OmniResponse: Online Multimodal Conversational Response Generation in Dyadic Interactions

Cheng Luo[1], **Jianghui Wang**[1], **Bing Li**[1*], **Siyang Song**[2], **Bernard Ghanem**[1]
[1]King Abdullah University of Science and Technology, [2]University of Exeter

Abstract

In this paper, we introduce Online Multimodal Conversational Response Generation (OMCRG), a novel task designed to produce synchronized verbal and non-verbal listener feedback online, based on the speaker's multimodal inputs. OMCRG captures natural dyadic interactions and introduces new challenges in aligning generated audio with listeners' facial responses. To tackle these challenges, we incorporate text as an intermediate modality to connect audio and facial responses. We propose OmniResponse, a Multimodal Large Language Model (MLLM) that autoregressively generates accurate multimodal listener responses. OmniResponse leverages a pretrained LLM enhanced with two core components: Chrono-Text Markup, which precisely timestamps generated text tokens, and TempoVoice, a controllable online text-to-speech (TTS) module that outputs speech synchronized with facial responses. To advance OMCRG research, we offer ResponseNet, a dataset of 696 detailed dyadic interactions featuring synchronized split-screen videos, multichannel audio, transcripts, and annotated facial behaviors. Comprehensive evaluations on ResponseNet demonstrate that OmniResponse outperforms baseline models in terms of semantic speech content, audio-visual synchronization, and generation quality. Our dataset, code, and models are publicly available at `https://omniresponse.github.io/`.

1 Introduction

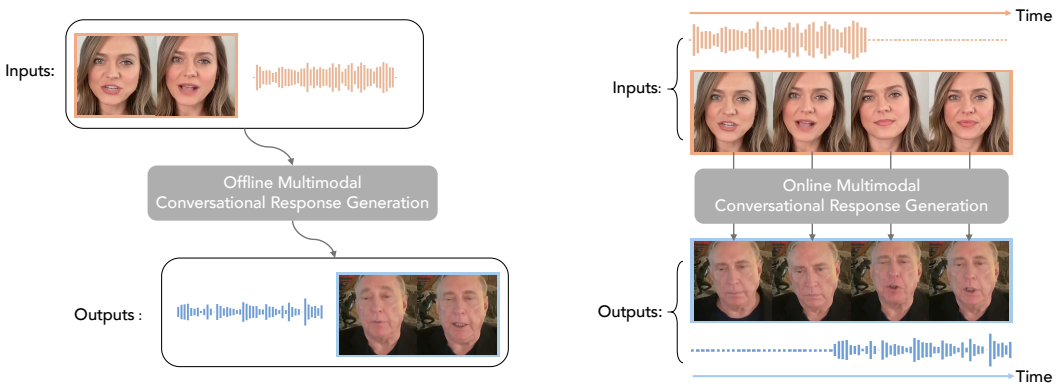
Generating realistic human conversational responses has substantial potential across numerous applications, spanning from human-computer interactions [39], immersive metaverse experiences [30], to mental health interventions [31]. However, human communication is inherently multimodal and complex. In face-to-face interactions, speakers convey their messages not only through spoken language but also through non-verbal cues, such as lip movements and facial expressions. Correspondingly, listeners provide multimodal responses consisting of verbal (e.g., audible affirmations or disapprovals) and non-verbal responses (e.g., subtle head nods). While considerable efforts [10, 67] have been dedicated to modeling text dialogue, particularly in language-based interfaces [35], modeling multimodal conversational interactions has been much underexplored.

In this paper, we explore a new task: learning to simultaneously generate verbal and non-verbal listener [2] responses in an online dyadic conversation setting, conditioned on the speaker's verbal and non-verbal inputs (see Figure 1). We refer to this task as Online Multimodal Conversational Response Generation. Although various audio-to-video generation methods (e.g. talking head generation [82, 84, 79]) have shown impressive performance, these methods focus on synthesizing visual content aligned with input audio signals, which ignores explicitly modeling multimodal conversational interactions. Recent studies [41, 47, 61] propose to generate facial reactions for a

*Corresponding author.

[2]Previous studies [8, 23] defined a speaker–listener framework for dyadic interactions, in which the listener both attends to the speaker's utterances and provides verbal and nonverbal feedback.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Offline Multimodal Conversational Response Generation

(b) Online Multimodal Conversational Response Generation

Figure 1: **Illustration of the new OMCRG task.** (a) In offline tasks, the generation model generates the listener's full response only after receiving the entire input sequence from the speaker. (b) Differently, OMCRG task requires sequentially processing the speaker's incoming input and generating multi-modal responses for the listener on the fly.

listener; however, these methods overlook verbal responses, which are essential to engage in dialogue fully.

The OMCRG task is complex and poses major challenges in three aspects. First, it is non-trivial to directly achieve synchronization between the generated audio and facial reactions of the listener for OMCRG task. As revealed in existing talking-head works [82, 66], achieving precise alignment between facial motion and audio is already challenging, even when the entire audio signal is given. In contrast, OMCRG is to generate both audio and facial reactions simultaneously and incrementally. Such online and multimodal generation settings make face-audio synchronization much more difficult, due to the high variability and semantic ambiguity of audio modality. Second, due to the online setting, the model has to reason over partial speaker input and generate audio-visual responses on the fly, which requires both powerful audio-visual understanding and generation abilities. While powerful pre-trained models have been developed for language and vision, audio modeling remains comparatively underdeveloped, making it more challenging to generate expressive and appropriate audio and facial reactions. Third, the lack of high-quality datasets for dyadic multimodal interaction significantly hinders the development of OMCRG.

We address the above challenges by proposing a unified framework, OmniResponse, which autoregressively generates high-quality multimodal listener responses. Rather than directly synchronizing generated audio and facial reactions, our key insight is to introduce text as an intermediate modality for the OMCRG task. Compared with audio, text offers clearer semantics and reduces uncertainty, making it more tractable for learning multimodal reaction generation. However, text is a static modality without inherent temporal information, posing challenges for synchronizing spoken words with visual frames in an autoregressive generation setting. To overcome this, we introduce a Multimodal Large Language Model (MLLM) augmented with two innovative modules: Chrono-Text and TempoVoice. The Chrono-Text module temporally anchors generated textual tokens by incorporating additional tokens (markers) that explicitly encode time, ensuring alignment between words and visual frames. TempoVoice is a controllable, online text-to-speech module designed to produce synchronized audio from these temporally annotated textual embeddings, ensuring accurate synchronization between audio and facial reactions.

In addition, we construct a high-quality dataset named ResponseNet, comprising 696 dyadic conversation pairs. Each pair includes synchronized split-screen video streams of both speaker and listener, multichannel audio recordings, verbatim text transcriptions, and detailed facial-behavior annotations (*i.e.*, facial expressions and head movements). Through extensive retrieval for scarce dyadic video data, rigorous content filtering, meticulous camera-shift alignment, and manual annotation, ResponseNet delivers a unique and valuable resource for benchmarking OMCRG.

Our contributions are summarized as follows: (1) we present OmniResponse, the first online model to jointly process and generate synchronized streams of conversational human behavior, establishing a

foundation for future work in human–agent interaction; (2) we introduce ResponseNet, an annotated dyadic conversation dataset and benchmark, enabling standardized evaluation of OMCRG models.

2 Related Work

Facial Reaction Generation. Facial reaction generation (FRG) [63, 83, 60] is a particularly challenging new task as it requires to predict the non-deterministic human facial reactions under different contexts. Early FRG approaches [27, 28] relied on Generative Adversarial Networks (GANs) [43, 21] typically conditioned the generation process on the speaker visual-speech behaviors. Since FRG is a non-deterministic process (*i.e.*, different facial reactions can be triggered by the same speaker behavior [63]), recent advances have shifted towards more sophisticated generative frameworks. For example, Ng *et al.* [47] introduces a non-deterministic approach based on Variational Autoencoders (VAEs) [32], which enabled sampling diverse human facial motions. This work was complemented by a novel dataset containing paired recordings of active speakers and silent listeners, providing essential training data for modeling natural reactions. Zhou *et al.* [83] developed a specialized speaker-listener video dataset for head motion generation, which is somewhat limited by its relatively short clip durations (median length of 9.0 sec) and modest dataset scale (1.58 hours total), and thus constraining their model's ability to learn long-term temporal dependencies. More recent works have attempted to address these limitations through innovative architectural choices or larger-scale datasets [61, 62]. Luo *et al.* [41, 15] and Zhu *et al.* [85] proposed transformer-based [70] VAE and diffusion models [64, 24], respectively, training them on a hybrid collection of videos from three different human-human dyadic interaction datasets [12, 57, 50].

Spoken Dialogue Models. Spoken dialogue models generate natural speech responses in real-time, requiring systems to process both verbal content and paralinguistic elements of communication. Early approaches including AudioPALM [58], Spectron [46], and SpeechGPT [77] adopted pipelines combining automatic speech recognition (ASR), text generation, and text-to-speech (TTS) synthesis. However, their requirement to complete the entire response before the speech generation makes them unsuitable for live human-computer interactions. Recent developments [44, 18, 49] have shifted towards end-to-end approaches that directly model speech-to-speech generation. Representative examples include Moshi et al. [18] and dGSLM [49], which operate as full-duplex speech dialogue systems capable of processing continuous speaker input while generating appropriate vocal responses. While these advances are significant, they focus exclusively on speech and text modalities, overlooking the crucial visual aspects of human communication. Even recent work by Park et al. [51] that includes visual-speech data is limited to intermittent speaker-listener interactions.

Autoregressive Generative Model. Transformer-based autoregressive models [70] have revolutionized numerous domains in AI, demonstrating remarkable success in language modeling [10, 67], multi-modal processing [40, 3, 34, 45], and generative tasks [56, 81, 74, 73, 72, 65]. Their success can be attributed to their inherent scalability and ability to unify multi-modal training under a single autoregressive objective, enabling seamless integration of different data modalities. The adaptation of transformers to visual tasks was pioneered by approaches such as VQVAE [69] and VQGAN [19], which introduced effective methods for quantizing visual information into discrete tokens. They align visual generation with the successful paradigm of language modeling by employing decoder-only transformers to predict sequences of image tokens. Subsequent research [13] has focused on enhancing both the efficiency of tokenization processes [42, 37] and sampling procedures [76], while simultaneously scaling up model architectures to handle increasingly complex tasks.

3 Methodology

Problem Definition. Let $\mathbf{F}_t^s$ and $\mathbf{A}_t^s$ be the speaker's facial and audio cues at time t, respectively. Given the speaker's streaming facial sequence $\mathbf{F}_{1:t}^s$ and audio sequence $\mathbf{A}_{1:t}^s$ from time 1 to t, the goal of OMCRG is to online generate facial reactions $\mathbf{F}_t^l$ and audio feedback $\mathbf{A}_t^l$ at time step t. Such multi-modal generation has been much less underexplored, different from recent works [83, 41, 18, 77] mainly focusing on single-modal response generation. To provide natural responses, it is crucial to ensure that the generated facial reactions and audio are temporally synchronized and react appropriately to the speaker. However, this is significantly challenging due to the inherent difficulty of online audio-visual understanding and generation.

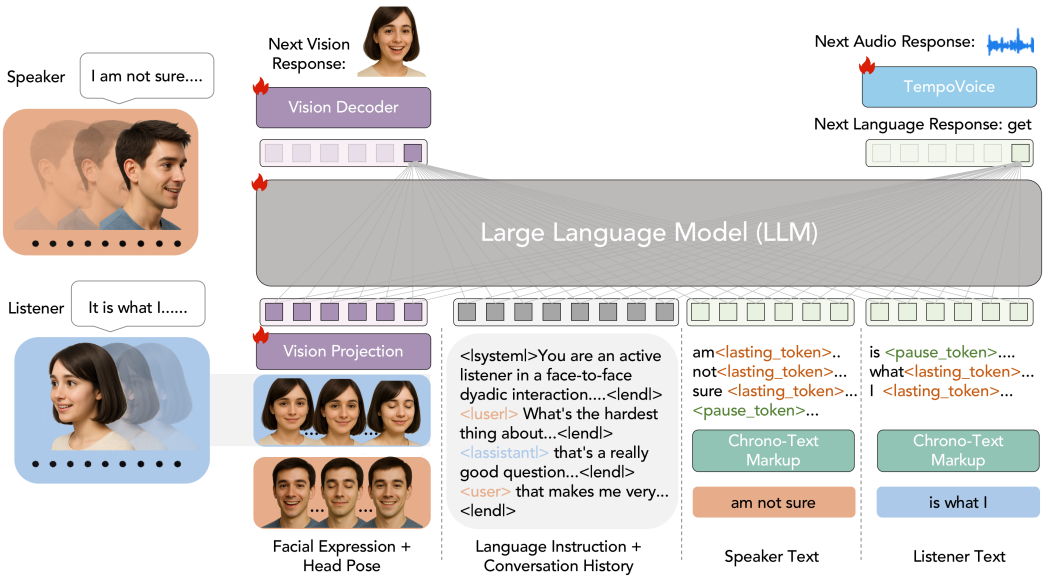


Figure 2: **Overview of the proposed OmniResponse**. The model takes textual conversational history and newly coming multimodal information (e.g., facial cues) from the speaker and listener as input, and generates temporally synchronized facial and textual responses for the listener by leveraging a pre-trained LLM enhanced with our proposed Chrono-Text Markup. The generated text embeddings are converted into audio synchronized with the facial response by the proposed TempoVoice module.

Instead of generating audio and visuals directly, we treat text as an intermediate modality and decompose OMCRG into two subproblems: (i) joint text–and–face response generation—producing temporally aligned facial reactions $\mathbf{F}_t^l$ and textual responses $\mathbf{W}_t^l$; and (ii) synchronous text-to-speech synthesis—converting $\mathbf{W}_t^l$ into audio waveform segments $\mathbf{A}_t^l$ that are aligned with the facial reactions. However, because text lacks explicit temporal information, achieving tight alignment with facial and audio streams is challenging for both subproblems. We address this issue with two novel modules.

Overview. We present OmniResponse, a novel framework for the OMCRG task (see Figure 2), where OmniResponse is a new MLLM enhanced by two proposed key components: *Chrono-Text Markup* and *TempoVoice*. In particular, our OmniResponse leverages the capability of a pretrained LLM to understand and interpret the speaker's multimodal inputs and autoregressively generate meaningful responses in terms of textual and facial responses. To address the lack of temporal information in text, the proposed *Chrono-Text Markup* embeds explicit temporal marks between text tokens, endowing the input and output text with time-aware embeddings and ensuring precise alignment with the generated facial reactions. Furthermore, the proposed *TempoVoice* generates audio responses temporally synchronized with both the generated textual response and the listener's facial movements.

3.1 OmniResponse

Model Architecture. As shown in Figure 2, OmniResponse processes multiple modalities from the speaker and the listener, temporally aligns different modalities, and outputs synchronous multimodal responses to the speaker. In particular, at each time step t, OmniResponse consumes: (1) *Static text inputs*: a task-specific instruction prompt W_{instruct} and the conversation history prior to time τ ($\tau < t$), denoted $W_{\text{history},<\tau}$; and (2) *Temporal inputs*: the previously generated facial features of the listener $\hat{\mathbf{F}}^l_{\tau:t-1}$, the facial features of the speaker $\mathbf{F}^s_{\tau:t-1}$ and the accumulated text sequences from both participants ($\mathbf{W}^s_{\tau:t-1}$, $\hat{\mathbf{W}}^l_{\tau:t-1}$) over the interval $[\tau, t-1]$. Using these inputs, OmniResponse predicts the next facial features $\hat{\mathbf{F}}^l_t$, the verbal response $\hat{\mathbf{W}}^l_t$, and the corresponding speech segment $\hat{\mathbf{A}}^l_\mu$ in the current frame, ensuring precise temporal alignment in all modalities. Formally, we defined this process as:

$$\{\hat{\mathbf{F}}^l_t, \hat{\mathbf{A}}^l_\mu, \hat{\mathbf{W}}^l_t\} = \mathcal{M}\big(W_{\text{instruct}}, W_{\text{history},<\tau}, \mathbf{F}^s_{\tau:t-1}, \hat{\mathbf{F}}^l_{\tau:t-1}, \mathbf{W}^s_{\tau:t-1}, \hat{\mathbf{W}}^l_{\tau:t-1}\big).$$

Vision Projection. We introduce the vision projection layer to enable the pretrained LLM (Phi-3.5 mini-instruct with 3.8B parameters [1]) to process visual facial features. The layer is implemented as a multilayer perceptron (MLP) that maps the the listener's and speaker's past facial features $\hat{\mathbf{F}}^l_{1:t-1}$ and $\mathbf{F}^s_{1:t-1}$ into embedding features $\mathbf{V}_{1:t-1}$ aligned with the LLM token space. During autoregressive generation, the MLLM employs causal self-attention [70] to model temporal dependencies between the next token and previous one, and outputs the next listener vision embedding $\hat{\mathbf{V}}^l_t$.

Vision Decoder. A learnable vision decoder, comprising transformer layers, converts $\hat{\mathbf{V}}^l_t$ back into the original coefficient space to produce the predicted listener facial coefficients $\hat{\mathbf{F}}^l_t$. Subsequently, a pre-trained visual renderer maps these visual coefficients to 2D frames, using a given portrait image. Please refer to the appendix for additional details.

Chrono-Text Markup. Visual frames inherently encode temporal information, whereas text tokens are static and lack any temporal dimension. Additionally, visual frames and textual tokens typically differ in length due to their fundamentally different modalities, making unified autoregressive prediction challenging. To resolve this mismatch, we propose *Chrono-Text Markup*, a novel yet straightforward approach that explicitly embeds temporal information into textual data, aligning the textual sequence precisely with the visual frame sequence. Unlike prior approaches such as TimeMarker [14], which inserts timestamps only between visual frames or the method by Ng et al. [48], which integrates timestamp embeddings into textual tokens, our method employs only two special markers, ensuring that the textual and visual sequences have identical lengths. Specifically, we insert two special tokens into the transcript: `[PAUSE]` to denote silent intervals between utterances, and `[LASTING]` to indicate that the previous textual word continues speaking to the current time. Each text token is placed between pause and lasting tokens.

Multimodal Context Modeling. Our synchronous Multimodal LLM integrates both static and dynamic inputs: *Static inputs*: the instruction prompt and the accumulated conversation history. *Dynamic inputs*: frame-aligned visual embeddings and timestamped textual tokens for both speaker and listener. All tokens are jointly processed by an *omni-attention* mechanism that enforces causal, cross-modal interactions. Under this operation, each visual token attends to preceding visual tokens and to text tokens marked by chrono-text markers at earlier timestamps; similarly, each dynamic text token attends to past visual and textual tokens. However, this omni-attention prevents dynamic tokens from looking at future tokens. This ensures the generation adheres to temporal dynamics and cross-modal interactions. Meanwhile, static tokens remain globally accessible, ensuring that every dynamic update remains guided by the overarching instructions.

TempoVoice. Generating natural speech that is precisely synchronized with text and facial frames poses a significant challenge. To address this, we introduce a dedicated synthesis pipeline, *TempoVoice*. Our framework begins by combining the listener's voiceprint, extracted via the Spark-TTS global tokenizer [71] to capture speaker identity, with the hidden states of the generated text (see Figure 3). We then apply sinusoidal positional encodings to the merged embeddings. Since audio-token sequences typically differ in length from visual frames and textual tokens, we prepend a series of zero-initialized placeholder tokens, each endowed with positional information. These placeholders serve as queries in a cross-attention module within a Transformer decoder, attending over the fused text–voice representations. This mechanism enables fully synchronous, autoregressive generation of audio tokens in lockstep with visual frames and text tokens. Finally, a linear projection layer maps the decoder outputs to logits over the discrete audio-codec vocabulary.

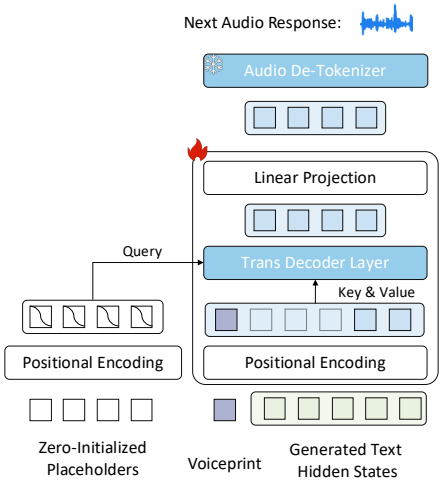


Figure 3: **Architecture of TempoVoice.** TempoVoice transforms textual hidden-state embeddings into audio segments.

The decoder logits are then quantized into discrete audio semantic tokens $\hat{\mathbf{A}}_\mu$, as defined by the Spark-TTS audio tokenizer [71]. Conditioned on these semantics and the global speaker-identity embeddings, the tokenizer reconstructs the continuous waveform segment.

3.2 Training Objectives

To train OmniResponse, the training objective is a weighted combination of text generation loss $\mathcal{L}_{\text{text}}$, vision reconstruction $\mathcal{L}_{\text{vision}}$, and audio generation loss $\mathcal{L}_{\text{audio}}$:

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \lambda_{\text{vision}}\mathcal{L}_{\text{vision}} + \lambda_{\text{audio}}\mathcal{L}_{\text{audio}}, \tag{1}$$

where λ_{vision} and λ_{audio} are the scaling factors balancing text, vision, and audio loss terms.

Text Generation Loss. The text loss encourages accurate next-token prediction conditioned on both speaker context and past listener states:

$$\mathcal{L}_{\text{text}} = -\sum_{t} \log p_\theta\big(W_t^l \mid W_{\text{instruct}},\, W_{\text{history},<\tau},\, \mathbf{F}^s_{\tau:t-1},\, \hat{\mathbf{F}}^l_{\tau:t-1},\, \mathbf{W}^s_{\tau:t-1},\, \hat{\mathbf{W}}^l_{\tau:t-1}\big). \tag{2}$$

Vision Reconstruction Loss. To align predicted and ground-truth facial dynamics, we apply an ℓ_2 reconstruction loss on the listener's feature embeddings:

$$\mathcal{L}_{\text{vision}} = \sum_{t} \big\|\hat{\mathbf{F}}^l_t - \mathbf{F}^l_t\big\|_2^2. \tag{3}$$

Audio Generation Loss. The audio loss operates over discrete semantic tokens $\mathbf{A}^l_\mu$, indexed by μ, which correspond to frame indices $t = \mu k$ (k is the downsampling factor). We maximize the likelihood of each token conditioned on previous audio semantics and the listener's hidden states:

$$\mathcal{L}_{\text{audio}} = -\sum_{\mu} \log p_\theta\big(\mathbf{A}^l_\mu \mid \mathbf{A}^l_{<\mu},\, \mathbf{H}_{t-k+1:t}\big), \tag{4}$$

where $\mathbf{H}_{t-k+1:t}$ denotes the model's hidden representations for the corresponding listener text tokens $\hat{\mathbf{W}}^l_{t-k+1:t}$. This formulation ensures coherent alignment across modalities throughout generation.

4 Dataset Construction

Existing publicly available dyadic video datasets do not satisfy the requirements of the OMCRG task (See Figure 1). For example, mono-view talking-head datasets and offline dialogue corpora (e.g., MultiDialog [51]) do not offer split-screen recordings that capture speaker and listener simultaneously. Others, such as IEMOCAP [11], feature predominantly side profile views recorded in noisy environments and provide only mixed audio channels, thus preventing separate analysis of each participant's speech. Furthermore, datasets such as ViCo [82], ICD [47], and REACT2024 [61] lack comprehensive textual annotations, suffer from low video resolution [82, 11, 61], or exhibit inconsistent spoken languages [61]. To fill the dataset gap, we introduce ResponseNet that comprises 696 temporally synchronized dyadic video pairs, totaling over 14 hours of natural conversational exchanges. Each pair provides high-resolution (1024×1024) frontal-face streams for both speaker and listener, along with separated audio channels to support fine-grained analysis of verbal and nonverbal behavior. Table 1 shows ResponseNet is the only dataset that satisfies the key requirements: (1) online video streaming, (2) separate audio channels, and (3) textual word-level annotations for both participants.

The construction of ResponseNet follows a rigorous workflow that integrates automated tools with extensive human-in-the-loop curation. (1) Initially, split-screen videos featuring simultaneous appearances of speaker and listener are sourced from YouTube according to predefined topic and quality criteria. These clips are then filtered to remove low-resolution, noisy, or frequent camera transitions. (2) Human annotators perform a thorough review to correct camera-view mis-alignments and ensure precise temporal synchronization between streams. (3) Next, mixed-channel audio tracks are automatically separated into discrete speaker and listener channels using speaker separation tools such as MossFormer2 [80] and subsequently verified and refined by experts. Finally, word-level transcripts are generated via automatic speech recognition [55] and meticulously proofread to guarantee accuracy. By combining automation with meticulous manual oversight across data sourcing, preprocessing, alignment, audio separation, and annotation, this pipeline yields a high-quality, richly annotated dyadic video corpus ideally suited for multimodal conversational response generation.

Table 1: **Comparison of conversation datasets.** and denote speaker and listener data respectively. ResponseNet provides complete multimodal data (speaker+listener) with their separated audios.

Dataset	Video	Audio	Text	Online	Separated Audios	# Dialogues	Total Duration
MultiDialog [51]	+	+	+	✗	✓	8,733	339.7h
ICD [47]	+	+	✗	✓	✓	182,132	72h
ViCo [83]	+		✗	✓	✗	483	1.6h
REACT2024 [62]	+	+	✗	✓	✓	5,919	71.8h
IEMOCAP [11]	+	+	+	✓	✗	151	11.5h
ResponseNet	+	+	+	✓	✓	696	14.2h

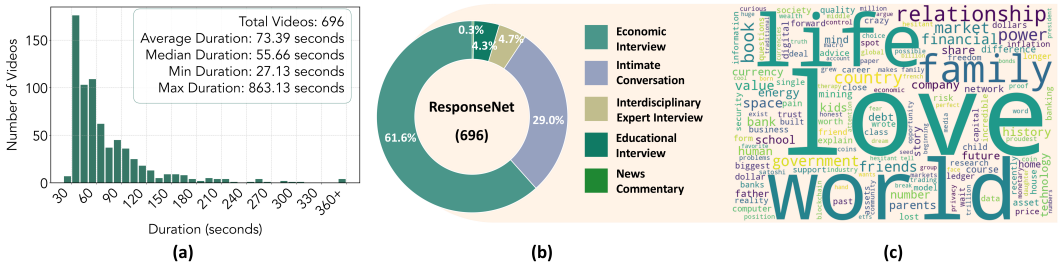


Figure 4: **Statistics of ResponseNet**. (a) Distribution of video clip durations. (b) Distribution of dyadic conversation topics. (c) Word cloud of spoken words in dyadic conversations.

The statistics of ResponseNet are shown in Figure 4. The durations of speaker-listener video clips range from 27.13 seconds (short conversations) to 863.13 seconds (long conversations) in ResponseNet. Figure 4.(a) shows that the average clip duration in ResponseNet is 73.39 seconds, significantly longer than that of other dyadic datasets such as REACT2024 (30 seconds), and ViCo (9 seconds). This extended duration ensures that each clip captures sufficient conversational exchanges. Figure 4.(b) illustrates that the conversations span a diverse range of topics, including professional discussions (e.g., economic interviews, news commentaries), emotionally driven interactions (e.g., intimate conversations), educational settings (e.g., teaching interviews), and interdisciplinary expert discussions. Figure 4.(c) presents a word cloud highlighting the most frequent words in the conversations. Such diversity shows that ResponseNet captures rich and varied human-human interactions rather than being restricted to narrow or monotonic conversation patterns.

5 Experiments

Implementation Details. Our framework was implemented using PyTorch [52] and trained on four NVIDIA Tesla A100 GPUs. The model optimization was performed using the AdamW optimizer [33] with a learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 10^{-4}, accompanied by a cosine learning rate scheduler. Training was executed with a batch size of one for 2,000 epochs. Additionally, we fine-tuned the LLM using the LoRA [26] technique with a LoRA rank of 64 and a LoRA alpha value of 16. More implementation details are provided in the Appendix.

Evaluation Metrics. Quantitatively evaluating the quality of multimodal response generation remains non-trivial. We thereby employ comprehensive metrics to evaluate generation results across text, audio, and visual modalities. For text response, we use METEOR [9], BERTScore$_{F1}$ [78], and ROUGE-L [38] to measure how *appropriate* and *natural* the generated responses are, based on reference responses from the ResponseNet test set. We also adopt Distinct-2 [36] to evaluate *diversity* through the ratio of unique bi-grams. For audio response, we adopt UTMOSv2 [6], a neural MOS predictor that estimates the perceptual naturalness, and employ LSE-D [54, 16] (Lip–Speech Error Distance) to evaluate *synchronization* between generated speech and lip movements. For facial response, we compute Fréchet Distance (FD) [4] between real and generated facial-feature distributions, and Fréchet Video Distance (FVD) [68] to assess the spatial–temporal *visual quality* of generated video sequences.

Table 2: **Quantitative Results on ResponseNet test set.**

Model	Text				Audio		Video	
	METEOR ↑	BERTScore$_{F1}$ ↑	ROUGE-L ↑	Distinct-2 ↑	LSE-D ↓	UTMOSv2 ↑	FD ↓	FVD ↓
Ground-Truth	–	–	–	0.835	8.96	1.56	–	–
Offline Text Dialogue Generation System								
GPT-4o [2]	0.167	0.805	0.079	0.928	–	–	–	–
GPT-4 [2]	0.163	0.822	0.082	0.960	–	–	–	–
GPT-o1 [2]	0.189	0.822	0.113	0.948	–	–	–	–
Qwen-7B-Chat [7]	0.167	0.807	0.090	0.920	–	–	–	–
Claude-Sonnet-4 [5]	0.183	0.807	0.101	0.966	–	–	–	–
Gemini-2.5-Flash [17]	0.175	0.824	0.085	0.932	–	–	–	–
DeepSeek-R1 [22]	0.173	0.815	0.078	0.981	–	–	–	–
Online Auditory Dialogue Generation System								
Moshi [18]	0.120	0.818	0.078	0.499	–	2.21	–	–
Facial Reaction Generation System								
ReactFace [41]	–	–	–	–	–	–	32.72	340.28
ViCo [83]	–	–	–	–	–	–	57.13	325.65
Online Multimodal Conversational Response Generation Baseline								
LSTM [25]	0.042	0.716	0.000	0.000	9.72	1.21	**6.51**	320.92
Audio-visual LLM	0.030	0.662	0.020	0.155	10.03	1.32	580.86	681.55
OmniResponse (Ours)	**0.141**	**0.806**	**0.081**	**0.882**	**9.56**	**1.41**	15.46	**314.94**

5.1 Quantitative Results

To the best of our knowledge, few works have explored the OMCRG task before. We build two baselines and compare them in Table 2: (1) LSTM-based method employing a recurrent neural network [25] for temporal sequence modeling; (2) Audio-visual LLM taking speaker–listener audio and visual inputs and leveraging pre-trained LLM to generate audio–visual frames autoregressively. Table 2 further reports the generation performance of representative single-modality baselines, including offline, text-only dialogue models (e.g., GPT variants [2], Qwen-7B-Chat [7], Claude-Sonnet-4 [5] (version 2025-05-14), Gemini-2.5-Flash [17], and DeepSeek-R1 [22] (version 2025-05-28)), online audio-only generation models (e.g., Moshi [18]), and facial reaction generation approaches [41, 83]. Different from these methods focusing on generating a single modality, our method enables online, synchronized generation across audio, visual, and textual modalities for modeling human conversation.

Table 2 shows that our OmniResponse achieves the best performance in dialogue speech content (METEOR, BERTScore$_{F1}$, ROUGE-L, Distinct-2), audio quality (UTMOSv2), audio–visual synchronization (LSE-D), as well as temporal consistency and visual quality (FVD). Although the LSTM baseline achieves a lower FD owing to its tendency to produce repetitive static visual output, it fails to generate rich, synchronized multimodal responses. Audio-Visual LLM does not incorporate the text modality, compared to our method. Consequently, Audio-Visual LLM achieves much lower speech content quality (METEOR and BertScore$_{F1}$) and struggles with audio–visual synchronization (LSE-D) than our method. Although Audio-Visual LLM leverages a powerful LLM, it is still challenging to directly synchronize generated audio with facial reactions, especially in the absence of a strong audio foundation model.

Our OmniResponse model significantly outperforms Audio-visual LLM across all evaluated metrics, including non-verbal ones. These results demonstrate that introducing text as an intermediate modality greatly enhances the naturalness and realism of non-verbal responses, as reflected by the FD and FVD scores. Moreover, we introduce a novel framework that effectively adapts pre-trained LLMs for audio–visual generation with the proposed Chrono-Text Markup and Tempo Voice.

5.2 Qualitative Results

Figure 5 presents a qualitative result. The synthesized listener remains silent while the speaker is speaking, but then produces an immediate or delayed response at the end of each speaker turn. This behavior demonstrates that OmniResponse effectively captures the temporal dynamics of online dyadic conversation and generates responses at appropriate timestamps. For example, between 100.97 and 132.05 s, the listener interjects briefly between 120.13 and 121.57 s in response to the speaker's ongoing content, reflecting natural human conversational interaction. In contrast, a

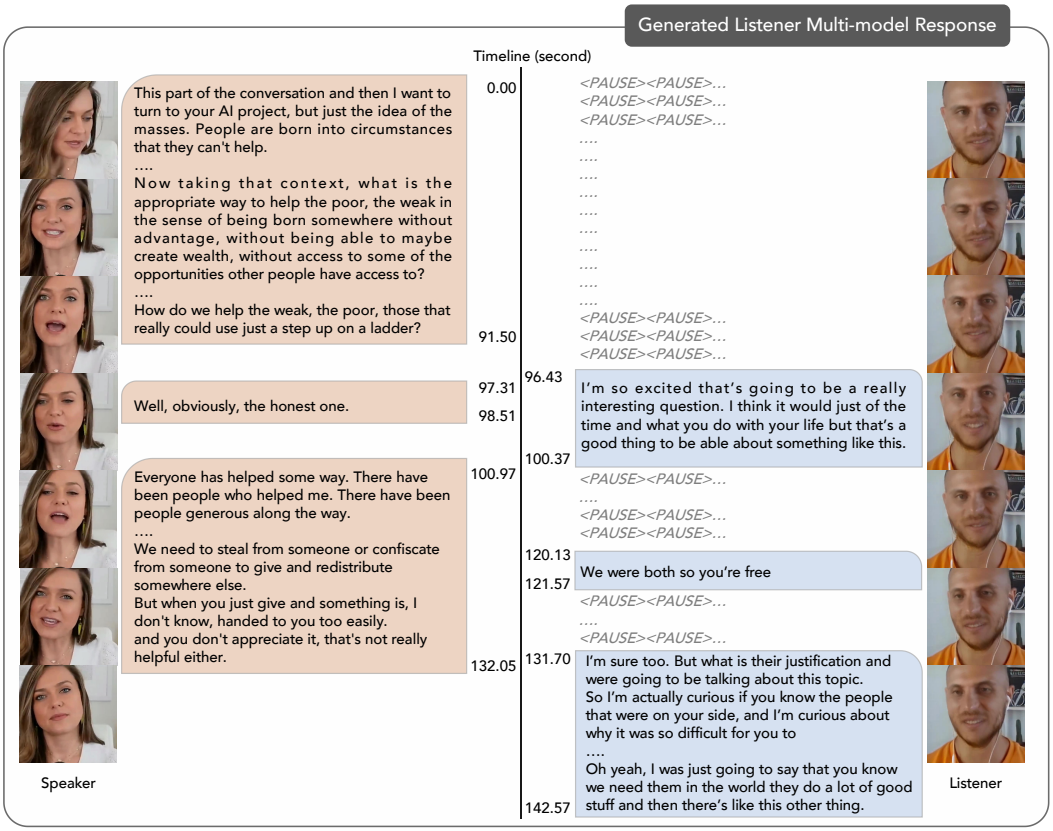


Figure 5: **Qualitative Results.** Given the speaker's audio and video streams and corresponding utterances (left), OmniResponse autoregressively generates synchronized visual, audio, and textual response streams (right). For clarity, `[LASTING]` tokens are removed from the generated dialogue.

conventional pipeline that integrates Automatic Speech Recognition (ASR), dialogue generation, TTS, and talking-head components waits for a predefined silence threshold before producing an offline multimodal response, thus diminishing conversational behaviors such as interruptions, backchannels, questions, and immediate feedback. In contrast, OmniResponse maintains the continuous flow of dyadic conversation by continuously modeling and generating synchronized time series streams of textual, visual, and audio outputs.

5.3 Ablation Studies

Effectiveness of Chrono-Text Markup. We construct baselines removing the proposed Chrono-Text Markup from our OmniResponse. In the baselines, each predicted word is assigned a timestamp indicating when it emerges; if this timestamp falls within a temporal window around the current time, the word is retained and appended to the spoken output; otherwise, it is discarded. As shown in the last rows of Table 3, incorporating Chrono-Text Markup significantly improves audio-visual synchronization, reducing the LSE-D score from 11.51 to 9.56. In addition, it enhances the semantic alignment of speech with conversational context, increasing METEOR from 0.122 to 0.141 and BERTScore$_{F1}$ from 0.766 to 0.806. Improvements in FD and UTMOSv2 further indicate that Chrono-Text Markup boosts the quality of the generated audio and facial responses. These results demonstrate the effectiveness of Chrono-Text Markup in generating high-quality multimodal responses.

Effectiveness of TempoVoice. To study the effect of our TempoVoice, we remove it from our framework and instead directly feed the hidden states, which are trimmed or padded to match the target audio length, into a multi-layer perceptron to predict audio token logits. As shown in Table 3, removing TempoVoice degrades audio–visual synchronization and reduces the quality of generated audio responses, where UTMOSv2 drops from 1.41 to 1.23, and LSE-D increases from 9.56 to 11.91.

Table 3: Ablation study on the effects of the proposed Chrono-Text Markup and TempoVoice.

Chrono-Text Markup	Tempo Voice	METEOR	BERTScore$_{F1}$	LSE-D	UTMOSv2	FD
✗	✗	0.090	0.755	13.64	1.21	596.27
✓	✗	0.128	0.778	11.91	1.23	19.58
✗	✓	0.122	0.766	11.51	1.39	23.42
✓	✓	**0.141**	**0.806**	**9.56**	**1.41**	**15.46**

Table 4: User study (A/B preference; higher is better). Each cell shows the percentage of participants preferring *Ours*.

Criteria	Ours vs. LSTM	Ours vs. Audio–Visual LLM
Speech Content Appropriateness	75.5%	81.6%
Audio Speech Quality	77.6%	85.7%
Visual Quality	67.3%	93.4%
Audio–Visual Synchronization	91.8%	95.9%

These results highlight the importance of TempoVoice in temporally aligning audio with the other modalities and enhancing the quality of the generated audio.

5.4 User Study

We conducted a user study with 49 participants (28 male, 21 female). Each subject viewed 16 randomly ordered clips and rated speech content appropriateness, audio speech quality, visual quality, and audiovisual synchronization. All participants were proficient in English (53.1% reported advanced proficiency or daily-communication ability). Educational attainment was high: 95.9% held at least an undergraduate degree, 44.9% held a master's degree, and 18.4% held a Ph.D. Ages were distributed as follows: 14.3% under 25, 34.7% aged 26–35, 24.5% aged 36–45, and 26.5% aged 46–55. In direct A/B preferences, "Ours" achieved a minimum preference of 67.3% (speech content appropriateness vs. LSTM) and a maximum of 95.9% (audiovisual synchronization vs. Audio–Visual LLM).

6 Conclusion and Discussion

We have presented OmniResponse, an online multimodal generation model that produces verbal and nonverbal listener responses to a speaker's multimodal behaviors. OmniResponse integrates techniques for processing multimodal inputs, synchronizing across modalities, and aligning responses with the speaker's content. To enable evaluation of this task, Online Multimodal Conversational Response Generation in Dyadic Interactions, we introduce ResponseNet, a dataset containing parallel recordings of speaker and listener streams. Our model and dataset lay the foundation for future research in this emerging field. Experimental results demonstrate that OmniResponse significantly increases speech semantic content, audio-visual synchronisation, audio and visual quality.

Limitations. While our approach performed well on the evaluated datasets, the remaining challenges include the proposed approach (e.g., its results) may largely depend on the quality and diversity of training data, replying on accurate speaker–listener segmentation and can be negatively affected in noisy or overlapping conversations. Additionally, generating well-aligned multi-modal responses remains difficult in fast-changing or emotionally rich interactions, while our paper lacks fairness analysis. Future work will focus on improving these aspects.

Risks and Potential Misuse. This system is developed for multi-modal conversational AI, but certain risks should be acknowledged. For instance, realistic synthetic contents could be misused [59] for impersonation or misleading information. During real-time human-user interactions, users may also develop misunderstandings or excessive reliance on the system without proper contents control. To avoid these risks, we recommend clear labeling of the generated contents, appropriate usage monitoring, and the inclusion of protective measures [20, 75] (e.g., Deepfake Detection [53, 29]) against potential misuse.

Acknowledgments. This work is supported by the KAUST Center of Excellence for Generative AI under award number 5940. The computational resources are provided by IBEX, which is managed by the KAUST Supercomputing Core Laboratory.

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 Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[4] Helmut Alt and Michael Godau. Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(01n02):75–91, 1995.

[5] Anthropic. Introducing claude 4 (opus 4 and sonnet 4). `https://www.anthropic.com/news/claude-4`, 2025. Model announcement and overview.

[6] 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.

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[8] MM Bakhtin. The problem of speech genres. 1987.

[9] 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.

[10] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

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

[12] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 350–359, 2017.

[13] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

[14] Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024.

[15] Luo Cheng, Song Siyang, Yan Siyuan, Yu Zhen, and Ge Zongyuan. Reactdiff: Fundamental multiple appropriate facial reaction diffusion model. *arXiv preprint arXiv:2510.04712*, 2025.

[16] 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.

[17] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[18] 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.

[19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[20] Yuan Gan, Jiaxu Miao, Yunze Wang, and Yi Yang. Silence is golden: Leveraging adversarial examples to nullify audio control in ldm-based talking-head generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13434–13444, 2025.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[23] Dirk KJ Heylen. Understanding speaker-listener interaction. In *10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009*, pages 2151–2154. International Speech Communication Association (ISCA), 2009.

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[26] 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.

[27] Yuchi Huang and Saad M Khan. Dyadgan: Generating facial expressions in dyadic interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–18, 2017.

[28] Yuchi Huang and Saad M Khan. Generating photorealistic facial expressions in dyadic interactions. In *BMVC*, page 201, 2018.

[29] Naseem Khan, Tuan Nguyen, Amine Bermak, and Issa Khalil. Unmasking synthetic realities in generative ai: A comprehensive review of adversarially robust deepfake detection systems. *arXiv preprint arXiv:2507.21157*, 2025.

[30] Do Yuon Kim, Ha Kyung Lee, and Kyunghwa Chung. Avatar-mediated experience in the metaverse: The impact of avatar realism on user-avatar relationship. *Journal of Retailing and Consumer Services*, 73:103382, 2023.

[31] Everlyne Kimani, Timothy Bickmore, Ha Trinh, and Paola Pedrelli. You'll be great: virtual agent-based cognitive restructuring to reduce public speaking anxiety. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*, pages 641–647. IEEE, 2019.

[32] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024.

[35] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[36] 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.

[37] Xiang Li, Hao Chen, Kai Qiu, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*, 2024.

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

[39] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)*, 4(4):1–28, 2013.

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

[41] 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.

[42] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

[43] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[44] Kentaro Mitsui, Koh Mitsuda, Toshiaki Wakatsuki, Yukiya Hono, and Kei Sawada. Pslm: Parallel generation of text and speech with llms for low-latency spoken dialogue systems. *arXiv preprint arXiv:2406.12428*, 2024.

[45] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

[46] Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*, 2023.

[47] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022.

[48] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10083–10093, 2023.

[49] Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, et al. Spirit-lm: Interleaved spoken and written language model. *arXiv preprint arXiv:2402.05755*, 2024.

[50] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1–12, 2021.

[51] Se Jin Park, Chae Won Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeong Hun Yeo, and Yong Man Ro. Let's go real talk: Spoken dialogue model for face-to-face conversation. *arXiv preprint arXiv:2406.07867*, 2024.

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[53] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024.

[54] 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.

[55] 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.

[56] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[57] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

[58] 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. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

[59] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

[60] Siyang Song, Micol Spitale, Xiangyu Kong, Hengde Zhu, Cheng Luo, Cristina Palmero, German Barquero, Sergio Escalera, Michel Valstar, Mohamed Daoudi, et al. React 2025: the third multiple appropriate facial reaction generation challenge. *arXiv preprint arXiv:2505.17223*, 2025.

[61] Siyang Song, Micol Spitale, Cheng Luo, German Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth Andre, et al. React2023: The first multiple appropriate facial reaction generation challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9620–9624, 2023.

[62] Siyang Song, Micol Spitale, Cheng Luo, Cristina Palmero, German Barquero, Hengde Zhu, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth André, and Hatice Gunes. React 2024: the second multiple appropriate facial reaction generation challenge. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 2024.

[63] Siyang Song, Micol Spitale, Yiming Luo, Batuhan Bal, and Hatice Gunes. Multiple appropriate facial reaction generation in dyadic interaction settings: What, why and how? *arXiv preprint arXiv:2302.06514*, 2023.

[64] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[65] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[66] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024.

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

[68] 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.

[69] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[70] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[71] 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.

[72] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

[73] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

[74] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[75] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*, 2023.

[76] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

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

[78] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[79] 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.

[80] S Zhao, Y Ma, C Ni, C Zhang, H Wang, TH Nguyen, K Zhou, J Yip, D Ng, and B Ma. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. arxiv 2024. *arXiv preprint arXiv:2312.11825*.

[81] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

[82] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306, 2019.

[83] 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.

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

[85] Hengde Zhu, Xiangyu Kong, Weicheng Xie, Xin Huang, Linlin Shen, Lu Liu, Hatice Gunes, and Siyang Song. Perfrdiff: Personalised weight editing for multiple appropriate facial reaction generation. In *ACM Multimedia 2024*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's main claims are clearly supported by its contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in our supplementary materials (appendix).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our work focuses on applications of multimodel generative model.

Guidelines: The paper does not include theory assumptions or proofs

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed implementation information in the main paper, including model architecture, training objectives, optimization settings, and evaluation protocols and so on in our supplementary mateirals.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in the supplemental material?

Answer: [No]

Justification: All code and data will be made available upon acceptance of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided implementation details in our main manuscript and more relevant611 information in our supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are discussed in our supplementary materials

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We discuss safeguards in the supplementary materials

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new dataset and codebase as part of this work. Due to ongoing internal approval, the full release will be made publicly available upon paper acceptance.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We collected videos from platforms that explicitly grant permission for research usage. The videos do not contain personally sensitive or private content, and are used in accordance with each platform's terms of service. No direct interaction or compensation with individuals was involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This study does not involve any direct interaction with human subjects. All data were collected from publicly available sources

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We used a large language model solely for minor linguistic improvements.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.