Towards Human-like Multimodal Conversational Agent by Generating Engaging Speech

Anonymous ACL submission

Abstract

Human conversation involves language, speech, and visual cues, with each communication medium providing complementary information. For instance, speech often conveys a vibe or tone not fully captured by text alone. While multimodal LLMs focus on generating text responses from diverse inputs, less attention has been given to generating natural and engaging speech. In this paper, we propose a more human-like agent that generates speech responses based on conversation mood and responsive style 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 which are used to generates speech covering para-lingual information. In experimental results, our model demonstrates the effectiveness of utilizing both visual and audio modalities in conversation and generating lively speech.

1 Introduction

In real life, people communicate through multimodal signals and interpret others' nonverbal cues. For instance, whether someone is smiling, crying, shouting, or frowning when saying 'thank you' can indicate various feelings from gratitude to irony. (Chu et al., 2018). This highlights the importance of multimodal understanding—where words, facial expressions, and vibes all contribute to the interpretation of meaning. Moreover, individuals often adjust their responses based on this context, not just in what they say but also in how they deliver their messages, reflecting the nuanced dynamics of tone, emphasis, and delivery. In this paper, we focus on developing a multimodal conversational agent capable of recognizing and responding to these cues. Such agents have diverse applications, including enhancing customer service by interpreting tone and emotions and improving online education by adapting teaching styles based on students' nonverbal engagement and comprehension.

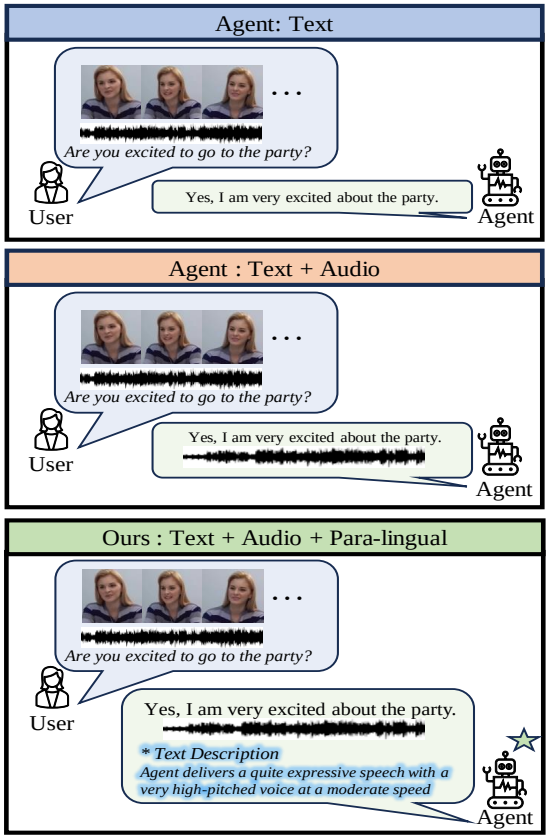

Figure 1: A dialogue example of multi-sensory conversation: (Top) text-only agent, (Middle) text and audio agent, (Bottom) text, audio, and para-linguistic agent.

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. Despite these advances, these models are currently only capable of generating text responses.

Integrating a text-to-speech (TTS) module is a straightforward approach to achieving multimodal communication. 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 considers paralingual information reflecting the communication mood. To address these challenges, we present a new dataset and propose our novel conversational model.

Developing the proposed conversational agent requires a substantial corpus of multimodal interactive conversation data of considerable scale, which presents a significant challenge due to the limitations in the available datasets, such as their smaller size or the absence of certain modalities like audio. To overcome these limitations, we present a new dataset called *MultiSensory Conversation (MSenC)* dataset. Our dataset is a carefully curated collection of about 31,000 utterances extracted from educational YouTube videos. These videos encompass natural and dynamic conversational scenarios with clear and high-quality speech, making them well-suited for training the model.

The creation of such a 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 utilize LLM and instruction tuning. This guides our model to generate voice descriptions that reflect the desired speech atmosphere.

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 (MSenC) 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

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

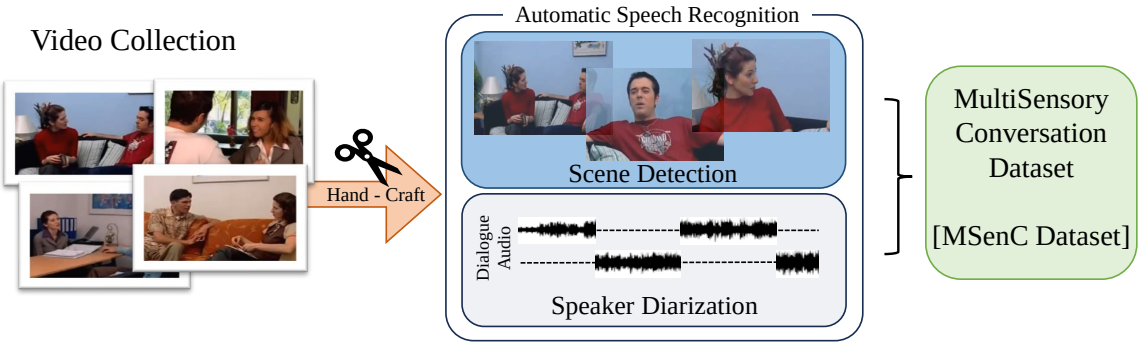

Figure 2: The illustration depicts the creation process of the MultiSensory Conversation dataset. Initially, raw video segments are manually divided into dialogue units. Subsequently, each utterance undergoes automatic speech recognition to refine further segmentation, supported by scene detection and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

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 pre-trained 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 shows the potential on both single and multi-speaker (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

Most existing multimodal conversation datasets, like AVSpeech (Ephrat et al., 2018), MEAD (Wang et al., 2020), and MovieChat (Chu et al., 2018), focus on single-speaker utterances and lack comprehensive multimodal features. 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 none exist with this specific focus. One notable dataset is the MELD (Poria et al., 2018) that provides facial images and audio but is designed for emotional analysis, which may lead to imprecise audio splitting and audience noise, making it unsuitable 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 MSensory Conversation dataset.

To address these limitations, we have taken the initiative to develop our novel dataset, the MultiSensory Conversation (MSenC) dataset depicted in Figure 2. This dataset, sourced from YouTube and designed for educational purposes, ensures that there is no background music and that each scene includes a person. It features natural conversations with rich visual and auditory elements that enhance the contextual understanding of dialogue situations. These videos offer a diverse range of voice features and interactions across various scenarios and contexts, crucial for developing robust models. Moreover, the educational content ensures that the spoken English is clear and high-quality, avoiding issues of low fluency. The dataset construction pipeline that will be made publicly available is also easily expandable to other languages and video types.

3.1 Preprocessing

3.1.1 Dialogue Split

Manually segmenting over 36 hours of videos based on speech is a challenging task. However, it is crucial to carefully check for any unnecessary parts to ensure the content is suitable for learning conversations. So we proceeded to manually segment the dialogue by human to ensure fairness and accuracy. The criteria guiding the separation of dialogues were as follows: 1) When multiple

dialogues occurred within the same 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 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.

3.2 Metadata Processing

3.2.1 Speaker Assignment

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 3 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 densities 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. To assess the quality of our process, we calculated the accuracy of our method by manually labeling 20 dialogues, which included a total of 602 samples, resulting in an accuracy of 95.49%. This demonstrates the effectiveness of our speaker assignment method.

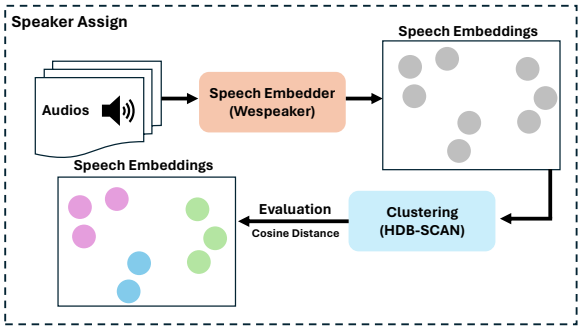

Figure 3: Illustration of speaker assignment pipeline. We obtain speech embeddings and perform clustering.

3.2.2 Speech Description

Since our goal is speech generation, we extracted para-lingual information that accurately describes speech. Parler-TTS (Lyth and King, 2024) is a text-to-speech system that transforms text into speech, incorporating detailed descriptions such as gender, pitch, speaking style, etc. This system provides methods for generating descriptions, allowing us to add annotations using the dataspeech [5] library. The resulting dataset includes discrete annotation tags and a coherent audio description of the spoken characteristics. For the MSenC dataset, we extract pitch, speech monotony, speaking pace, and reverberation while excluding noise level. Especially for gender, which is needed for generating speech descriptions but cannot be derived directly, we conduct gender recognition [6] from the raw speech, achieving an F1 score of 99.93%. Subsequently, the LLaMa-3 (Dubey et al., 2024) generates descriptions that reflect the mood of the conversation based on these annotations.

3.3 Data Statistic

The statistics are presented in Table 1. To summarize, we divided the video content into 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. More details with data samples can be found in Appendix B.

[1] distil-whisper/distil-large-v3
[2] https://github.com/Breakthrough/PySceneDetect
[3] pyannote/speaker-diarization-3.1
[4] pyannote/wespeaker-voxceleb-resnet34-LM
[5] https://github.com/huggingface/dataspeech
[6] alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech

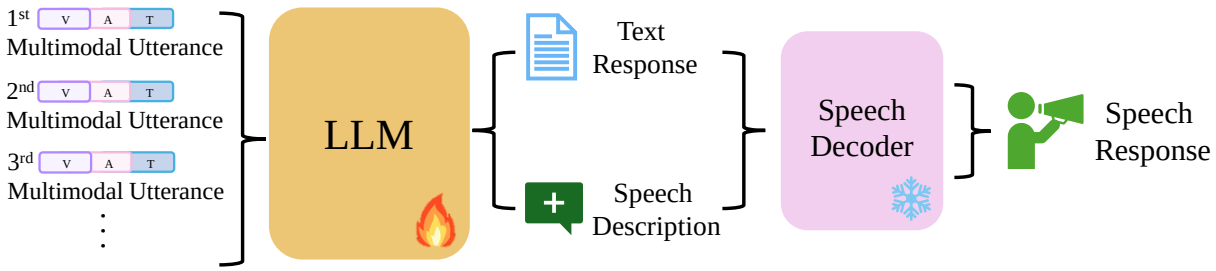

Figure 4: Overview of our model architecture. The multimodal utterances, composed of text, audio, and video features, are inputted into LLM. Then LLM generates text response and speech description. These outputs are then processed by speech decoder, which produces speech response.

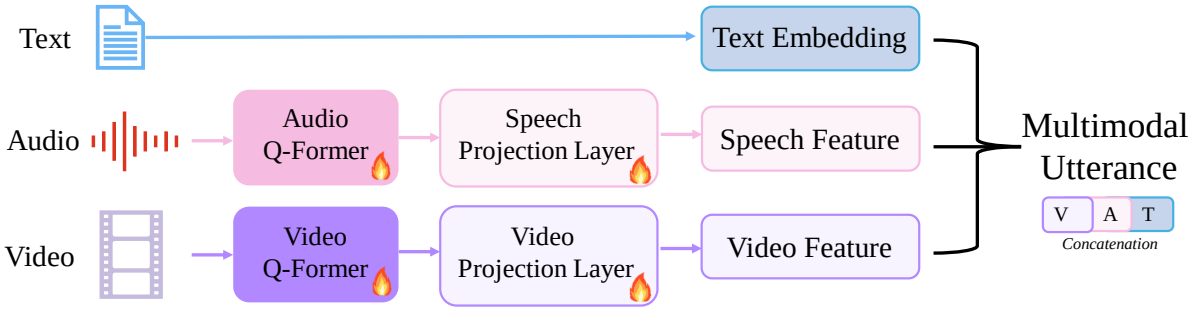

Figure 5: Workflow of multimodal encoding. Text, audio, and video inputs are processed independently. Text is converted into text embedding, audio is processed into speech feature via the audio Q-Former and speech projection layer, and video is processed into video feature via the video Q-Former and video projection layer. These features are concatenated to form a multimodal utterance, integrating information from multiple modalities.

4 Method

We develop a model capable of processing data from multiple modalities and generating engaging speech within a large language model. Figure 4 shows the overview of our architecture. Our model takes in a set of images, audio, and text as a single utterance input and generates a responsive textual sentence output along with a voice description. We denote our dataset as $D = \left\{d^a, d^v, d^l\right\}$ where a represents the acoustic modality, v the visual modality, and l the linguistic modality, with d indicating dialogue. Each dialogue consists of a set of utterances. Let $d^m = \{u_1^m, u_2^m, ..., u_t^m, u_{t+1}^m\}$ represent a single dialogue, where t denotes the order of utterance, and m presents each modality. Note that the dataset includes multiple dialogues, but they are independent.

For single utterance $u_t^m = \left\{u_t^a, u_t^v, u_t^l\right\}$, the video and audio modalities are processed through their respective Q-Former to generate representation vectors. The processed utterances are then combined with the conversation history $\{u_1^m, u_2^m, ..., u_{t-1}^m, u_t^m\}$ and fed into the LLM. Ultimately, the model generates the output $\left\{\hat{u}_{t+1}^l, \hat{desc}_{t+1}\right\}$ in textual format where $desc$ represent speech description.

4.1 Multimodal Understanding

In Video-LLaMA (Zhang et al., 2023b), the video and audio modalities are trained on separate Q-Former. This Q-Former shares the same structure as Blip-2 (Li et al., 2023) and has demonstrated strong performance by enhancing computational efficiency and model stability. It produces fixed-size features, which simplifies the integration of multimodal data and ensures consistent input sizes for subsequent processing. To initialize the video Q-Former and audio Q-Former, we adopt the pre-trained 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 list of

Modality	Datasets							
	MSenC				MELD (Poria et al., 2018)			
	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. "B" stands for BLEU score.

image features as a conversation scene. For video sampling, we uniformly extract three frames per second. In contrast, the audio Q-Former takes the entire speech feature 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, the consistent output length characteristic of the Q-Former helps prevent the length mismatch between video and audio.

The features processed through each Q-Former are projected into the embedding space of the text using a linear layer for each. These features are then concatenated with features from other modalities and treated as utterance features.

4.2 Speech Description Generation

If the model can understand the intention of a single utterance containing multimodal information, it can also interpret the conversation mood through dialogue history. We utilize the LLM capable of understanding sequential information from dialogue history. The sequence of utterance features serves as the dialogue history and is fed into the LLM to comprehensively understand the context.

The response generated by the model, which considers the conversation mood, is primarily text-based. To provide richer communication, we train our model to incorporate not only linguistic information but also para-linguistic cues by describing voice. Our model first generates the response text and then produces a description based on that text. To accomplish this, we introduce instruction tuning, a new process where voice descriptions are created after the language model generates responses. We also provide instructions to specify which speaker is delivering each utterance, enabling the model to respond appropriately or continue the previous utterance. More details about instruction tuning are presented in Appendix C.

4.3 Training Loss

We use a target response sentence paired with its corresponding audio description. The cross-entropy loss is then computed between the target and the model output, as illustrated in Equation 1, using the concatenation operation denoted by $\|$.

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

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 and generating paralinguistic description.

5 Experiment

5.1 Experimental Setup

In our experiments, we used two datasets: the MSenC dataset and MELD (Poria et al., 2018). We extract visual features using 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. Lastly, We utilize Mistral-7B (Jiang et al., 2023) as our LLM backbone and utilize Parler-TTS (Lyth and King, 2024) as our speech decoder. More details about implementations and computational resources are described in Appendix A.

5.2 Text Analysis

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, we aim to assess how metrics alter as we integrate information from diverse modalities into the existing large

	Ours	Ours (w.o.desc)	Tie	Ours	StyleTTS2	Tie	Ours	HierSpeech++	Tie
Emotional	**54.6%**	39.2%	6.2%	**53.4%**	35.8%	10.8%	**56.0%**	32.2%	11.8%
Suitability & Engagement	**53.4%**	41.4%	5.2%	**48.2%**	42.4%	9.4%	**54.8%**	38.2%	7.0%
Conversation Naturalism	**56.0%**	38.6%	5.4%	**43.6%**	41.2%	15.2%	**59.6%**	32.4%	8.0%

Table 3: User study results compared with text-to-speech models on the MSenC dataset.

language model. We calculate the metrics on the text response-only portion of the output, not on the concatenated voice description. This ensures that the metrics accurately reflect the quality of the text generation without being influenced by the voice description. Table 2 shows the impact of audio and video modality processed through the Q-Former. METEOR (Banerjee and Lavie, 2005) measures not just word overlap but also semantic meaning, while the ROUGE (Lin, 2004) measures coherence and flow. According to the MSenC dataset result, incorporating additional modalities enhances the quality of the responses, indicating a positive effect on multimodal understanding. Specifically, using audio, video, and text together shows the highest performance, suggesting that the responses are more appropriate, contextually accurate, and natural-sounding.

Similar results are observed with the MELD (Poria et al., 2018), where incorporating audio and video inputs also results in the highest performance. However, the scale of the score is smaller, which implies that the MSenC dataset is more suitable for the tasks we presented.

5.2.2 Qualitative Analysis

Text-based LLMs have demonstrated remarkable capabilities in generating text based solely on textual input but struggle to interpret emotional context when the same text is presented with different emotions. For instance, they might interpret "Hello, how are you?" the same way, whether spoken happily or sadly, as it lacks access to non-verbal cues like tone of voice or facial expressions. In the quantitative analysis of evaluating the 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, 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 6. The figure provides two dialogues from different scenarios,

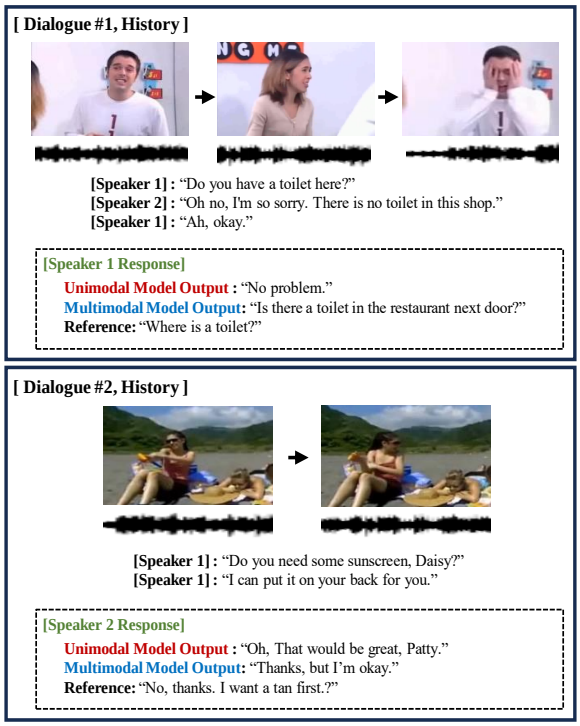

Figure 6: Qualitative evaluation on multimodality. We evaluate the results on our MSenC dataset.

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 indicate an urgent situation. In Dialogue #2, the output text adapts based on information from the video. Our model generates more natural responses and demonstrates a better understanding of the context than the unimodal model, resulting in responses that closely match the reference context.

5.3 Speech Analysis

5.3.1 User Study

We conducted a human evaluation to assess the conversational speech style generated by our system, focusing on the impact of paralinguistic information. We use Amazon Mechanical Turk and ask three questions in each comparison as described below. The assessment involved 5 judges and 100 generated samples across three criteria: emotional,

Model	Accuracy
Ours	15.10%
Ours (w.o. description)	11.20%
StyleTTS2 (Li et al., 2024)	13.72%
HierSpeech++ (Lee et al., 2023)	12.54%

Table 4: Result of evaluating emotional continuity in conversations on the MSenC dataset.

engagement & suitability, and conversation naturalism. The baselines are StyleTTS2 (Li et al., 2024), HierSpeech++ (Lee et al., 2023), and Parler-TTS (Lyth and King, 2024), which generate speech without natural language prompt. Our comparisons are made against the baseline model, which is inferred using the official checkpoints. The Speech sample was generated from the test set of the MSenC dataset. As in Table 3, our model shows superior results on every criterion, which demonstrates the effectiveness of our approach to generate expressive speech reflecting conversation mood. The template and more details can be found in Appendix D.

5.3.2 Emotion Classification

In this experiment, we classify the emotions of each utterance and then calculate the accuracy based on whether their emotions are matched or not. The process assumes if the emotion of an utterance aligns with the previous one, we consider it as empathetic. This approach helps to highlight the continuity and empathy within a dialogue. In this setup, we use the MSenC dataset and employ a pretrained speech emotion classification model[7] from HuggingFace. We performed emotion classification using one of eight emotions: angry, calm, disgust, fearful, happy, neutral, sad, and surprised. Our model generates each speech and then compares it with previous speech to assess consistency. Table 4 demonstrates that our model outperforms the baseline models in maintaining consistent emotional expression across the conversation.

5.3.3 Qualitative Analysis

In the qualitative analysis of voice description evaluation, we demonstrate our model's ability to generate speech that conveys how to say the text content. We present a comparative analysis of our model's outputs against those of the reference one in Figure 7. The figure presents two dialogues from different scenarios, showing that our model generates similar descriptions in terms of pace, pitch, and tone. This consistency leads to more contextually appropriate conversations and engaging speech responses.

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

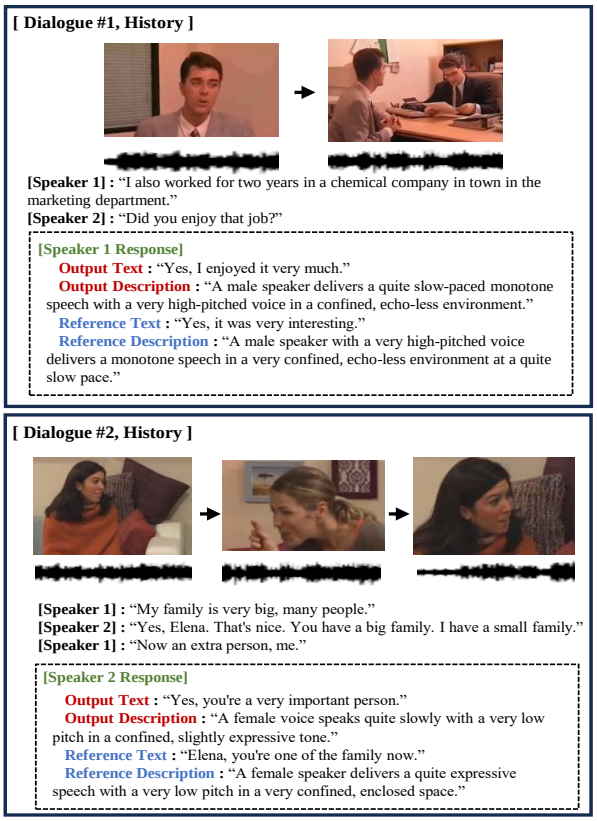

Figure 7: Qualitative evaluation of description. We evaluate the results on our MSenC dataset.

6 Conclusion

We study a dialogue model with visual and audio inputs from a speaker and generate engaging speech, 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 method 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.

7 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. The exploration of the model's performance across different languages and cultures is also limited due to the current lack of extensive and appropriate data.

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.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

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 has spent 30 hours.

B MSenC Dataset Details

In this section, we show further details of the new MSenC dataset. The histograms of video durations and word count can be found in Figure 8. Note that many videos begin with greetings such as "Hello" or "Good Morning", which contribute to a higher word count due to their conciseness. More detailed examples of the dataset can be found in Figure 9. Additionally, we evaluated gender bias within our dataset in Table 5. This involved analyzing the distribution of male and female speakers across different conversational contexts. The result shows about 1:1.5 of rate ensuring that the trained conversational system remains equitable and reliable across diverse gender groups.

C Instruction-tuning

We've introduced instruction tuning in our training process. We provide comprehensive instruction initially and give speaker ID information for each utterance. At the end, we give additional instruction for generating voice descriptions. Figure 10 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.

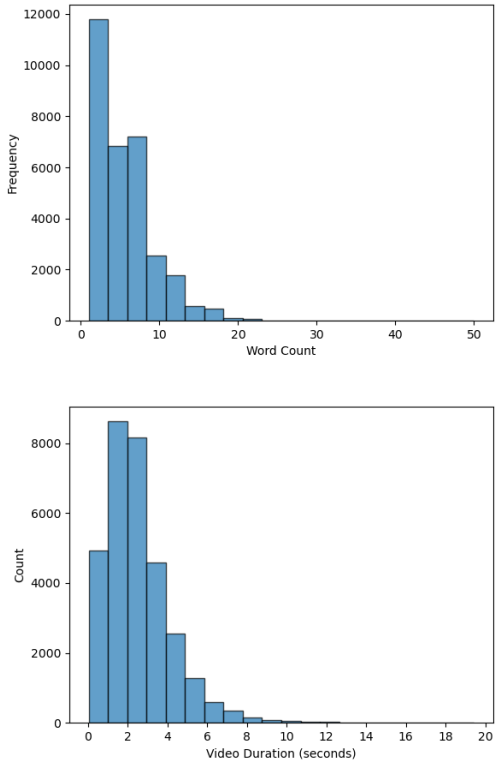

Figure 8: We report the histogram of video duration in seconds and the histogram of word count in words.

	Male	Female
Train	10,267	15,357
Validation	12,97	1,848
Test	985	1,655
Total	12,549	18,860

Table 5: This table shows gender bias within our MSenC dataset.

D Details of Human Evaluation

We present our experimental setup as follows: history is limited to a maximum of five entries, and the history is provided through video content. Participants in the experiment are presented with three response options: 'Speech 1', 'Speech 2', 'Tie'. The evaluation criteria for the responses are as follows: "emotional" assesses how well the response conveys emotions and connects with the conversation partner's feelings, while also measuring the energy, liveliness, and interactivity of the response. "suitability and engagement" evaluates how appropriately the response fits within the context and evaluates the level of active and attentive participation in the conversation. "conversation naturalism" reflects the overall smoothness of the conversation, ensuring that the interaction feels natural, effortless,

Figure 9: MSenC Dataset sample

	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 6: Result of LLM fine-tune on MSenC 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 7: Result of LLM fine-tune on MELD dataset.

and genuine. To ensure representative samples, we evaluate 100 output samples. This approach guarantees that our evaluation encompasses a diverse range of responses, contributing to the overall reliability of our findings. The template for human evaluation is provided in Figure 11.

E LLM fine-tuning

We investigated the impact of fine-tuning a large language model with parameter efficient fine-tuning at Table 6, 7. This indicates that after fine-tuning, the model exhibited enhanced conversational capabilities compared to its pre-fine-tuned state.

$$I \;\; H_1\,[utt_1] \;\; H_2[utt_2] \;\; \cdots \;\; H_n[utt_n] \;\; R$$

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 10: Example 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.

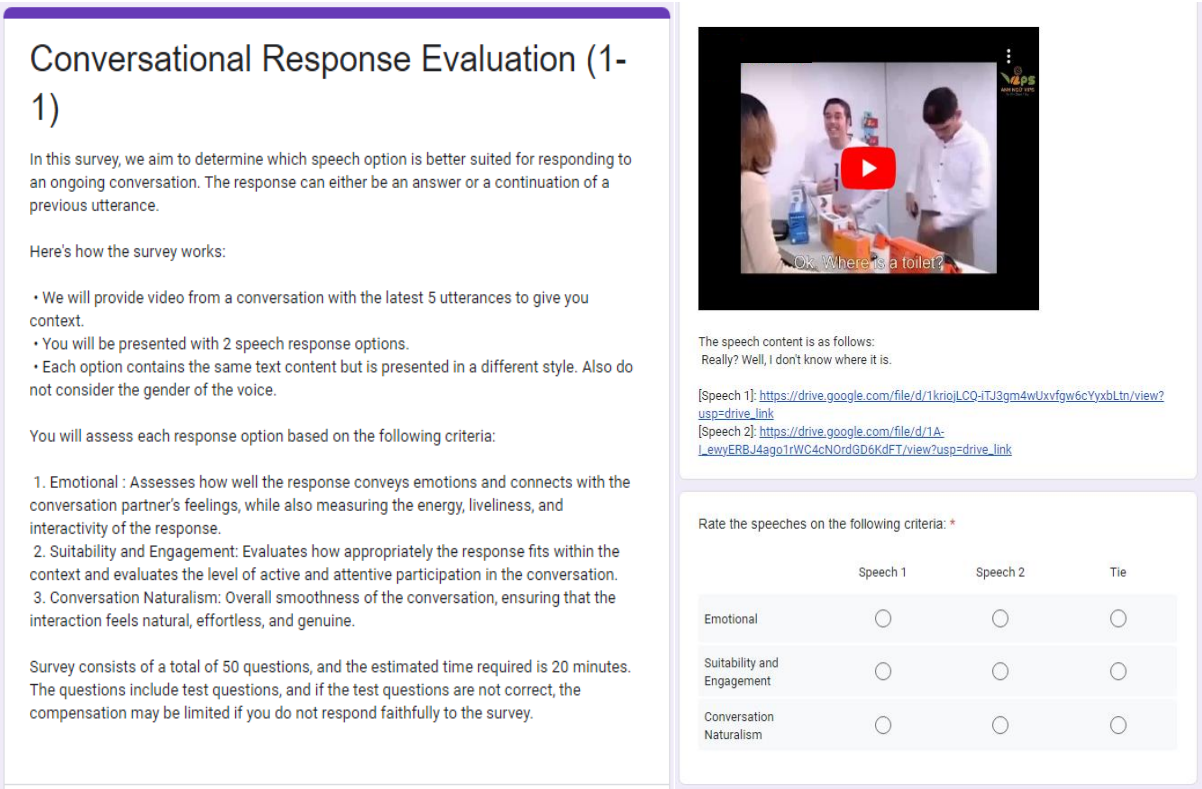

Figure 11: Human evaluation template.