

---

# OmniResponse: Online Multimodal Conversational Response Generation in Dyadic Interactions

## – Appendix

---

Cheng Luo<sup>1</sup>, Jianghui Wang<sup>1</sup>, Bing Li<sup>1\*</sup>, Siyang Song<sup>2</sup>, Bernard Ghanem<sup>1</sup>

<sup>1</sup>King Abdullah University of Science and Technology, <sup>2</sup>University of Exeter

### A Implementation Details and Hyperparameters

Table 1: Implementation details and hyperparameters.

| Setup                                |   |
|--------------------------------------|---|
| Batch Size                           | 1   |
| Training Epoch for the Unified Stage | 1500  |
| Training Epoch for A/V finetuning    | 500   |
| Warmup Epoch                         | 100   |
| Large Language Model                 | Phi-3.5 Mini-Instruct [1] (3.8B)                          |
| Text Tokenizer                       | Phi-3.5 Mini-Instruct [1] Tokenizer                       |
| Audio Tokenizer                      | Spark-tts [20] BiCodec                                    |
| Facial Coefficients                  | MediaPipe [15] facial blendshapes + transformation matrix |
| Lora Rank                            | 64  |
| Lora Alpha                           | 16  |
| Optimizer                            | AdamW   |
| Learning Rate                        | $2.0 \times 10^{-5}$                                      |
| Model Parameters $N_{\text{param}}$  | 4.5B  |
| $\beta_1$                            | 0.9   |
| $\beta_2$                            | 0.999   |
| $\lambda_{\text{vision}}$            | 1.0   |
| $\lambda_{\text{audio}}$             | 100   |

Tab. 1 summarizes the key hyperparameters used in our experiments. For the core architecture, we employ the Phi-3.5 Mini-Instruct large language model [1] for multimodal fusion and dialogue reasoning. Input modalities are processed as follows: the audio waveform is tokenized into discrete representations using the BiCodec component of Spark-tts [20], while text is tokenized using the Phi-3.5 Mini-Instruct tokenizer, augmented with special tokens such as [PAUSE] and [LASTING]. Visual features are extracted using the widely adopted MediaPipe toolkit [15], yielding 52-dimensional facial blendshape coefficients to capture local facial movements and a 12-dimensional transformation matrix representing head pose dynamics.

Model optimization is performed using the AdamW optimizer [11], with an initial learning rate of  $2.0 \times 10^{-5}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and a weight decay of  $1.0 \times 10^{-4}$ . The batch size is set to 1, and a cosine learning rate scheduler is applied throughout training. The model is first trained end-to-end—including all components (*i.e.*, LLM, vision projection, decoder, and TempoVoice) for 1,500 epochs, with a 100-epoch warmup phase. To enable efficient adaptation of the large language model, we employ the LoRA fine-tuning strategy [10] (rank 64, alpha 16), while all other parameters of OmniResponse are jointly optimized. Subsequently, a dedicated fine-tuning stage is performed for the audio and visual components (*i.e.*, vision projection, decoder, and TempoVoice) over an additional 500 epochs.

---

\*Corresponding author.

## B Methodological Details

In this section, we provide a comprehensive overview of OmniResponse, highlighting its architectural design and the key technical innovations, namely, Chrono-Text Markup and TempoVoice.

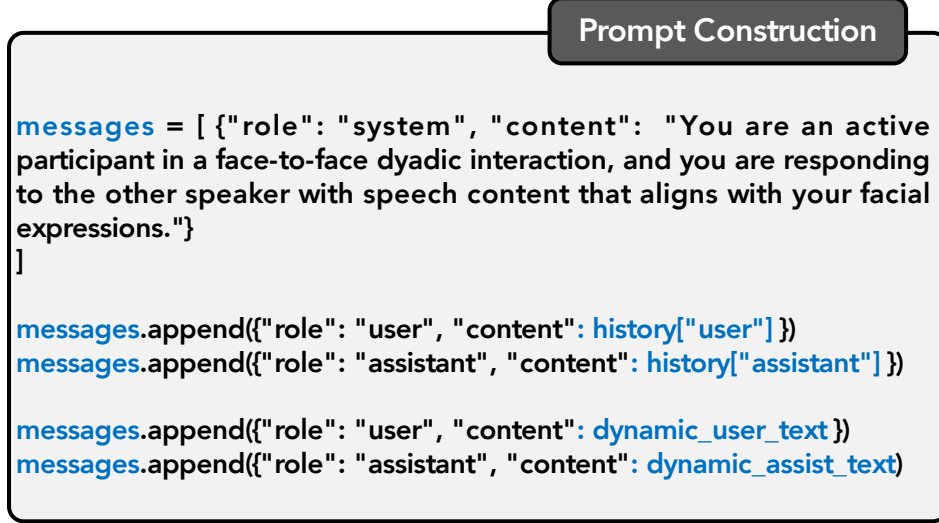


Figure 1: **Illustration of Prompt Construction.** The final prompt (messgae) is composed of a system prompt, the conversation history from the speaker (user) and listener (assistant), and dynamic speaker/listener text processed by our Chrono-Text Markup module.

### B.1 Network Architecture

OmniResponse is composed of several interconnected modules: a vision projection layer for encoding the visual frames of both the speaker and listener, a large language model [1] for fusing visual features, textual instructions, and conversational history, and a Chrono-Text Markup module for temporal alignment of text tokens. The model jointly predicts the next visual token, text token, and audio response token. A vision decoder layer reconstructs the listener’s visual frame from the predicted visual token, while the TempoVoice module converts textual embeddings into audio waveforms.

**Vision Projection Layer.** The Vision Projection Layer, denoted as  $M_{\text{vis-proj}}(\cdot)$ , encodes the previously predicted visual frames of the listener  $\hat{\mathbf{F}}_{\tau:t-1}^l$  together with the speaker’s visual frames  $\mathbf{F}_{\tau:t-1}^s$ , and projects them into a sequence of embedding features  $\mathbf{V}_{\tau:t-1}$  over the temporal interval  $[\tau, t-1]$ . Here,  $\tau$  is the starting index of the considered time window, which limits the number of temporal visual tokens and reduces computational overhead.

The process is formulated as follows:

$$\mathbf{V}_{\tau:t-1} = M_{\text{vis-proj}}(\hat{\mathbf{F}}_{\tau:t-1}^l, \mathbf{F}_{\tau:t-1}^s) \quad (1)$$

The projection module  $M_{\text{vis-proj}}$  can be instantiated either as a multilayer perceptron that processes the concatenated visual features of the speaker and listener:  $[\hat{\mathbf{F}}_{\tau:t-1}^l, \mathbf{F}_{\tau:t-1}^s]$  (where  $[\cdot]$  denotes concatenation), or as a transformer-based layer, where the listener’s visual features serve as queries, and the speaker’s visual features act as keys and values within a cross-attention mechanism.

This architecture enables effective temporal fusion of visual information from both conversational participants, providing context for subsequent response generation.

**Vision Decoder.** The vision decoder consists of a two-layer Transformer Decoder that processes the predicted embeddings  $\hat{\mathbf{V}}_{\tau+1:t}^l$  generated by the large language model for the first  $t-\tau$  positions, and maps them to the facial coefficient space  $\hat{\mathbf{F}}_{\tau+1:t}^l$ .

Subsequently, a pre-trained visual renderer converts these facial coefficients into 2D facial frames, conditioned on a given portrait image. The renderer is trained on a large-scale web video dataset and

is utilized as a tool to synthesize photorealistic images by mapping the predicted facial expression and head pose coefficients to high-quality 2D visuals.

**Static Text.** The large language model accepts both visual and textual inputs. The textual inputs include static text. Specifically, static text contains the instruction prompt  $W_{\text{instruct}}$  and the conversation history  $W_{\text{history}, < \tau}$ . The construction process for the instruction prompt is illustrated in Figure 1. The final prompt comprises the static system message (serving as the assistant’s instruction) and the conversation history between the speaker (user) and the listener (assistant) up to time  $\tau$ . This static text is provided to the LLM following the visual coefficients.

## B.2 Chrono-Text Markup

In addition to the static instruction and conversation history, we also supply the model with dynamic text annotated with precise timing information. Figure 1 illustrates how we interleave static and dynamic text when constructing each prompt: the static text preserves long-term context, while the dynamic text encodes exactly when each word occurs and how long silences last.

To achieve this, Chrono-Text Markup introduces two special tokens, [PAUSE] and [LASTING], into the token stream according to the timestamps in our dataset. At each frame timestamp: If neither the speaker nor listener is uttering a word, we insert a [PAUSE] token; When speech is present, we emit the actual word tokens (e.g., “I”, “am”) and then append one or more [LASTING] tokens to occupy the remainder of that word’s duration in the timeline. Here is an example show in Figure 2.

The large language model also generates dynamic text predictions. By encoding precise timing information into these text embeddings, the subsequent audio synthesis produces segments that are more tightly synchronized with the spoken content.

## B.3 Multimodal Context Modeling.

Our synchronous Multimodal LLM splits its inputs into **static** and **dynamic** streams and fuses them via a causally-aware omni-attention mechanism (See Figure 3):

- **Static inputs:** the instruction prompt and full conversation history, encoded as global tokens that remain unmasked and accessible at every time step.
- **Dynamic inputs:**
  - Frame-aligned visual embeddings.
  - Temporal text tokens for both speaker and listener, processed with Chrono-Markup.

All tokens enter a single omni-attention block enforcing strict causality *across* and *within* modalities:

- Visual tokens attend only to earlier visual tokens, to text tokens that precede the current frame and to all the static text tokens.
- Dynamic text tokens attend only to past visual tokens and past text tokens, and to all the static text tokens.
- Future dynamic tokens are masked out to preserve temporal integrity.
- Static tokens remain unmasked, ensuring that each update stays guided by the overarching instruction and dialogue context.

This design yields tightly synchronized, temporally coherent cross-modal interactions while maintaining global guidance.

## B.4 TempoVoice

TempoVoice is designed to transform generated textual tokens into temporally synchronized audio waveforms. Given the hidden representations corresponding to the listener’s text tokens,  $\mathbf{H}_{\tau:t}$ , TempoVoice generates audio tokens  $\mathbf{A}_{\lfloor \frac{\tau}{k} \rfloor : \mu}$ , which are then converted into continuous audio waveforms using an audio tokenizer.

The process is defined as follows:

$$\mathbf{A}_{\lfloor \frac{\tau}{k} \rfloor : \mu} = \text{TempoVoice}(P_{\lfloor \frac{\tau}{k} \rfloor : \mu}, [\mathbf{A}_{\text{voiceprint}}, \mathbf{H}_{\tau:t}]) \quad (2)$$

## Dynamic Text

[PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 Why [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] do [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] I [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] think [LASTING]  
 [LASTING] you're [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 in [LASTING] my [LASTING] life? [LASTING] [LASTING] [LASTING]  
 [LASTING] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 Okay. [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE] [PAUSE]  
 [PAUSE] That [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] brought [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] up [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] [LASTING] [LASTING] [LASTING] two [LASTING] [LASTING]  
 things [LASTING] [LASTING] [LASTING] [LASTING] [LASTING] [LASTING]  
 [LASTING] in [LASTING] [LASTING] my [LASTING] [LASTING] [LASTING]  
 mind. [LASTING] [LASTING]

Figure 2: Example of Dynamic Text.

where  $P_{\lfloor \frac{\tau}{k} \rfloor : \mu}$  denotes the positional encodings for positions  $\lfloor \frac{\tau}{k} \rfloor$  to  $\mu$ ,  $\mathbf{A}_{\text{voiceprint}}$  represents the voiceprint embeddings, and  $\mathbf{H}_{\tau:t}$  are the generated textual embeddings over the interval  $[\tau, t]$ . Here,  $[\cdot]$  indicates concatenation along the temporal axis.

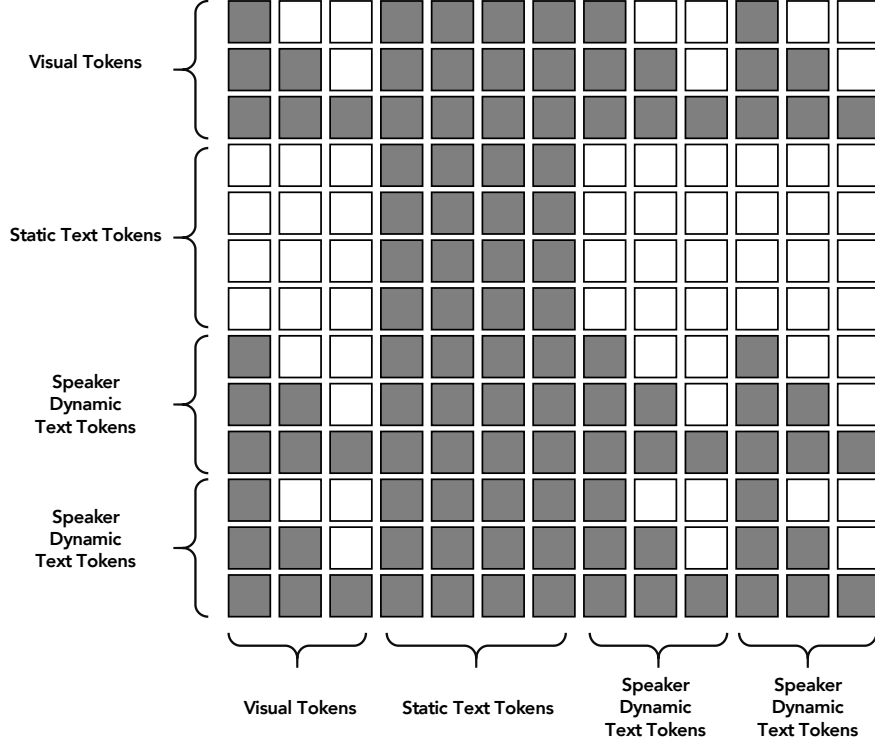


Figure 3: **Illustration of Multimodal Context Modeling.** Each visual token attends to all preceding visual tokens and static and dynamic text tokens annotated by Chrono-Text markers at earlier timestamps. Similarly, each dynamic text token attends to all past visual and textual tokens, enabling rich cross-modal context integration.

The resulting audio tokens  $\mathbf{A}_{\lfloor \frac{\tau}{k} \rfloor; \mu}$  are subsequently transformed into audio waveforms using the BiCodec module from Spark-TTS.

## B.5 Inference Speed

We benchmark our model at 15.6 FPS (64 ms latency) on a single NVIDIA A100 80GB, without deployment optimizations (e.g., flash-attention, multi-GPU parallelism, distillation, or quantization), indicating headroom for real-time deployment.

Table 2: Runtime and modality comparison.

| Method         | FPS $\uparrow$ | Generation Paradigm              | Audio Support            | Input Conditions                                |
|----------------|----------------|----------------------------------|--------------------------|---|
| Real Video     | —              | —                                | —                        | —   |
| SadTalker [22] | 1.81           | Offline full-sequence generation | Pre-recorded audio input | Video only; pre-recorded audio of same identity |
| Hallo [7]      | 0.13           | Offline full-sequence generation | Pre-recorded audio input | Video only; pre-recorded audio of same identity |
| <b>Ours</b>    | <b>15.62</b>   | Online, frame-by-frame           | Dynamically generated    | Video + Audio + Text; live partner audio/video  |

## C ResponseNet Dataset

### C.1 Construction Pipeline

To build the **ResponseNet** dataset, we design a three-stage pipeline encompassing data collection, data processing, and data refinement, as illustrated in Figure 4. This structured process ensures high-quality, temporally aligned multimodal data suitable for online conversational response generation tasks.

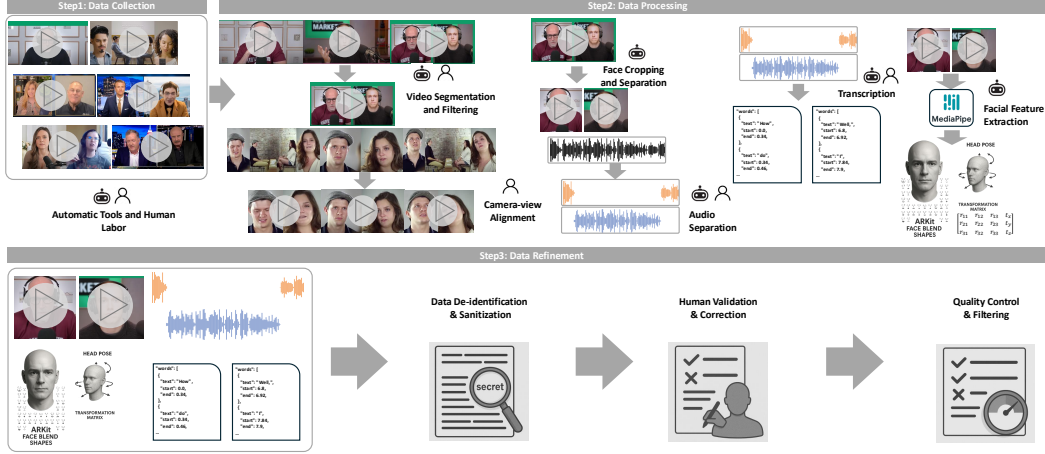


Figure 4: Illustration of Dataset Construction Pipeline.

**Step 1: Data Collection.** We begin by sourcing dyadic conversational videos from diverse public domains, including interviews, podcasts, and online discussions. Candidate videos are selected through a combination of automatic filtering tools and human curation to ensure conversational structure and speaker clarity. These videos include one speaker and one listener. The data is manually labeled for speaker turns, and high-resolution videos are retained for downstream visual analysis.

**Step 2: Data Processing.** This step extracts synchronized multimodal data from the raw videos. First, *Video Segmentation and Filtering* isolates segments with clear speaker-listener interaction using face detection and quality heuristics. We apply *Camera-view Alignment* to standardize the perspective, especially in multi-camera recordings. Next, we conduct *Face Cropping and Separation* to isolate individual speaker and listener views. In parallel, the audio track is separated and segmented using speaker diarization and voice activity detection. However, automatic tools could lead to bad cases, we correct these separated audio tracks manually. We then apply an ASR system (whisper [18]) for *Transcription* to obtain timestamped word-level alignments. Subsequently, we extract facial behavior features using MediaPipe [15], yielding per-frame ARKit blendshape coefficients and 3D head pose transformation matrices for both speaker and listener tracks.

**Step 3: Data Refinement.** To ensure privacy and label accuracy, we conduct multi-level cleaning. First, in the *De-identification and Sanitization* stage, we mask personally identifiable information (PII) and redact sensitive content from transcripts and audio. Then, *Human Validation and Correction* is performed to manually inspect and correct transcription errors, alignment mismatches, and feature inconsistencies. Finally, a *Quality Control and Filtering* phase discards corrupted or ambiguous segments, yielding a clean, high-quality dataset with tightly aligned audio, visual, and textual modalities.

Overall, the pipeline enables reliable construction of multimodal dialogue samples with rich facial dynamics and accurate verbal content, supporting the development of real-time response generation models.

## C.2 Dataset Statistics

The dataset is partitioned into training, validation, and test splits following the standard ratio of 6:2:2. Specifically, we ensure that the distributions of conversation topics, speaker identities, and recording conditions are balanced across each subset to avoid potential biases and to facilitate robust evaluation. The detailed statistics of the video stream pairs in each split are summarized in Table 3.

Each data sample consists of a synchronized pair of video streams representing a dyadic conversational interaction. The train, validation, and test splits are disjoint with respect to participant pairs to ensure fair evaluation and to prevent data leakage. This stratified partitioning enables comprehensive benchmarking of model performance across diverse conversational scenarios.

Table 3: Data split of video stream pairs in our dataset.

| Split        | Number of Video Stream Pairs | Proportion (%) |
|--------------|------------------------------|----------------|
| Train        | 417                          | 59.9           |
| Validation   | 139                          | 20.0           |
| Test         | 140                          | 20.1           |
| <b>Total</b> | 696                          | 100.0          |

We additionally analyze the dataset’s demographic diversity. The Table 4 summarizes identities, gender balance, ethnic distribution, and age bands.

Table 4: Demographic statistics of our dataset.

| Category   | Count / Share   |
|------------|---|
| Identities | 161 unique identities   |
| Gender     | Female: 93 (57.8%), Male: 68 (42.2%)  |
| Ethnicity  | White: 122 (75.8%), Black: 24 (14.9%), Asian: 15 (9.3%)   |
| Age bands  | 10–19: 10 (6.2%), 20–29: 63 (39.1%), 30–39: 51 (31.7%),<br>40–49: 17 (10.6%), 50–59: 17 (10.6%), 60–69: 2 (1.2%), 70+: 1 (0.6%) |

### C.3 Privacy Considerations

The YouTube platform enforces strict content moderation policies to prevent the dissemination of violent or harmful material. In addition, according to YouTube’s copyright guidelines<sup>2</sup>, the use of copyrighted material for research purposes typically qualifies as fair use, permitting reuse without the need for explicit permission from the copyright holder. Together, these factors ensure that our dataset collection and usage align with established privacy and ethical standards.

## D Evaluation Protocol

### D.1 Evaluation Metrics

Quantitative evaluation of multimodal response generation is inherently challenging due to the need to assess multiple aspects of quality across different modalities. To provide a comprehensive assessment, we employ a suite of metrics spanning text, audio, and visual outputs.

#### Text Metrics.

- **METEOR** [5]: Measures the alignment between generated and reference responses by considering synonymy, stemming, and word order, providing a nuanced evaluation of semantic adequacy.
- **BERTScore**<sub>F1</sub> [21]: Computes the similarity between generated and reference texts based on contextual embeddings from a pretrained RoBERTa model [14], offering a robust measure of semantic similarity.
- **ROUGE-L** [13]: Evaluates the longest common subsequence between generated and reference responses, reflecting fluency and content overlap.
- **Distinct-2** [12]: Calculates the proportion of unique bi-grams in the generated responses, serving as an indicator of output diversity and lexical richness.

#### Audio Metrics.

- **UTMOSv2** [4]: A neural mean opinion score (MOS) predictor that estimates the perceptual naturalness and intelligibility of the generated speech.

<sup>2</sup><https://www.youtube.com/howyoutubeworks/policies/copyright/#fair-use>

- **LSE-D (Lip–Speech Error Distance)** [17, 6]: Measures the temporal alignment and synchronization between generated audio and corresponding lip movements, reflecting audio-visual coherence.

### Visual Metrics.

- **Fréchet Distance (FD)** [3]: Computes the distributional distance between real and generated facial feature embeddings, assessing the realism of static visual features.
- **Fréchet Video Distance (FVD)** [19]: Quantifies the spatial-temporal quality of generated video sequences by comparing their feature distributions to those of real videos, thus evaluating overall video realism and consistency.

By leveraging these complementary metrics, we are able to rigorously assess the *appropriateness*, *naturalness*, *diversity*, and *synchronization* of generated responses across modalities, enabling a thorough benchmarking of model performance on the ResponseNet test set.

## D.2 Baseline Methods

As this is the first work addressing online multimodal conversational response generation (OMCRG), we compare OmniResponse with a diverse set of prior methods that target single-modality generation, as well as several strong multimodal baselines.

Specifically, we include the following baselines:

- **Offline Text Dialogue Generation Systems:** State-of-the-art large language models, including GPT-4o, GPT-4, and GPT-o1 [2], are evaluated for their ability to generate text responses in offline settings. These models only produce text outputs, without audio or visual generation.
- **Online Auditory Dialogue Generation System:** Moshi [8] is adopted as a representative model for generating spoken responses in real time, focusing exclusively on audio outputs.
- **Facial Reaction Generation Systems:** ReactFace [16] and ViCo [23] serve as facial reaction generation baselines, producing only visual (facial) responses based on the conversational context.
- **Online Multimodal Conversational Baselines:** To provide a direct comparison for OMCRG, we construct two multimodal baselines:
  1. A **LSTM-based model** [9] employing a recurrent neural network for temporal sequence modeling across modalities. The LSTM takes visual-audio-text modalities of speaker as inputs, and outputs listener’s visual and audio modalities.
  2. An **Audio-visual LLM baseline** that takes both speaker and listener audio–visual inputs and autoregressively generates audio-visual responses of the listener via a pre-trained large language model [1].

While previous approaches focus primarily on generating a single modality, OmniResponse is designed to produce synchronized and coherent responses across text, audio, and visual channels in an online setting.

## E Additional Experiments

### E.1 On SyncNet Metrics (LSE-D and LSE-C)

Table 5: SyncNet metrics. ↓ denotes lower is better; ↑ denotes higher is better.

| Method           | LSE-D ↓     | LSE-C ↑      |
|------------------|-------------|--------------|
| LSTM             | 9.72        | 0.157        |
| Audio–Visual LLM | 10.03       | 0.269        |
| <b>Ours</b>      | <b>9.56</b> | <b>0.371</b> |



Our evaluation follows SyncNet and its two metrics: LSE-D (lip-sync error distance; lower is better) and LSE-C (lip-sync confidence; higher is better). We adopt LSE-D in Table 5 using the official SyncNet evaluation script. We did not use LSE-C as a primary metric because it is not well suited to our setting: in multimodal conversation, the listener is often silent while the speaker talks. These silent spans are appropriate reactions but cause SyncNet’s confidence to drop, making the averaged LSE-C noisy for our task.

## **F Broader Impacts**

Our work contributes to the development of more intuitive and responsive multi-modal dialogue systems, with potential applications in education, healthcare, assistive communication, and companion. These technologies may improve access to information, support inclusive interaction, and enhance user experience across diverse contexts and scenarios. We encourage responsible research practices that prioritize transparency, user safety, and alignment with social values to ensure that such systems serve the public good.

## **G Responsibility and License**

We acknowledge full responsibility in the event of any rights infringement. The dataset is distributed under the Creative Commons CC BY-NC-SA license, permitting use with attribution for non-commercial purposes and requiring derivative works to be shared alike.

## References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] Helmut Alt and Michael Godau. Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(01n02):75–91, 1995.
- [4] Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 818–824. IEEE, 2024.
- [5] Satanjeev Banerjee and Alon Lavie. 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, 2005.
- [6] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017.
- [7] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21086–21095, 2025.
- [8] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.
- [9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. 1(2):3, 2022.
- [11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [12] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.
- [13] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [15] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.
- [16] Cheng Luo, Siyang Song, Weicheng Xie, Micol Spitale, Linlin Shen, and Hatice Gunes. Reactface: Multiple appropriate facial reaction generation in dyadic interactions. *arXiv preprint arXiv:2305.15748*, 2023.

- [17] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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, 2020.
- [18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- [19] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.
- [20] Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025.
- [21] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [22] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 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, 2023.
- [23] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: a benchmark dataset and baseline. In *European Conference on Computer Vision*, pages 124–142, 2022.