Towards Human-like Multimodal Conversational Agent by Generating Engaging Speech

Anonymous ACL submission

Abstract

Human conversation is usually conducted with language, speech, and visual information. Each communication medium contains rich information and complementary to others, for example, speech (para-lingual) may contain vibe that is not well represented in language. Multimodal LLM consider multimodal information and aim to generate text responses. However, generating more natural and engaging speech response has received little attention even though response only with text cannot give a rich conversation experience. In this paper, we suggest a more human-like agent that makes a speech response based on the conversation mood and responsive style information. Our model is trained to generate text responses along with voice descriptions from multimodal conversation environment. With the voice description, the model generates speech covering para-lingual information. To achieve this goal, we first build a novel multi-sensory conversation dataset mainly focused on speech to enable conversational agents to generate natural speech communication. Then we propose our multimodal LLM based model for generating both text response and voice description. In experimental results, our model demonstrates the effectiveness of utilizing both visual and audio modalities in conversation and generating lively speech.

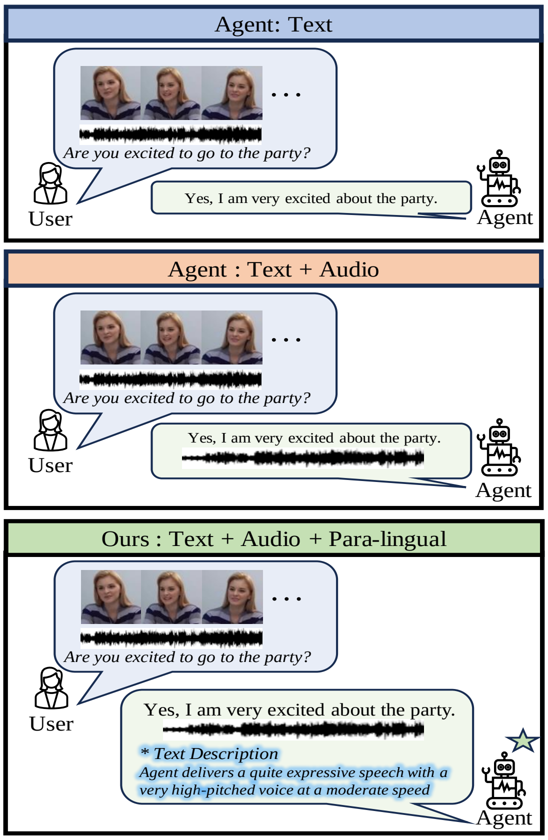

Figure 1: A dialogue example of multi-sensory conversation. (Top) represents text only responsive agent. (Middle) represents text and audio responsive agent. (Bottom) represents text and audio with para-linguistic responsive agent.

1 Introduction

"In real life, people make gestures and read other people's gestures when they communicate. Whether someone is smiling, crying, shouting, or frowning when saying 'thank you' can indicate various feelings from gratitude to irony. People also form their response depending on such context, not only in what they say but also in how they say it (Chu et al., 2018)". Multimodal conversational agents, which can understand both verbal and nonverbal cues such as gestures and tone of voice, have a wide range of potential applications across various domains. It can be employed in customer service interactions to enhance user experience. They can interpret circumstances and tone of voice to better understand customer emotions and address their concerns effectively. In online education, these agents could assist students by gauging their engagement and comprehension through nonverbal cues, adapting the teaching style accordingly.

Recently, communication with machines has become increasingly effective due to the remarkable success of Large Language Models (LLMs). Even considering just open-source models, we see significant advancements in various Question Answering (QA) systems. For instance, Text-based QA systems (Touvron et al., 2023) can understand and respond to text inputs. Visual QA systems (Liu et al., 2023) can interpret both text and image inputs. Video QA systems (Lin et al., 2023) can comprehend text and sequences of images. Audio-Video QA systems (Zhang et al., 2023b) can process text, video, and audio inputs. However, these models are currently only capable of generating text responses.

The easiest way to achieve multimodal communication may be combined with a Text-To-Speech (TTS) module. However, current TTS modules are inadequate for effective communication. For instance, TTS modules (Popov et al., 2021; Shen et al., 2023; Li et al., 2024) cannot generate speech that incorporates para-linguistic information reflecting the communication moo. To address these challenges, we propose our novel speech generation model with paralingual information.

The creation of such conversational model relies on exposure to a diverse range of multimodal conversations that seamlessly integrate textual, visual, and acoustic elements. To comprehend multimodal information in conversations, we adopt the BLIP-2 (Li et al., 2023) approach to ensure efficient cross-modal training. To capture variations in visual scenes within videos, we employ a pre-trained visual encoder to compute frame representations separately. A video Q-Former is then introduced to generate visual query tokens. For audio signals from the video, we utilize a pre-trained audio encoder and an audio Q-Former to learn effective auditory query embeddings. Finally, to generate conversational responses with paralinguistic components derived from the overall communication atmosphere, we use instruction tuning. This guides our model to generate voice descriptions that reflect the desired speech atmosphere.

In order to develop the proposed conversational agent, a substantial corpus of multimodal interactive conversation data of considerable scale is desirable. However, there are limitations in the dataset available for training the model such as smaller scale or missing modality like audio. To overcome these limitations, we present a new dataset called *MultiSensory Conversation (MSC)* dataset. Our dataset is a carefully curated collection of about 31,000 utterances extracted from educational YouTube videos. These videos encompass straightforward and natural conversational scenarios, making them well-suited for the training of our model.

The contributions of our work can be summarized as follows:

- To the best of our knowledge, we are first to study a dialogue model incorporating paralingual output in responses. We generate speech with paralinguistic information reflecting multimodal factors in conversation.
- We introduce the MultiSensory Conversation (MSC) dataset, a collection of around 31,000 utterances from educational YouTube videos, which will be publicly available to advance research in multimodal conversational agents.
- Our model effectively utilizes both visual and auditory modalities, producing natural and contextually appropriate speech responses, as validated by both quantitative metrics and qualitative assessments.

2 Related Work

2.1 Multimodal LLM

Large Language Models (LLMs) have demonstrated a high level of common knowledge (Achiam et al., 2023). Initial attempts to leverage this knowledge for vision-language tasks mainly involve adding visual information to LLMs. The common approach is to encode image features using a pre-trained vision model, project these features, and then directly input them into the LLM (Lin et al., 2023; Zhang et al., 2023b; Liu et al., 2024; Chen et al., 2023). Traditional vision-language datasets (Sharma et al., 2018; Schuhmann et al., 2022) are not designed for instruction-following tasks (Liu et al., 2024). To address this, detailed captions and object bounding box information are provided to the LLM, creating an instruction-following dataset. Models trained on this dataset exhibits impressive multimodal conversation abilities (Liu et al., 2024).

Beyond vision-language tasks, there have been efforts to integrate various modalities into LLMs. While vision-language tasks primarily focus on generating text from image inputs, there have also been attempts to generate other modalities using LLMs (Wu et al., 2023; Tang et al., 2024). These models try to retain the semantic information of the input but often struggle with consistency across

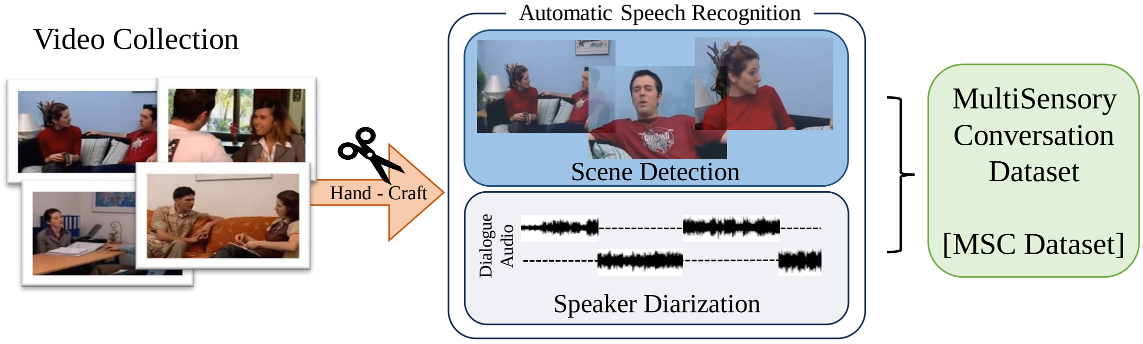

Figure 2: The illustration depicts the creation process of the MultiSensory Conversation dataset. Initially, raw video segments are manually divided into dialogue units. Subsequently, each utterance undergoes automatic speech recognition (ASR) to further refine segmentation, supported by scene detection and speaker diarization techniques.

modalities. Our approach enables speech interaction with LLMs without losing consistency by merging TTS systems, circumventing the aforementioned drawbacks.

2.2 Text-to-Speech

Diffusion models have gained traction in speech synthesis due to their potential for diverse speech sampling and fine-grained speech control (Zhang et al., 2023a). Probabilistic diffusion and large pretrained speech language models achieve human-level performance in synthesizing natural and diverse speech (Popov et al., 2021; Huang et al., 2022b,a). Toward human-level TTS systems, modeling speech styles as a latent random variable show the potential on both single and multispeaker (Li et al., 2022, 2024). Alternatively, natural language prompting of speaker identity and style has demonstrated promising results and provides an intuitive method of control (Lyth and King, 2024). Our approach follows the natural language prompt method for generating voice descriptions. By generating responsive voice descriptions that consider the conversation history, we can enhance the naturalness and contextual appropriateness of TTS outputs in dialogue systems.

3 Data

The majority of existing datasets for multimodal conversation primarily involve utterances consisting of single speakers, in the case of AVSpeech (Ephrat et al., 2018) and MEAD (Wang et al., 2020) where one speaker provides continuous utterances, or MovieChat (Chu et al., 2018) do not involve scene images or audios but the dataset has texts and facial landmarks. However, to effectively communicate in a more human-like way, a dataset that encompasses both looking at and conversing with human faces along with voice is desirable and no dataset has been curated with this precise focus in mind. One notable dataset is the MELD (Poria et al., 2018) that provides both facial images and audio. However, since it was initially designed for multimodal emotional analysis, it may not always achieve precise audio splitting, which could result in some parts of the speech being missing or cut off. Also, since it originated from the TV series Friends, most of the clips contain noise from audience reactions not adequate for training natural human-like speech generation models.

	Train	**Valid**	**Test**	**Total**
# of Dialogue	913	110	97	1120
# of Utterance	25624	3145	2640	31409
Duration	17.5h	2.1h	1.8h	21.5h

Table 1: Statistics of the MultiSensory Conversation dataset.

To address these limitations, we have taken the initiative to develop our novel dataset, the MultiSensory Conversation Dataset depicted in Figure 2. This dataset originated from YouTube, and because it is an educational video that allows people to communicate fluently in English, it consists of natural conversations containing abundant visual components for conversation such as background, human face, gestures and various aspects of voice features such as pitch, volume, timbre, and prosody.

3.1 Preprocessing

3.1.1 Dialogue Split

Manually segmenting over 36 hours of videos by speech is a challenging task for an individual. Also, it is necessary to check if any parts are not ap-

Figure 3: An example of MultiSensory Conversation dataset. This illustration shows text, audio, and videos from about 31,000 utterances obtained from educational YouTube videos.Dialogues within a single utterance are separated using ASR, scene detection, and speaker diarization techniques.

propriate for learning conversations. So we proceeded to partition the data into units of dialogue manually, aiming to address any existing inappropriateness. The criteria guiding the separation of dialogues were as follows: 1) When multiple dialogues occurred within a single context. 2) In instances where the scene transitioned to a different setting during the conversation. 3) When transitioning between similar scenes, provided that the individuals involved changed.

3.1.2 Utterance Split

To efficiently split dialogue videos into individual utterances, we can use a technique called Speaker Diarization, aimed at segmenting and indexing audio recordings by speaker identity and marking speech timestamps. However, it has some limitations, such as difficulty in accurately identifying speakers and overly fragmenting single utterances.

To address these issues, we incorporated Automatic Speech Recognition (ASR) with timestamp capabilities. In our approach, we utilized a pre-trained ASR model[1] that trains OpenAI's Whisper-large-v3 (Radford et al., 2023) on English-only data, providing more accurate and faster inference speeds. However, since this model is trained for audio clips up to 25 seconds long, it struggles to accurately timestamp longer clips. To overcome this, we applied a scene detector[2] to divide longer audio into shorter clips. For clips still exceeding 25 seconds, we employed speaker diarization[3]. This method allowed us to more effectively segment the entire video into distinct speech units, each corresponding to individual speakers. Figure 3 shows a sample of MSC dataset.

[1] distil-whisper/distil-large-v3
[2] https://github.com/Breakthrough/PySceneDetect
[3] pyannote/speaker-diarization-3.1

3.2 Metadata Processing

3.2.1 Speaker Assign

We assign a speaker ID to each video clip according to dialogue units. While speaker diarization is the desirable method for segmenting and indexing speakers to utterances, as mentioned earlier, it has limitations in speaker identification performance. We take an alternative approach to address this limitation: cluster the speech embedding. Figure 4 shows our approach. We obtain speech embeddings from each video clip using WeSpeaker [4] (Wang et al., 2023), a tool focused on speaker embedding learning, particularly for speaker verification tasks. By grouping speech embeddings, we perform clustering with the HDBSCAN (McInnes et al., 2017) algorithm, which can handle variable density and does not require specifying the number of clusters. We use cosine distance as the metric since most speaker verification systems utilize cosine similarity for evaluation. This method allows us to assign each entire utterance to individual speakers effectively.

3.2.2 Speech Description

Since our goal is speech generation, we decided to extract para-lingual information that accurately describes speech. Parler-TTS (Lyth and King, 2024) is a Text-to-Speech (TTS) system that transforms text into speech, incorporating detailed speech descriptions such as gender, pitch, speaking style, etc. This system provides methods for creating these descriptions, which we utilized in our process. From the MSC dataset, we extract pitch, gender, speech monotony, speaking pace, and reverberation excluding noise. To verify, especially for gender, we conduct gender recognition [5] from audio which

[4] pyannote/wespeaker-voxceleb-resnet34-LM
[5] alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech

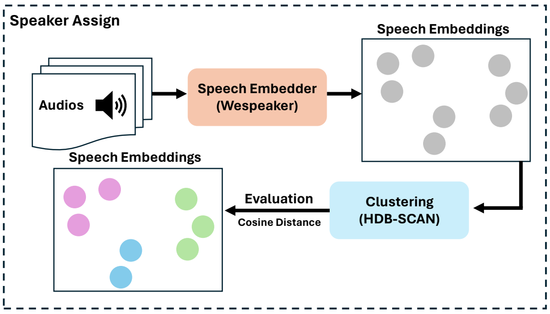

Figure 4: Illustration of our speaker assignment pipeline: we obtain speech embeddings using WeSpeaker and perform clustering with the HDB-SCAN algorithm.

shows 99.93 of F1 score. After that, we generate natural language descriptions of them.

3.3 Data Statistic

The statistics are presented in Table 1. To summarize, we divided the video content into a total of 1,120 dialogues and 31,409 utterances. The total video length is 21.5 hours. The average duration of an utterance is 2.46 seconds. You can find more details in Appendix B.

4 Model

We develop an end-to-end model capable of processing data from multiple modalities within a large language model (LLM). Our model takes in a set of images, audio, and text as inputs as a single utterance and generates responsive textual sentence with voice description. Figure 5 shows the overview of our architecture. We denote our dataset as $D = \left\{d^a, d^v, d^l\right\}$ where a is acoustic, v is visual, and l is language. And each dialogue consists of a set of utterances. Let $d^m = \{u_1^m, u_2^m, ..., u_t^m, u_{t+1}^m\}$ as single dialogue and t is the order of utterance, and m presents modality. Note that the dataset includes several dialogues, but they are independent of each other. For single utterance $u_t^m = \left\{u_t^a, u_t^v, u_t^l\right\}$, video and audio modalities go through each Q-Former to generate a representation vector. Then the processed utterance was brought together to LLM. LLM input is integrated with the conversation history $\{u_1^m, u_2^m, ..., u_{t-1}^m, u_t^m\}$. Ultimately, LLM generate the output $\left\{\hat{u}_{t+1}^l, \widehat{desc}_{t+1}\right\}$ which is text modality.

4.1 Multimodal Understanding

In Video-LLaMA (Zhang et al., 2023b), the video and audio data are trained on each Q-Former, which shares the same structure as Blip-2 (Li et al., 2023). To initialize the Video Q-Former and Audio Q-Former, we adopt the pretrained Q-Former from Blip-2. These models are fine-tuned to enable understanding of visual and auditory information in conversations. Within Q-Former, queries interact via self-attention layers and with frozen feature encoders via cross-attention layers. To match the extracted feature's dimension of video and audio to the dimension of pretrained Q-Former during the cross-attention process, we add a linear projection layer inside Q-Former. They extract a fixed number of output features from both the image encoder and audio encoder, regardless of the length of input video and audio. In the Video Q-Former, we consider the image feature list as a conversation scene. For video sampling, we uniformly extract three frames per second. However, in the Audio Q-Former, the entire feature of the speech is taken as input. While sampling is conducted for videos to reduce redundant information and improve efficiency, the same method cannot be applied to audio due to significant information loss. Nevertheless, Q-Former's consistent output length characteristic helps mitigate the miss-length issue between video and audio information. The features after Q-Former will concatenate with textual information obtained from the embedding token of LLM and treat it as an utterance feature.

4.2 Speech Description Generation

If the model can understand the intention of a single utterance containing multimodal information, reading conversation mood is possible with dialogue history. We utilized LLM capable of understanding dialogue history which is sequential information. The features processed through the Q-Former are projected into the embedding space of LLM using a linear layer. Additionally, we employed Instruction tuning to provide information about which speaker is delivering each utterance.

The response considering the conversation mood can be obtained in text format. But in order to provide richer communication, we've trained our model to reflect not just linguistic information but also para-linguistic cues by describing voice. Our model first generates the response text and then produces a description influenced by that text. To accomplish this, we've introduced instruction tuning, a new process where voice descriptions are created after the language model generates responses. We also give instructions about who should speak,

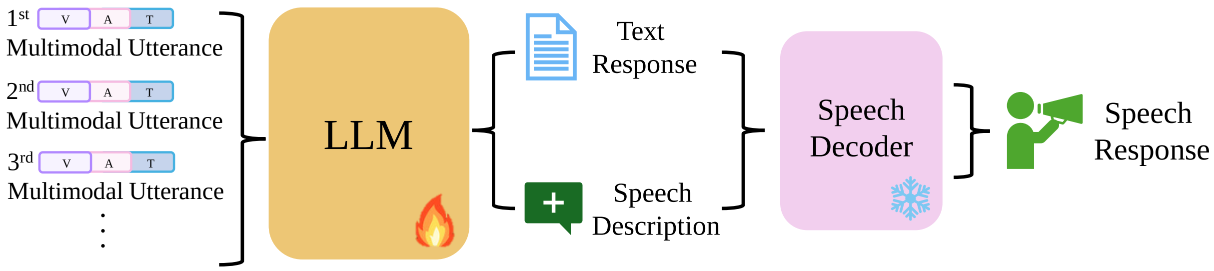

Figure 5: Overview of our model architecture. Multimodal Utterances, composed of text, audio, and video features, are input into LLM(Large Language Model). LLM generates Text Response and Speech Description. These outputs are then processed by Speech Decoder(TTS), which produces Speech Response.

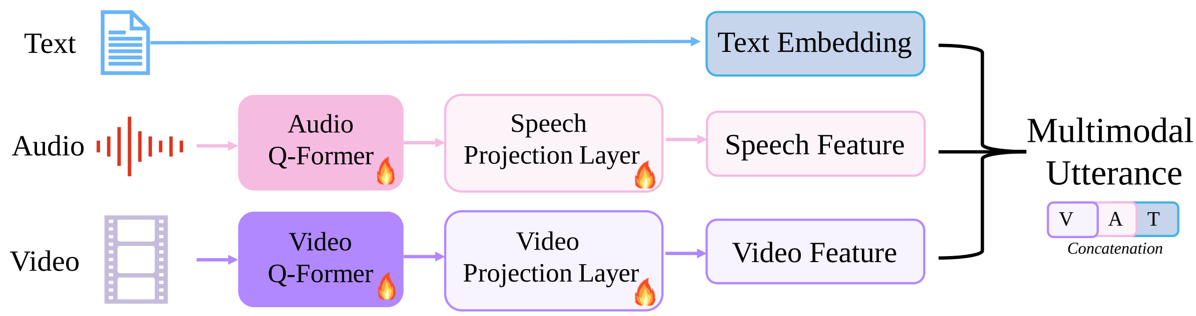

Figure 6: Workflow of Multimodal Encoding. Text, Audio, and Video inputs are processed independently. Text is converted into text embedding, audio is processed into Speech Feature via the Audio Q-Former and Speech Projection Layer, and Video is processed into video Feature via Video Q-Former and Video Projection Layer. These features are concatenated to form a Multimodal Utterance, integrating information from three modalities.

which makes a model response or continues the previous utterance. More details about instruction tuning are presented in Appendix C.

4.3 Training Loss

Our approach involves end-to-end training. Initially, we use a target reference sentence paired with its corresponding audio description. The cross-entropy loss is then computed between the target and the model output, as illustrated in Equation1, with the concatenation operation $\|$.

$$Loss = CE(u^l_{t+1} \,\|\, desc_{t+1}, \hat{u}^l_{t+1} \,\|\, \hat{desc}_{t+1}) \quad (1)$$

Our training primarily emphasizes the Video Q-Former and Audio Q-Former models. Furthermore, we fine-tune the large language model backbone with parameter efficient fine tuning (Hu et al., 2021) to specialize the model specifically for the conversation task.

5 Experiment

5.1 Experimental setup

5.1.1 Multimodal Feature Extraction

We obtained modality-specific data from each video segment, corresponding to an utterance unit. For visual data, we extract visual features with CLIP-VIT (Radford et al., 2021). This model has a strong alignment with text, having a potential impact on downstream tasks. The audio data extraction process gets an acoustic feature with WavLM (Chen et al., 2022). This model tries to solve full-stack downstream speech tasks with speech information including speaker identity, paralinguistics, and spoken content.

5.1.2 Evaluation

In our experiments, we used two datasets: our MSC dataset and the MELD dataset (Poria et al., 2018). To evaluate our model's performance, we employed several metrics commonly used in natural language processing. These included the BLEU score (Papineni et al., 2002), which measures n-gram overlap between machine-generated text and reference text. We also utilized METEOR (Banerjee and Lavie, 2005), designed to address limitations of BLEU by considering factors like synonymy, stemming, word order, and recall. Additionally, we employed ROUGE (Lin, 2004), which is particularly useful for evaluating the coherence and flow of summaries and translations. These metrics collectively provided a thorough assessment of our model's capa-

Modality	Datasets							
	MSC				MELD			
	B@1	B@3	METEOR	ROUGE	B@1	B@3	METEOR	ROUGE
Text	12.30	4.11	5.81	11.90	7.99	1.60	4.47	8.09
Text + Audio	12.96	4.82	6.27	11.83	9.10	2.11	4.35	8.24
Text + Video	14.62	4.78	6.63	13.38	5.62	1.00	2.53	4.03
Text + Audio + Video	**15.11**	**5.25**	**6.89**	**14.12**	**10.23**	**2.19**	**4.74**	**9.88**

Table 2: Ablation study on different modalities across two datasets. The text-only modality model represents a pure LLM that has been fine-tuned with each dataset.

bility to generate high-quality text outputs compared to reference data.

5.2 Text

5.2.1 Modality Ablation

Given that our model processes a multimodal input, comprehending the impact of each modality on its performance becomes crucial. Therefore, our aim is to assess how metrics alter as we integrate information from diverse modalities into the existing large language model. Table 2 shows the impact of audio and video features. According to the MSC dataset result, The addition of audio features and video features influences the enhancement of conversational generation outcomes. Furthermore, The incorporation of audio and video input noticeable increase in the score. It is the same for the MELD dataset, where incorporating audio and video inputs also results in the highest performance. However, the scale of the score is smaller, which implies that the MSC dataset is more suitable for the tasks we presented.

5.2.2 Qualitative Analysis

LLMs(Large Language Models) have demonstrated remarkable capabilities in generating text based solely on textual input. However, LLMs' understanding and response generation is limited when it comes to interpreting the emotional context behind the same textual content presented with different emotions. c For instance, text-based LLM might understand the sentence "Hello, how are you?" the same way, regardless of whether the speaker is happy or sad. Because it lacks access to non-verbal cues such as tone of voice, or facial expressions that convey these emotions.

In the Qualitative Analysis of evaluating multimodal model, we have demonstrated our model's capability to understand multimodality through metric scores. This is evident in the enhanced performance achieved by integrating text, audio,

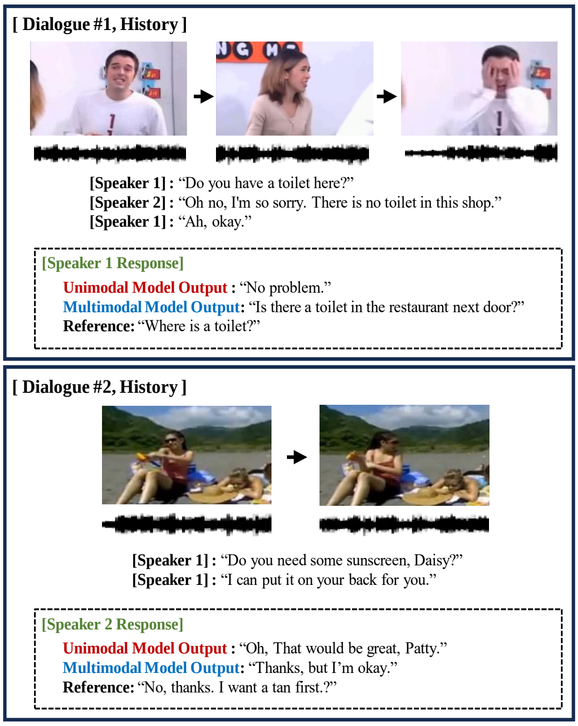

Figure 7: Qualitative Evaluation of Multimodality. We evaluate on our dataset, namely MultiSensory Conversation Dataset.

and video. However, it is worth noting that metrics alone might not capture the full essence in an open-domain scenario. Consequently, we present a comparative analysis of our model's outputs against those of the text-based unimodal model in Figure 7. Our model generates more natural responses and demonstrates a better understanding of the context than the unimodal model. The figure provides two dialogues from different scenarios, illustrating how the inclusion of additional modalities (audio and video) enables our model to produce more contextually appropriate and natural responses. In Dialogue 1, the speaker's gestures in the video and tone of voice in the audio clearly indicate an urgent situation. In Dialogue 2, the output text adapts based on information from the video, resulting in a generated

Model	Accuracy
Ours	15.10%
Ours (w.o. description)	11.20%
StyleTTS2	13.72%
HierSpeech++	12.54%

Table 3: Emotion classification result.

text that closely matches the reference.

5.3 Speech

5.3.1 Emotion Classification

In this experiment, we performed emotion classification using one of eight emotions: angry, calm, disgust, fearful, happy, neutral, sad, and surprised. The results demonstrated in Table 3. We use a pretrained model from Hugging Face for emotion classification [6]. The baselines, including StyleTTS2 (Li et al., 2024), HierSpeech++(Lee et al., 2023), and Parler-TTS(Lyth and King, 2024) which generated speech without natural language prompts. Our model generates each speech sample from text and voice descriptions and then compares it with previous speech samples to assess consistency. Results show our model outperformed the baseline models in maintaining consistent emotional expression across the conversation.

5.3.2 Qualitative Analysis

In the Qualitative Analysis of evaluating voice description, we have demonstrated our model's capability to generate consistent emotional description. We present a comparative analysis of our model's outputs against those of the reference one in Figure 8. The figure provides two dialogues from different scenarios, demonstrating our model generates similar descriptions in terms of pace, pitch, and tone which leads to producing more contextually appropriate and natural responses.

6 Limitation

One limitation of our model is its inability to generate speech with a speaker's identical voice as it appears in historical recordings. However, this does not pose an issue during inference, as the agent consistently uses the same voice. Potential risks include the copyright concerns associated with YouTube videos. Since sharing downloaded videos is prohibited, we only provide the preprocessing code to ensure compliance with copyright laws. This approach allows users to process their own legally obtained data without violating any terms of service or copyright regulations.

[6]ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition

Figure 8: Qualitative Evaluation of description. We evaluate on our dataset, namely MultiSensory Conversation Dataset.

7 Conclusion

We study a dialogue model with visual and audio inputs from a speaker, which is essential for a more human-like conversation model. We propose a novel dataset that is suitable and curated for training such a model. Then we propose a novel multi-sensory conversation model that outperforms the baseline in experiments and thus shows its effectiveness in both quantitative and qualitative evaluations. In the ablation study, we also demonstrate the importance of each modality we exploited. In the future, we aim to use and extend our model for a more human-like appearance by merging with Talking Face Generation from speech inputs (Zhou et al., 2020) (Zhou et al., 2021) (Zhang et al., 2023c) to considering emotional components (Peng et al., 2023) (Gan et al., 2023). We believe our approach contributes to more natural and human-like conversation and our proposed dataset may promote subsequent research in conversation models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Hang Chu, Daiqing Li, and Sanja Fidler. 2018. A face-to-face neural conversation model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7113–7121.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. 2023. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22634–22645.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. 2022a. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*.

Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022b. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Sang-Hoon Lee, Ha-Yeong Choi, Seung-Bin Kim, and Seong-Whan Lee. 2023. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *arXiv preprint arXiv:2311.12454*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2022. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *arXiv preprint arXiv:2205.15439*.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2024. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

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.

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. 2023. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

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.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2024. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. 2023a. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. *arXiv preprint arXiv:2303.13336*, 2:2.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023c. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661.

Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15.

A Implementation Details

We utilize Mistral-7B (Jiang et al., 2023) as our LLM backbone. We train our model with the following hyperparameters. We use a batch size of 6 and Adam optimizer with learning rate of 5e-5 and learning rate decay of 0.98. The video padding size is 50, audio padding size is 800. This size made the same number of utterances in a single dialogue history. We sample the video data, capturing frames at a rate of three per second for each utterance, while the audio remains unsampled. We set the maximum input length for LLM as 800 which can cover about 10 multimodal histories. They are truncated from the oldest history to prioritize focusing more on the latest utterance. Finally, we tuned the number of epochs on validation data and chose epoch 10. Our experimental environment was conducted using a single NVIDIA-A100 80G GPU. Training spent 30 hours.

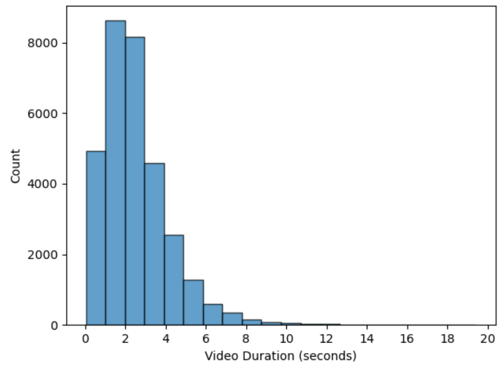

Figure 9: We report the histogram of video duration in seconds.

B MSC Dataset Details

In this section, we show further details of the new MSC dataset. The histograms of video durations and word count can be found in Figure 9, 10. Note that many videos begin with greetings such as "Hello" or "Good Morning", which contribute to a higher word count due to there conciseness. More detailed examples of the dataset can be found in Figure 11.

C Instruction-tuning

We give comprehensive instruction first and give speaker ID information for each of utterance. Lastly, we give another instruction for generating voice descriptions. Figure 12 shows a sample of instruction tuning. This sample demonstrates text input for easy understanding, though actual input includes not only text but also integrated text, audio, and video modalities.

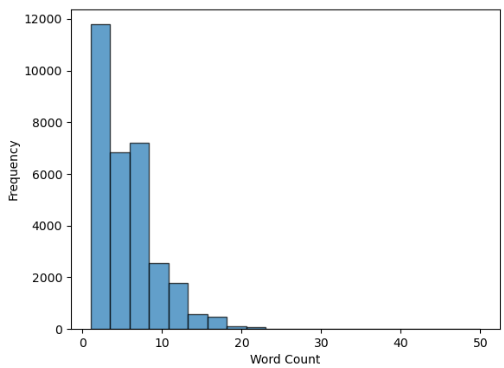

Figure 10: We report the histogram of word count in words.

D LLM fine-tuning

We investigated the impact of fine-tuning a large language model with parameter efficient fine-tuning at Table 4, 5. This indicates that after fine-tuning, the model exhibited enhanced conversational capabilities compared to its pre-fine-tuned state.

Figure 11: MSC Dataset details

$$\mathrm{I}\ \ \mathrm{H}_1\ [\mathrm{utt}_1]\ \ \mathrm{H}_2[\mathrm{utt}_2]\ \ \cdots\ \ \mathrm{H}_n[\mathrm{utt}_n]\ \ \mathrm{R}$$

I: ### Instruction: Generate a following response of this conversation.
H_1: ### Speaker_{1}: Hey, how was your vacation?
H_2: ### Speaker_{2}: It was very fun.
H_3: ### Speaker_{1}: Where did you go?
H_4: ### Speaker_{2}: I went to the beach.
H_5: ### Speaker_{1}: Who did you go with?
R: ###Generate a response format as [text of response (voice description)], Speaker_{2}: "

Figure 12: Sample of an LLM input with instructions. This sample demonstrates text input for easy understanding, though actual input includes not only text but also integrated text, audio, and video modalities.

	B@1	B@2	B@3	B@4	METEOR	ROUGE	SPICE	CIDEr
Ours w.o.ft	13.96	7.96	5.03	3.25	6.55	12.77	4.01	34.98
Ours	**15.11**	**8.57**	**5.25**	**3.35**	**6.89**	**14.12**	**4.02**	**38.53**

Table 4: impact of LLM fine-tune on MSC dataset.

	B@1	B@2	B@3	B@4	METEOR	ROUGE	SPICE	CIDEr
Ours w.o.ft	5.67	2.11	0.97	0.48	2.90	4.95	1.02	6.13
Ours	**10.23**	**4.33**	**2.19**	**1.21**	**4.74**	**9.88**	**2.25**	**16.63**

Table 5: impact of LLM fine-tune on MELD dataset.