Let's Go Real Talk: Spoken Dialogue Model for Face-to-Face Conversation

Anonymous ACL submission

Abstract

In this paper, we introduce a novel Face-to-Face spoken dialogue model. It processes audio-visual speech from user input and generates audio-visual speech as the response, marking the initial step towards creating an avatar chatbot system without relying on intermediate text. To this end, we newly introduce MultiDialog, the first large-scale multimodal (*i.e.*, audio and visual) spoken dialogue corpus containing 387 hours of approximately 10,000 dialogues, recorded based on the open domain dialogue dataset, TopicalChat. The MultiDialog contains parallel audio-visual recordings of conversation partners acting according to the given script with emotion annotations, which we expect to open up research opportunities in multimodal synthesis. Our Face-to-Face spoken dialogue model incorporates a textually pretrained large language model and adapts it into the audio-visual spoken dialogue domain by incorporating speech-text joint pretraining. Through extensive experiments, we validate the effectiveness of our model in facilitating a face-to-face conversation. All the data will be open-sourced.

1 Introduction

Spoken Dialogue System (SDS), often referred to as a conversational agent, engages in natural speech conversations with humans by recognizing speech from user input and providing contextually appropriate and accurate responses with speech. With spoken language as the primary interface, it has numerous applications for human-computer interactions such as customer service and voice assistants.

However, when people communicate face-to-face, we utilize not only audio but also visual information of the conversing partner to process spoken words and non-verbal cues (*i.e.*, facial expressions, gestures, and emotions) (Petridis et al., 2018; Hong et al., 2023). This multimodal information enhances understanding of the speech content and the speaker's intent. Furthermore, having a visual counterpart to audio can simulate a real face-to-face conversation experience, making the user feel more connected and engaged.

In this paper, we explore an audio-visual spoken dialogue system to facilitate direct face-to-face conversation for the first time. Central to the development of dialogue systems is the large amount of high-quality dialogue data. Current dialogue systems are predominantly text-based, driven by the abundance of text dialogue datasets (Lowe et al., 2015; Li et al., 2017; Zhang et al., 2018; Rashkin et al., 2018; Budzianowski et al., 2018; Zhou et al., 2018; Reddy et al., 2019; Lambert et al.; Ding et al., 2023; Köpf et al., 2023). Recently, several audio dialogue datasets have been released (Lee et al., 2023; Si et al., 2023; Nguyen et al., 2023a) which augment existing text dialogue data (Li et al., 2017; Budzianowski et al., 2018) with speech. However, those with visual components remain limited in scale, comprising less than 15 hours in total (Busso et al., 2008; Poria et al., 2018). Addressing this data gap, we introduce MultiDialog, the first large-scale audio-visual spoken dialogue corpus. It consists of 387 hours of audio-visual recordings of approximately 10,000 dialogues, derived from open-domain text dialogue dataset, TopicalChat (Gopalakrishnan et al., 2023) which is an extensive multi-turn dialogue corpus collected from real conversations covering 9 broad topics. The proposed MultiDialog consists of emotion annotations for each utterance and simultaneous recordings of both the listener and speaker, presenting opportunities for diverse research; from face-to-face dialogue system to talking face synthesis (Park et al., 2022; Zhang et al., 2023b), listener's face synthesis (Song et al., 2023; Zhou et al., 2023), and emotion-conditioned face synthesis (Goyal et al., 2023).

Based on the MultiDialog dataset, we propose the first audio-visual spoken dialogue model that

Dataset	# Dialogues	# Turns	Length (hrs)	Audio	Text	Video	Emotion
IEMOCAP (Busso et al., 2008)	151	10,039	12	✓	✓	✓	✓
DSTC2 (Henderson et al., 2014)	1,612	23,354	32	✓	✓	✗	✗
MELD (Poria et al., 2018)	1,433	13,000	13.7	✓	✗	✓	✓
DailyTalk (Lee et al., 2023)	2,514	23,774	21.7	✓	✓	✗	✗
Expresso (Nguyen et al., 2023a)	391	2,400	47	✓	✓	✗	✓
SpokenWOZ (Si et al., 2023)	5,700	203,074	249	✓	✓	✗	✗
MultiDialog	9,920	198,400	387	✓	✓	✓	✓

Table 1: Comparison of MultiDialog dataset with publicly available multimodal dialogue datasets.

can directly process audio-visual speech as user input and generate audio-visual speech as the output response. Motivated by the recent success of the direct spoken dialogue model using discretized speech tokens (Nguyen et al., 2023b; Zhang et al., 2023a), we introduce audio-visual (AV) speech tokens extracted by quantizing audio-visual speech features from a self-supervised model (Shi et al., 2021). Utilizing the AV speech tokens as pseudo texts, we integrate AV speech into a pretrained large-language model (LLM) (Zhang et al., 2022) through joint speech-text pretraining. The response is also returned in AV speech tokens, and it is synthesized into a talking face video for direct face-to-face interaction between the systems.

Our contributions are in three folds: (1) We introduce the first direct Face-to-Face dialogue model which processes multimodal speech from user input and generates multimodal speech as the output response, facilitating a face-to-face conversation system. (2) To build a face-to-face dialogue system, we propose the first large-scale multimodal (*i.e.*,audio, visual, and text) dialogue corpus, MultiDialog consisting of approximately 400 hours of audio-visual conversation streams. (3) We demonstrate that speech-text joint pretraining leveraging a pre-trained large language model improves upon direct initialization in retaining knowledge of the original large language model.

2 Related Work

2.1 Spoken Dialogue Dataset

In recent years, the development of speech dialogue datasets has played a pivotal role in understanding human behavior and building spoken dialogue systems that emulate real-life conversations. Early speech datasets focus on analyzing human behavior such as emotion and intent in speech, establishing the foundation for spoken dialogue systems. IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2018), comprising audio and video recordings of dialogues, are designed to study emotional dynamics in conversations. In addition to understanding emotions, DSTC2 (Henderson et al., 2014) presents telephone-based speech dialogues for dialogue state tracking to predict user's goals. Building upon datasets that study human behavior in speech, recent spoken dialogue datasets were built to model realistic dialogue systems. Expresso (Nguyen et al., 2023a) introduces speech dialogues spanning 26 expressive styles for natural speech synthesis. DailyTalk (Lee et al., 2023) and SpokenWOZ (Si et al., 2023) datasets introduce speech-text conversations for spoken dialogues. While existing works have contributed to advancing spoken conversation systems, dialogue datasets are limited in scale and solely consist of audio and text, thereby constraining the development of audio-visual spoken dialogue systems incorporating visual cues. To address these limitations, we expand the spoken dialogue in scale and to the visual modality and introduce a large-scale multimodal spoken dialogue dataset. A summary of existing multimodal dialogue datasets and MultiDialog is in Table 1.

2.2 Spoken Dialogue Models

Audio Language Model, driven by transformer-based architecture, has made remarkable strides in speech processing. By treating continuous speech as a discrete set of representations, speech can be effectively modeled as text, allowing the application of Natural Language Processing (NLP) techniques. While it has made notable progress in speech synthesis (Lakhotia et al., 2021; Borsos et al., 2023; Wang et al., 2023a; Hassid et al., 2023; Nachmani et al., 2023), speech translation (Barrault et al., 2023; Dong et al., 2023; Rubenstein et al., 2023), and speech recognition (Wang et al., 2023b), spoken dialogue system is a relatively unexplored field of research due to the scarcity of spoken dialogue datasets. Several works made an effort to tackle data issues by leveraging the power of large language models (LLMs). SpeechGPT

MultiDialog	Statistics
# dialogues	9,920
# turns	108,624
# utterances	217,248
avg # turns/dialogue	11.0
avg length/turns (s)	12.8
avg length/dialogue (s)	140.2
total length (hr)	387
# speakers	12
# dialogues/speaker	826.7

Table 2: Datailed statistics of MultiDialog

(Zhang et al., 2023a) first converts speech into discrete speech tokens, and then designs a three-stage training pipeline on paired speech data, speech instruction data, and chain-of-modality instruction data. AudioGPT (Huang et al., 2023) instructs LLMs to generate commands for controlling external tools before inputting them into the LLMs. d-GSLM (Nguyen et al., 2023b) models two-channel conversations to produce natural turn-taking conversations.

There are Multimodal Large Language Models (MM-LLM) (Wu et al., 2023; Gong et al., 2023) capable of processing both visual input and output. However, they are visual grounding dialogue systems that use visual information as supplementary for tasks such as image captioning and image editing. In contrast, we aim to build an audio-visual spoken dialogue system (*i.e.*, facial movement related to the speech) to enhance the understanding of speech content and enrich the communication experience, emulating a real face-to-face conversation.

3 MultiDialog Dataset

3.1 Preparation

To obtain audio-visual recordings of dialogues, we gathered 12 fluent English speakers, with varying gender, age, and nationality. The participants, aged 20 to 25, came from six different countries, with six female and six male actors. We derived dialogue scripts from the open-domain dialogue dataset, TopicalChat (Gopalakrishnan et al., 2023) which is a rich knowledge-grounded dataset collected from real human-human conversations. It spans eight broad topics including fashion, politics, books, sports, general entertainment, music, science & technology, and movies. It is annotated for eight emotions: Disgusted, Angry, Fearful, Happy, Sad, Surprised, Neutral, and Curious to dive deeper. The conversation partners don't have explicitly defined roles as 'speaker' or 'listener' so they interact naturally similar to how people engage in real-world conversations. Due to the topical variety, emotion annotation, and representation of natural human conversations, we chose TopicalChat as the foundation for constructing a multimodal dialogue dataset.

3.2 Recording

Data was recorded in a professional recording studio with a green screen and minimal background noise, shown in Appendix A.2. During a recording session, two conversation partners sat side-by-side and were recorded with a separate camera and a microphone. The camera position was adjusted according to the individual's height to capture the upper body, starting from the shoulders. The participants were asked to act according to a given script conveying the desired emotion annotation for each utterance. We specifically provided emotion instructions for visual cues based on the Facial Action Coding System (Ekman and Friesen, 1978) (*i.e.*, happy: cheek raiser, lip corner puller) and for audio cues based on prosody (*i.e.*, happy: high pitch, normal loudness). For recordings, we combined the emotion labels 'Neutral' and 'Curious to dive deeper' into a single label 'Neutral' due to the lack of visually apparent difference between the two. Moreover, when the turn passes to another participant, they naturally react while listening. Participants were instructed to press a button to proceed to the next utterance, which recorded the start and end times of each turn for post-processing. The audio streams were recorded in a mono WAV format at 48kHz and the video streams in full HD at 30fps.

3.3 Post-Processing

To refine the data, we had an annotator go through the audio-visual recordings to check if there were any misalignments between the audio and visual streams. We asked the annotator to manually adjust the misalignments by sliding the start time. Additionally, we filtered out recordings without either audio or visual streams. Then, we segmented the recordings into conversations and turns based on the recorded timesteps of each turn. The MultiDialog dataset consists of approximately 400 hours of audio-visual videos of 10,000 dialogues between 6 pairs of conversation partners. The final statistics of our dataset are shown in Table 2.

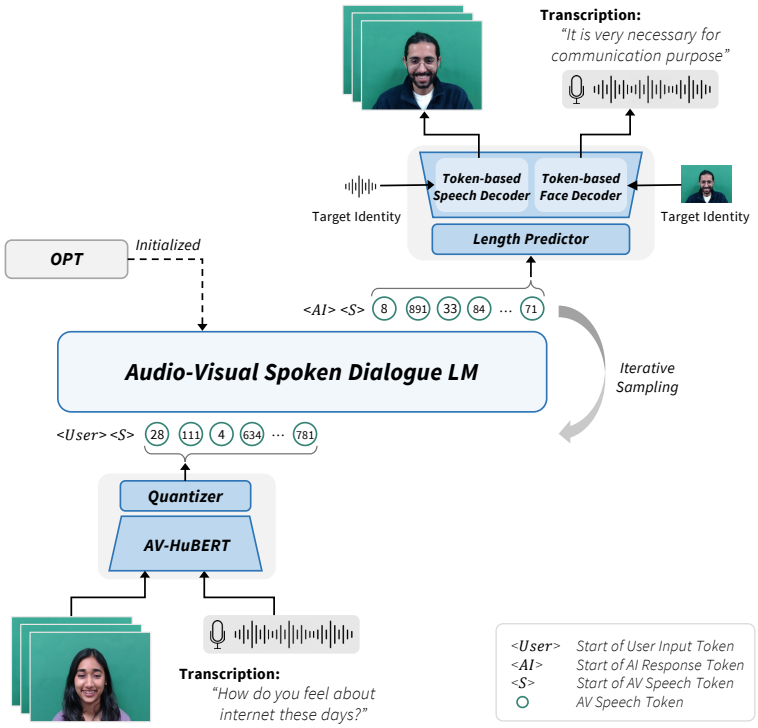

Figure 1: Overview of the proposed framework for multimodal spoken dialogue language modeling. With the AV speech tokens as the pseudo-texts, it can process audio-visual face video from the user input and generate corresponding response as audio-visual face video.

4 Audio-Visual Spoken Dialogue System

Based on the proposed MultiDialog dataset, we introduce an audio-visual spoken dialogue system that directly understands the audio-visual of the user's face video and generates appropriate responses with audio-visual face video. It consists of three main parts: 1) Encoding audio-visual speech into discrete representations, namely audio-visual (AV) speech tokens. 2) Conducting multimodal spoken dialogue language modeling using the AV speech tokens as pseudo texts. 3) Projecting the output AV speech tokens into the audio and visual space for direct face-to-face dialogue.

4.1 Audio-Visual Speech Encoding

By integrating both audio and visual modalities, we can improve the dialogue system's understanding of the speech content. This is because speech not only comprises auditory signals but also visual cues from the movements of the speaker's mouth. This visual information complements auditory signals, particularly in noisy environments, resulting in more robust performance (Afouras et al., 2018).

To this end, we adopt a unified approach to model both the audio and visual of talking face input into audio-visual speech tokens. Motivated by the recent success of utilizing discrete speech tokens extracted from self-supervised speech models (Schneider et al., 2019; Baevski et al., 2020; Hsu et al., 2021; Chung et al., 2021; Babu et al., 2021) in speech processing (Lakhotia et al., 2021; Lee et al., 2021; Maiti et al., 2023; Kim et al., 2023), we tokenize the audio and visual streams into audio-visual speech tokens (*a.k.a.* AV speech tokens). Specifically, we employ one of the multimodal speech models, AV-HuBERT (Shi et al., 2021), a state-of-the-art self-supervised framework for understanding speech by both seeing and hearing. It is trained on raw audio-visual face videos to predict discrete clusters from speech (Hassid et al., 2023). The audio-visual speech features are extracted and quantized into discrete tokens as in (Lakhotia et al., 2021; Popuri et al., 2022; Kim et al., 2024). By combining the visual cues and the auditory information, the audio-visual speech tokens extract both linguistic and phonetic information. Then, we treat the AV speech tokens as pseudo text to train our Audio-Visual Spoken Dialogue LM.

4.2 Audio-Visual Spoken Dialogue Language Modeling

As shown in Fig. 1, our audio-visual spoken dialogue language model is trained with the AV speech tokens on our MultiDialog dataset. Previous work

(Hassid et al., 2023) showed that initializing a speech language model with a textually pretrained language model (LLM) leads to better performance and faster convergence. Accordingly, we use a pretrained LLM, OPT-1.3B (Zhang et al., 2022) to initialize our model and combine the vocabulary of AV speech tokens with the original text vocabulary, as in (Zhang et al., 2023a; Nachmani et al., 2023; Maiti et al., 2023). This allows us to jointly model the probability of both AV speech tokens and text tokens t, where the loss can be represented as,

$$\mathcal{L} = -\sum_{i=1}^{N} \log p(t_i \mid t_1, ..., t_{i-1}), \qquad (1)$$

which is the negative log-likelihood of predicting the next token in the sequence of length N tokens.

Motivated by the joint speech-text training used in speech processing tasks such as speech translation, audio speech recognition, and text-to-speech synthesis (Cheng et al., 2023; Maiti et al., 2023; Dong et al., 2023; Wang et al., 2023b), we newly introduce a joint speech-text pre-training scheme tailored for spoken dialogue language modeling. In our setting, each dialogue $D = [T_1^{ai}, T_1^{user}, T_2^{ai}, T_2^{user}, \ldots, T_k^{ai}, T_k^{user}]$ consists of k rounds of turns T between two speakers which we randomly designate as the AI and the User. The goal of this pre-training is to effectively transform the text-based LLM into the AV speech token-based LLM, enabling it to produce relevant AV speech responses from the AI side given a conversation context. It proceeds in the following two stages:

The first stage is instructing the LLM to interpret and generate AV speech tokens. We segment the dialogue into turns T and prepare paired AV speech tokens and text tokens. We then concatenate the pair with their respective modality prefix tokens, <speech> and <text>, to indicate the beginning of text and AV speech tokens. Adding the reversed order of concatenation, we construct both audio-visual speech recognition (AVSR) and text-to-speech generation (TTS) training objectives as shown in Fig. 2(a) and (b). Only the embedding layer and the projection layer are trained in this first stage, which guides the LLM to understand and generate AV speech tokens while fully retaining the given LLM knowledge needed for dialogue generation.

The second stage is jointly learning the text and AV speech token-based dialogue. We select either one of the speakers as the AI which the model aims to predict and indicate the start of the response with additional speaker prefix tokens, <User> and <AI>. The Speaker prefix token is followed by modality prefix tokens, <Speech> and <Text>, to indicate whether the utterance is in AV speech or text as shown in Fig. 2(c). During the pretraining, we evenly mix the use of AV speech tokens and text which allows the model to utilize both token knowledge to generate dialogue response. We pretrain the entire model at this stage and we later finetune on pure AV speech token-based dialogue as in Fig. 2(d).

(a) AV Speech to Text Token (AVSR)

• <Speech> 7 495 123 495 7 21 495 21 495 ... 118
<Text> Are you a football fan?

(b) Text to AV Speech Token (TTS)

• <Text> I know. I think because the league prefers offense over defense nowadays. <Speech> 7 381 17 338 21 123 57 123 329 ... 11

(c) Mixed Text and AV Speech Token Dialogue

• <User> <Text> do you like drama? <AI> <Speech> 7 278 123 7 21 278 123 21 7 ... 212 <User> <Text> Yeah. Do you like Jim Carrey? <AI> <Speech> 7 278 123 278 123 278 161 7 406 ... 2 <User> <Text> Yes as they do not project positivity <AI> <Text> Yes. Do you listen to radio dramas?

(d) AV Speech Token Dialogue

• <User> <Text> 7 123 381 123 7 402 437 21 413 ... 38
<AI> <Speech> 7 278 123 7 21 278 123 21 7 ... 212
<User> <Text> 7 445 123 7 123 329 57 437 161 ... 2
<AI> <Speech> 7 278 123 278 123 278 161 7 406 ... 2
<User> <Text> 7 17 7 437 23 437 7 437 161 ... 225
<AI> <Text> 7 278 123 278 123 21 7 437 380 ... 385

Figure 2: Constructed data based on the MultiDialog dataset used for training the audio-visual speech dialogue model. (a-c) are joint pretraining of the audio-visual speech and text tokens and (d) is used to finetune the model.

4.3 Audio-Visual Generation

The generated AV speech tokens are projected to audio and visual to generate the response as a talking face video. As shown in Fig. 1, the audio-visual generator consists of a length predictor, a token-based speech decoder, and a token-based face decoder. Since our language model is trained with duplicate reduced AV speech tokens, we train a length predictor to first restore them back to their original length. The token-based speech decoder and token-based face decoder are adapted from an off-the-shelf audio generator (Kong et al., 2020) and a talking face generator (Prajwal et al., 2020) respectively, where we train them to process AV speech tokens as the input instead of raw audio. Additionally, we incorporate speaker identity information by extracting the speaker embedding (Jia

Evaluation Prompt

- <User> Hi, how are you doing today? <AI>
- <User> Hi, how are you doing today? <AI> I am good, thanks. I'm wondering if you are a football fan like me? I know a lot of people are sad the season is over, especially since fantasy football is such a hit. <User> I like football, but I don't get much chance to watch. usually I try to catch the SB. <AI>
- <User> Hi, how are you doing today? <AI> I am good, thanks. I'm wondering if you are a football fan like me? I know a lot of people are sad the season is over, especially since fantasy football is such a hit. <User> I like football, but I don't get much chance to watch. usually I try to catch the SB. <AI> What is the SB? I know if fantasy football it's all about the picks and people plan for weeks. <User> Super Bowl. Sorry, should have made that clear. I've never played fantasy sports. What's the goal? How do you win? What do you win? <AI>

Figure 3: Evaluation prompt of multimodal dialogue language modeling. It is written in text for illustration but the actual prompt is given as audio and visual.

et al., 2018) from a target identity sample audio. Also, the target identity's face and pose prior are utilized as in (Prajwal et al., 2020), to enable the generation of talking face video with desired identity.

5 Experimental Setup

5.1 Evaluation Metrics

We evaluate the semantic quality and the generation quality of both audio and video. For the semantic quality, we first generate transcriptions from the synthesized audio-visual output using an off-the-shelf ASR model (Shi et al., 2021), and employ standard metrics used for text-based dialogue generation: log-perplexity (PPL), BLEU, METEOR, F1, D-1, and D-2. The log-perplexity is calculated using Dialo-GPT model (Zhang et al., 2019) and it is calculated for each utterance and averaged across the test set. To measure the generation quality of video, we adopt metrics used for TFG. This includes Fréchet Inception Distance (FID) (Heusel et al., 2017) to measure visual quality, and LSE-C and LSE-D (Prajwal et al., 2020) to measure the audio-visual synchronization. To evaluate the acoustic quality, we compute speaker similarity (SIM) between the given target sample and generated speech using the WavLM-Base model for speaker verification (Chen et al., 2021). Please refer to the appendices for a detailed explanation of each metric.

5.2 Implementation Details

To encode AV speech tokens, we crop the video into the mouth region of size 96×96 using a face detector (Deng et al., 2020) and a facial landmark detector (Bulat and Tzimiropoulos, 2017), and resample the audio to 16kHz. We take English-trained AV-HuBERT (Shi et al., 2021) and finetune it to predict corresponding target clusters from HuBERT tokenizer (Hassid et al., 2023) which operates at 25Hz with 500 clusters. We train it for 100k steps on 6 A6000 GPUs with a maximum token length of 2,000.

We initialize the model with a pre-trained language model, OPT-1.3B (Zhang et al., 2022). We first pretrain the input embedding layer and the projection layer on AVSR and TTS objectives for 200K steps. Then, we continue training the entire model on a mixture of text and AV speech token dialogue for 5K steps, followed by finetuning for additional 3K steps on AV speech token dialogue only. We use a max token length of 700 on 4 A6000 GPUs.

The audio-visual generator is trained using ground truth AV speech tokens. The token-based speech decoder and length predictor are jointly trained for 450K steps with a batch size of 32. For training the token-based face decoder, we employ the reprogramming strategy in (Choi et al., 2023) and train an adapter layer consisting of two layers of transformer encoder to bridge between the AV speech tokens and the corresponding audio features of the TFG model (Prajwal et al., 2020). It is trained for 250K steps with a batch size of 256. We additionally incorporate a face enhancer (Wang et al., 2021) to upsample the generated face video into high resolution.

5.3 Baselines

Since there is no previous method that can directly perform audio-visual spoken dialogue synthesis, we compare with the recently proposed spoken dialogue systems, Speech-GPT (Zhang et al., 2023a) and d-GSLM (Nguyen et al., 2023b). They support only audio speech at both input and output. Additionally, we build a cascade system by integrating a series of off-the-shelf pre-trained models: AVSR (Anwar et al., 2023), LM (Tang et al., 2022), TTS (Casanova et al., 2022), and TFG (Prajwal et al., 2020). Please note the objective of the comparisons with the cascaded method is not to achieve state-of-the-art performance, but rather to assess the extent to which the performance of the proposed system can be attained through the direct strategy. For a fair comparison, we finetune SpeechGPT and d-GSLM on our MultiDialog dataset and we use a dialogue language model (Tang et al., 2022) trained on TopicalChat as the LM of the cascade system.

Method	Input Modality	Output Modality	Semantic Evaluation					
			PPL ↓	BLEU ↑	METEOR ↑	F1 ↑	D-1 ↑	D-2 ↑
• ***Ground Truth***								
GT AV Speech Token	–	–	1054.643	76.326	0.565	0.474	0.947	0.996
• ***Cascaded System***								
AVSR + LM + TTS + TFG	AV	AV	1157.586	47.287	0.075	0.100	0.959	0.977
• ***Spoken Dialogue System***								
SpeechGPT (Zhang et al., 2023a)	A	A	930.401	20.536	0.0640	0.0542	0.743	0.876
d-GSLM (Nguyen et al., 2023b)	A	A	1085.265	8.197	0.065	0.064	0.883	0.876
• ***Audio-Visual Spoken Dialogue System***								
Scratch	AV	AV	1898.864	13.305	0.058	0.064	0.945	0.955
+ LLM initialized	AV	AV	1237.757	17.098	0.059	0.058	0.936	0.963
+ AVSR/TTS Pretraining	AV	AV	1068.904	22.090	0.062	0.066	0.943	0.965
+ Mixed Text-AV Speech Pretraining	AV	AV	1248.001	24.094	0.063	0.065	0.945	0.957

Table 3: Comparison of the semantic quality between state-of-the-art spoken dialogue systems. Note that our proposed method is the only method that supports both audio and visual at the input and output of the dialogue system without relying on intermediate text.

6 Results

6.1 Semantic Evaluation

To accurately assess the semantic quality of the generated response, we employ the evaluation strategy used for text-based dialogue language models. We conduct evaluations on the test set of MultiDialog, where the model is prompted to sequentially generate a response for each turn in the conversations. Sample evaluation prompts are illustrated in Figure 3. The generated response is then transcribed into text and compared against the ground truth response to evaluate its semantic quality. As shown in Table 3, compared with the state-of-the-art spoken dialogue systems, SpeechGPT (Zhang et al., 2023a) and d-GSLM (Nguyen et al., 2023b), our proposed method performs the best in BLEU, D-1, and D-2 which demonstrates that our method can generate contextually coherent and diverse response. SpeechGPT has the highest PPL because it is trained on an extensive amount of speech data and PEFT-finetuned (Hu et al., 2021) on the MultiDialog, which allows it to generate more fluent speech but fails to match with the reference response as indicated by the lower BLEU score. Also, it requires generating text transcription of the input to generate the response in text first. Notably, our proposed method stands as the first approach to directly recognize and generate response in audio-visual speech video, without requiring intermediate text generation.

6.2 Ablation on the Pretraining Scheme

We analyze the pretraining scheme used for our audio-visual spoken dialogue model in the lower section of Table 3. The results demonstrate that initializing the model with a textually pretrained LLM yields improved semantic quality, which is further enhanced by AVSR/TTS pretraining. Simply training the embedding layer and projection layer to predict corresponding AV speech tokens and text tokens improves the response. However, when incorporating mixed text-AV speech token pretraining, we observe an overall enhancement in semantic quality, albeit with a slight decrease in the PPL score, which we attribute to the model's increased complexity and adaptability to multimodal inputs.

Method	FID ↑	LSE-C ↑	LSE-D ↓	SIM ↑
• ***Cascade System***				
AVSR + LM + TTS + TFG	30.581	7.041	7.640	0.433
• ***Spoken Dialogue System***				
SpeechGPT (Zhang et al., 2023a)	-	-	-	0.194
d-GSLM (Nguyen et al., 2023b)	-	-	-	0.211
• ***Audio-Visual Spoken Dialogue System***				
Proposed	30.323	7.298	7.390	0.624

Table 4: Evaluation of the audio and visual generation quality. Note that we evaluate the reconstructed audio and visual output of selected 300 videos from the test set.

6.3 Audio and Visual Evaluation

We evaluate the audio and visual generation quality in Table 4. Our token-based speech decoder, enriched with speaker embedding, achieves the highest speaker similarity score (SIM). When assessing visual quality, we compared it with the cascaded system, which generates audio-visual videos from TFG (Prajwal et al., 2020). While our FID score is comparable, our approach exhibits superior audio-visual synchronization, due to the utilization of dis-

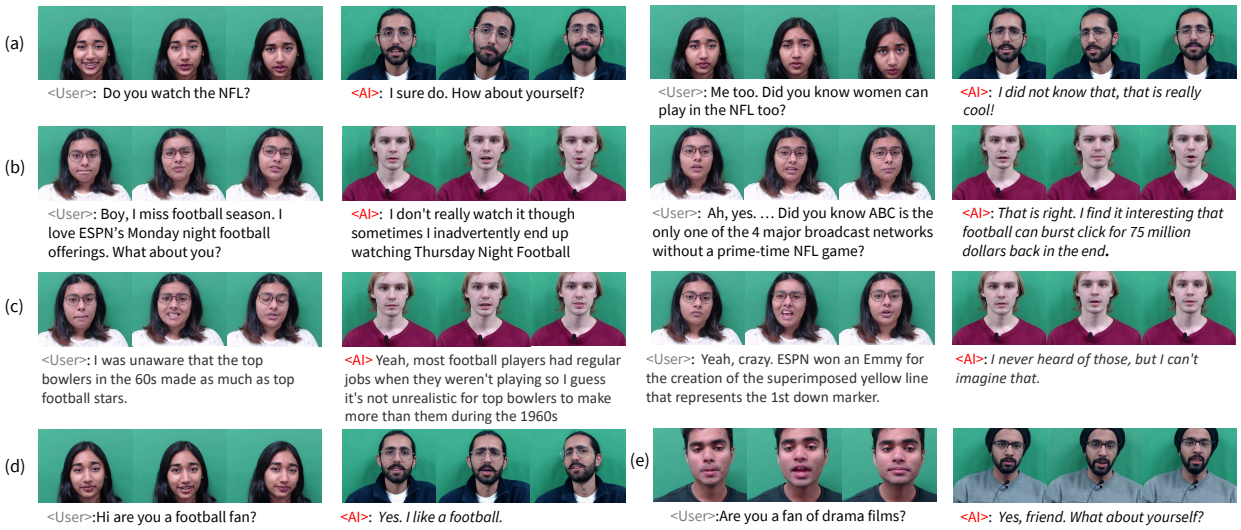

Figure 4: Audio-visual dialogue generation results of the proposed method, where the last turn is the generated audio-visual response. Note that we have randomly sampled three video frames from each turn for illustration. (a-c) are conversations with four turns and (d-e) are with two turns, The generated response is in italics and we provide ASR transcriptions below.

Method	Input Modality	SNR (dB)			
		-5	0	5	clean
Proposed	A	11.340	14.751	21.143	23.089
	AV	13.853	18.144	21.186	24.094

Table 5: Dialogue response generation performance (BLEU) with different input modalities under acoustic noise corruption with different SNR levels (dB).

cretized audio-visual tokens, which provide clearer alignment between the audio and visual components.

In Figure 5, we show the generated audio-visual response between the two partners along with transcriptions generated with ASR (Shi et al., 2021). Given a conversation context, our model generates the next response that is contextually coherent and adequate. For example, in Figure 5 (a), it answers the question asked by the user in the previous turn and responds accordingly about the chatting topic, NFL. Please refer to the demo for more demonstrations of the generated response.

6.4 Robustness to Acoustic Noise

In Table 5, we analyze the effectiveness of incorporating additional visual modality into the dialogue system. Following (Shi et al., 2021), we corrupt the input speech with random noise of varying SNR levels. Compared with audio-only input, audio-visual input enhances the robustness of the system as indicated by less degradation of the performance under noisy conditions. It further demonstrates that our system is applicable for real use in unstable speech input scenario.

7 Conclusion and Limitation

We introduce a novel Face-to-Face spoken dialogue model that directly processes audio-visual speech from the user input and generates audio-visual speech response. This is the first step toward creating a talking face avatar chatbot system, without intermediate text in the generation process. In addition, we release MultiDialog, the largest multimodal dialogue dataset to date with tri-modality (*i.e.*, audio, visual, and text) spoken dialogue data. As it is an extensive dataset that captures real human-human conversation covering broad topics, we believe it brings diverse research opportunities for multimodal synthesis. The limitation of our work is that although our dataset provides emotion labels for each utterance, we have not yet made use of them in this work. We can further incorporate emotion knowledge by recognizing the emotion from the user's face to generate more emotion-aware response which can be reflected in the speech content and facial expression. Also, since our data provides parallel recordings of the speaker and the listener, we can simultaneously model the generation of both faces to enable smooth and continuous conversation.

References

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727.

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *arXiv preprint arXiv:2303.00628*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

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.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing.

Yong Cheng, Yu Zhang, Melvin Johnson, Wolfgang Macherey, and Ankur Bapna. 2023. Mu^{2}slam: Multitask, multilingual speech and language models. In *International Conference on Machine Learning*, pages 5504–5520. PMLR.

Jeongsoo Choi, Minsu Kim, Se Jin Park, and Yong Man Ro. 2023. Reprogramming audio-driven talking face synthesis into text-driven. *arXiv preprint arXiv:2306.16003*.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Qianqian Dong, Zhiying Huang, Chen Xu, Yunlong Zhao, Kexin Wang, Xuxin Cheng, Tom Ko, Qiao Tian, Tang Li, Fengpeng Yue, et al. 2023. Polyvoice: Language models for speech to speech translation. *arXiv preprint arXiv:2306.02982*.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A

vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.

Sahil Goyal, Sarthak Bhagat, Shagun Uppal, Hitkul Jangra, Yi Yu, Yifang Yin, and Rajiv Ratn Shah. 2023. Emotionally enhanced talking face generation. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, pages 81–90.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. 2023. Textually pretrained speech language models. *arXiv preprint arXiv:2305.13009*.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18783–18794.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

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, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31.

Minsu Kim, Jeongsoo Choi, Dahun Kim, and Yong Man Ro. 2023. Many-to-many spoken language translation via unified speech and text representation learning with unit-to-unit translation. *arXiv preprint arXiv:2308.01831*.

Minsu Kim, Jeong Hun Yeo, Jeongsoo Choi, Se Jin Park, and Yong Man Ro. 2024. Multilingual visual speech recognition with a single model by learning with discrete visual speech units. *arXiv preprint arXiv:2401.09802*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Nathan Lambert, Nazneen Rajani Lewis Tunstall, and Tristan Thrush. Huggingface h4 stack exchange preference dataset. 2023. *URL https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences*.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. 2021. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2023. Voxtlm: unified decoder-only models for consolidating speech recognition/synthesis and speech/text continuation tasks. *arXiv preprint arXiv:2309.07937*.

Eliya Nachmani, Alon Levkovitch, Julian Salazar, Chulayutsh Asawaroengchai, Soroosh Mariooryad, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2023. Lms with a voice: Spoken language modeling beyond speech tokens. *arXiv preprint arXiv:2305.15255*.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. 2023a. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023b. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. 2022. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2062–2070.

Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018. End-to-end audiovisual speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6548–6552. IEEE.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. In *Proc. Interspeech*.

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*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2021. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Luchuan Song, Guojun Yin, Zhenchao Jin, Xiaoyi Dong, and Chenliang Xu. 2023. Emotional listener portrait: Realistic listener motion simulation in conversation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20782–20792. IEEE.

Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Mvp: Multi-task supervised pre-training for natural language generation. *arXiv preprint arXiv:2206.12131*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023b. Viola: Unified codec language models for speech recognition, synthesis, and translation. *arXiv preprint arXiv:2305.16107*.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023b. 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.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.

Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, and Tiejun Zhao. 2023. Interactive conversational head generation. *arXiv preprint arXiv:2307.02090*.

A MultiDialog Dataset

A.1 Participant Recruitment

Prior to recording our dataset, we received an IRB approval to collect facial video, speech, and text data to build human multimodal dialogue technology. We recruited students at a university who were fluent in English and could fulfill the designated portion of the dialogues. A recruitment notice included general information about TopicalChat, the dataset to be recorded, wage and responsibilities of the participants, and potential effects and contributions of building a multimodal dialogue dataset. After receiving 25 applications, interviews were conducted on all applicants. During the interview, we notified that we will be collecting audiovisual data of the participant during recording sessions, which will be released to the research field in the future. We also collected participant information such as race, sex, nationality and age, agreement to and assessed the English fluency, ability to read and act out a given dialogue script with emotions, and time availability of each participant. Two interviewees in charge of the dataset collection selected actors by ranking each participant on a scale of 1 to 5 on each criterion and considering the diversity of participant demographic. Thus, six female and six male actors from six different countries, and age varying from 20 to 25 were selected.

After all participants were selected, we held an orientation to guide participants on the recording procedure. For a single recording session of three hours, two participants were scheduled to film 50 to 60 conversations in TopicalChat. The number of conversations to film in a session was calculated based on a trial recording session, in which two speakers filmed approximately 60 conversations in a three-hour period, including breaks. Participants learned how to navigate through the dialogue display program to start and end recording conversations, and proceed to the next utterance. The display program showed the conversation script along with the corresponding emotion for each utterance, and the remaining number of conversations to film in the current session. We notified each participant to attach a microphone about 15 to 20 cm from their mouth and adjust the camera to the shoulder level before recording. Lastly, we collected consent forms for providing personal information for compensation and informed consent forms for human subject research participants.

A.2 Recording Setup

Fig. 5 shows the studio setup for recording sessions.

Figure 5: Recording studio setup for MultiDialog dataset

A.3 Dataset Statistics

Table 2 shows detailed statistics of MultiDialog. MultiDialog consists of 9,920 human-human conversations, 106,624 turns, 218,248 utterances, totalling to approximately 387 hours of audiovisual dialogue data. A single dialogue contains multiple turns, where each turn includes two utterances. An utterance is an instance of speech by one person followed by silence or another person speaking. In our dataset, a conversation averaged 11.0 turns, 21.9 utterances, 140.2 seconds in length. 12 speakers were paired to record an average of 826.7 dialogues per person.

B Evaluation Metrics

BLEU (Post, 2018) evaluates the fluency and adequacy of generated responses based on n-gram overlap. A higher BLEU score indicates a more natural and engaging dialogue model.
PPL (Bengio et al., 2000) measures how well a language model predicts the generated response. A lower perplexity indicates that the model is more confident and accurate in predicting the next word, suggesting higher quality in generating coherent and contextually relevant responses.
DISTINCT-n (Li et al., 2016) evaluates the diversity of generated response by calculating the percentage of unique n-grams in the set of responses. Specifically, D-1 measures the percentage of unique unigrams in the generated text, while D-2 measures the percentage of unique bigrams.
METEOR (Banerjee and Lavie, 2005) (Metric for Evaluation of Translation with Explicit Ordering) evaluates the quality of generated response by computing the alignment-based precision and recall between the generated output and the ground truth, considering synonyms and paraphrases.
F1 (Banerjee and Lavie, 2005) combines the accuracy of the generated response (precision) and the coverage of the relevant response (recall). It provides a balanced measure of how well the model performs in generating relevant and accurate responses.

C Generation Results

Please refer to the submitted demo video for more generation results.