang. 2021. Smartmeeting: Automatic meeting transcription and summarization for in-person conversations. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2777–2779.

Haochen Tan, Han Wu, Wei Shao, Xinyun Zhang, Mingjie Zhan, Zhaohui Hou, Ding Liang, and Linqi Song. 2023. Reconstruct before summarize: An efficient two-step framework for condensing and summarizing meeting transcripts. In *Proc. EMNLP 2023)*, pages 13128–13141.

Thilo Von Neumann, Christoph Boeddeker, Tobias Cord-Landwehr, Marc Delcroix, and Reinhold Haeb-Umbach. 2024. Meeting recognition with continuous speech separation and transcription-supported diarization. In *Proc. ICASSPW 2024*, pages 775–779. IEEE.

Shasha Xie and Yang Liu. 2010. Using confusion networks for speech summarization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–54.

Shuaishuai Ye, Peiyao Wang, Shunfei Chen, Xinhui Hu, and Xinkang Xu. 2022. The royalflush system of speech recognition for m2met challenge. In *Proc. ICASSP 2022*, pages 9181–9185.

Takuya Yoshioka, Igor Abramovski, Cem Aksoylar, Zhuo Chen, Moshe David, Dimitrios Dimitriadis, Yifan Gong, Ilya Gurvich, Xuedong Huang, Yan Huang, Aviv Hurvitz, Li Jiang, Sharon Koubi, Eyal Krupka, Ido Leichter, Changliang Liu, Partha Parthasarathy, Alon Vinnikov, Lingfeng Wu, Xiong Xiao, Wayne Xiong, Huaming Wang, Zhenghao Wang, Jun Zhang, Yong Zhao, and Tianyan Zhou. 2019. Advances in online audio-visual meeting transcription. In *Proc. ASRU 2019*, pages 276–283.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, et al. 2022. M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge. In *Proc. ICASSP 2022*, pages 6167–6171.

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

Qinglin Zhang, Chong Deng, Jiaqing Liu, Hai Yu, Qian Chen, Wen Wang, Zhijie Yan, Jinglin Liu, Yi Ren, and Zhou Zhao. 2023. Mug: A general meeting understanding and generation benchmark. In *Proc. ICASSP 2023*, pages 1–5.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proc. NAACL 2021*, pages 5905–5921.

A Appendix

A.1 Data Construction

A.1.1 Recording Devices

The microphone array is integrated into the iFLYTEK Smart Office Book X3, configured in a rectangular topology with dimensions of 197mm in length and 134mm in width. This array comprises 8 omnidirectional microphones symmetrically distributed along the two width edges. Each microphone captures audio at a sampling rate of 16kHz and a resolution of 32bits. An Insta360 Panoramic Sports Camera X3 is positioned upright adjacent to the microphone array. This rectangular camera measures 46mm in width and 114mm in length, with two fisheye lenses at the top of the front and back surfaces and 2 omnidirectional microphones on the sides. The setup is mounted on a stand, positioned 30-40cm above the table surface. The outputs include MP4 files with 360-degree panoramic video at 3840 × 1920 pixels and 30fps, accompanied by 2-channel audio recorded at 48kHz and 16-bit. The headset microphone collected near-field speech at 44.1kHz and 16-bit resolution.

A.1.2 Summarization Details

The prompt starts with an identity assumptions statement that directs the model to take on the role of a senior secretary, ensuring a professional tone throughout. Then, the prompt outlines a specific format for the meeting transcript, where each line includes start and end timestamps, a speaker identifier, and the corresponding dialogue. Lastly, we establish clear requirements for the length and style of the summary. Recognizing the varying lengths of meetings in MISP-Meeting, we implement two summary formats: brief and detailed. The brief summary will skillfully distill key information into a concise format of 200–300 characters, with coherent logic and succinct wording. The detailed summary will then build on this foundation, offering a richer narrative with additional insights, capturing between 800–1200 characters of substantive content.

Experts refinement of ChatGPT-o1 outputs addresses three dimensions:

Size	**Tiny**	**Small**	**Middle**	**Large**	**Total**
Area (in m^2)	0–18	18–36	36–60	60–150	0–150
Train	5	5	3	2	15
Dev	1	1	1	1	4
Eval	1	1	1	1	4

Table 6: Distribution of meeting rooms in the MISP-Meeting dataset.

Model	**Strategy**			**Param**	**MACs**
	GSS	**FT**	**AV**	**(M)**	**(G/30s)**
Whisper-Large-v3	×	×	×	1541.57	1272.87
	✓	×	×	1541.57	1272.87
	✓	✓	×	670.02	988.28
	✓	✓	✓	716.96	1249.88

Table 7: Comparison of parameters and multiply-accumulate operations (MACs) for Whisper-v3-Large models employing various improvement strategies. The corresponding CER results are reported in Table 4. **FT** and **AV**: Fine-tuning and Audio-visual.

1. Completeness and Unity: Ensuring all essential information from the original content is included and consistent across both the brief and detailed summary versions.

2. Level of Detail and Length Constraints: Striking a balance between sufficient granularity and adherence to the required word count for each summary.

3. Clarity and Conciseness: Improving readability by refining sentence structures and removing redundant information.

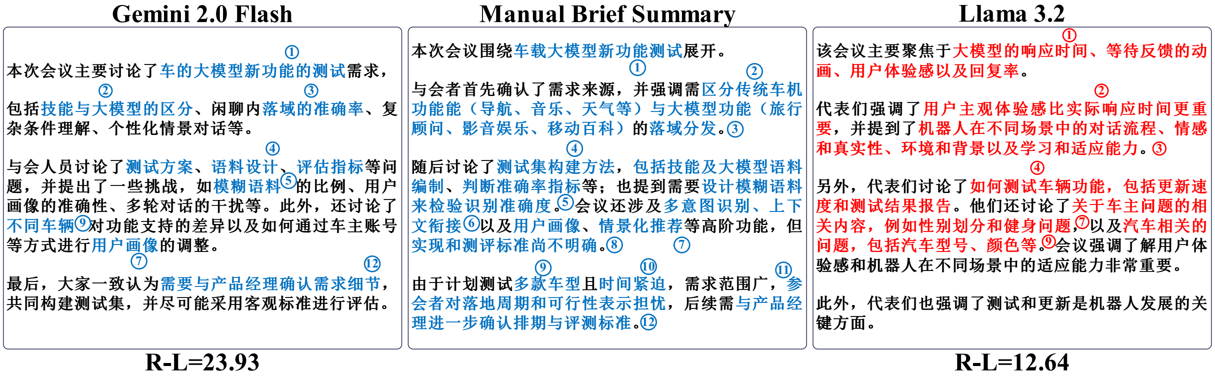
A.1.3 Distribution of Meeting

All meeting rooms are categorized based on size, as shown in Table 6. The dataset covers a range of room sizes, from tiny spaces of 0-18 m^2 to large spaces of 60-150 m^2. The distribution of meeting rooms is balanced across the training, validation, and evaluation sets, ensuring training and evaluation across various room sizes and acoustic environments.

A.2 Experiments Details

A.2.1 Recognition Models and Training Details

We fine-tune the Whisper-v3-Large model on the MISP-Meeting training set by freezing the encoder and re-initializing an attention-based decoder with 6 Transformer layers ($n_{\text{head}} = 8, d_{\text{model}} =$



Gemini 2.0 Flash

本次会议主要讨论了车的大模型新功能的测试①需求，

包括技能与大模型的区分②、闲聊内落域的准确率③、复杂条件理解、个性化情景对话等。

与会人员讨论了测试方案、语料设计④、评估指标等问题，并提出了一些挑战，如模糊语料⑤的比例、用户画像的准确性、多轮对话的干扰等。此外，还讨论了不同车辆⑨对功能支持的差异以及如何通过车主账号等方式进行用户画像⑦的调整。

最后，大家一致认为需要与产品经理确认需求细节⑫，共同构建测试集，并尽可能采用客观标准进行评估。

R-L=23.93

Manual Brief Summary

本次会议围绕车载大模型新功能测试展开。

与会者首先确认了需求来源①，并强调需区分传统车机功能能（导航、音乐、天气等）②与大模型功能（旅行顾问、影音娱乐、移动百科）的落域分发。③

随后讨论了测试集构建方法④，包括技能及大模型语料编制、判断准确率指标等；也提到需要设计模糊语料来检验识别准确度。⑤会议还涉及多意图识别、上下文衔接⑥以及用户画像⑦、情景化推荐等高阶功能，但实现和测评标准尚不明确。⑧

由于计划测试多款车型⑨且时间紧迫⑩，需求范围广，⑪参会者对落地周期和可行性表示担忧，后续需与产品经理进一步确认排期与评测标准。⑫

Llama 3.2

该会议主要聚焦于大模型的响应时间①、等待反馈的动画、用户体验感以及回复率。

代表们强调了用户主观体验感比实际响应时间更重要②，并提到了机器人在不同场景中的对话流程、情感和真实性、环境和背景以及学习和适应能力。③

另外，代表们讨论了如何测试车辆功能④，包括更新速度和测试结果报告。他们还讨论了关于车主问题的相关内容，例如性别划分和健身问题，⑦以及汽车相关的问题，包括汽车型号、颜色等。⑨会议强调了解用户体验感和机器人在不同场景中的适应能力非常重要。

此外，代表们也强调了测试和更新是机器人发展的关键方面。

R-L=12.64

Figure 7: An example illustrating high-scoring and low-scoring brief summaries from Gemini 2.0 Flash and Llama 3.2, respectively. The numbers within each circle represent the key points identified in the manual summary. Blue and red highlights denote alignment and discrepancies with the manual summary, respectively.

$512, d_{ffn} = 2048$). During fine-tuning, we leverage a joint CTC loss with a weight of $\lambda = 0.3$ and the Adam optimizer configured with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate init with 6.0×10^{-4} and linearly warms up for the first 6000 steps, followed by a decline proportionally corresponding to the step number's inverse square root. Adopted data augmentation strategies include speed perturbation, SpecAug, and continuous-segment splicing.

The AVSR model is expanded upon our previously established audio-only model by introducing a visual branch consisting of a ResNet-18 with a 3D-CNN head followed by 3 Transformer layers. For audio-visual fusion, we employ an attention-based cross-modal fusion method. Each fusion layer incorporates a cross-attention block within the Transformer layer of the audio branch, utilizing the embedding generated by the visual branch as queries, while the audio embeddings provide the keys and values.

The parameter counts and multiply-accumulate operations (MACs) of the recognition models are summarised in Table 7. The GSS front-end incurs no additional learnable parameters or computational overhead. During fine-tuning, a lightweight decoder is adopted, reducing the overall parameter count and MACs by 56.54% and 22.36%, respectively, relative to the baseline model. Introducing a visual encoder in the AVSR model increases parameters and MACs by 7.01% and 26.47% when compared with the audio-only model. Nevertheless, both metrics remain lower than those of the baseline model by 54.14% parameters and 1.81% MACs.

We conducted three independent training runs with distinct random seeds to ensure statistical reliability. The best-performing checkpoint on the development set from each run was retained for evaluation. The final metrics represent the mean performance of these three optimal checkpoints. All experiments were performed on an NVIDIA A100 GPU cluster (4 × 80GB), requiring approximately 72 hours per training instance.

A.2.2 Examples

Figure 7 contrasts error-handling strategies in the summarization of the noisy meeting transcript through representative examples from Gemini 2.0 Flash (high-scoring) and Llama 3.2 (low-scoring). The visualization reveals two distinct error propagation patterns: Gemini 2.0 Flash selectively excludes ambiguous content when encountering recognition errors, prioritizing precision over recall, while Llama 3.2 compensates for information gaps through unsupported extrapolation, introducing hallucinations.

Figure 8 illustrates an far-filed overlapping example selected from the evaluation set randomly and the comparison between audio-only and audio-visual recognition results. The target speech overlaps with the interfering speech. Consequently, the target recognition result of the audio-only model incorporates two interference characters. However, the audio-visual recognition model effectively suppresses these artifacts by filtering inconsistent interference segments with lip movements. Otherwise, lip movements also mitigate acoustic attenuation in far-field conditions by reinforcing place/manner of articulation cues.

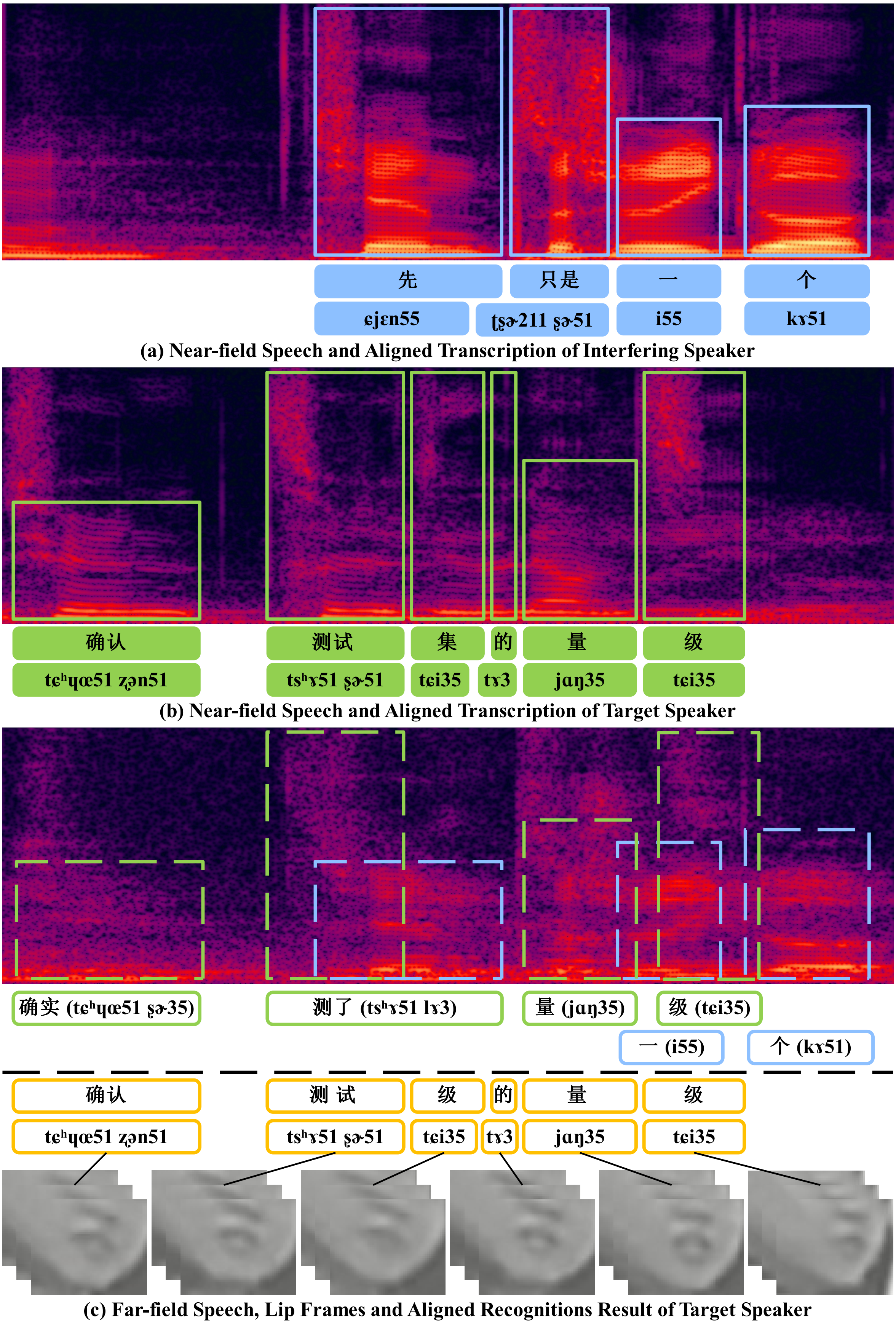


Figure 8: An example demonstrating how visual modality aids in extracting and filling the target speaker's components in far-field overlapped speech. The far-field speech has been enhanced using GSS, and the recognition results correspond to the fine-tuned audio-only and audio-visual results shown in Table 4.