Making Visual Dialogue More Engaging: A New Task, Method, and Metric

Guanghui Ye[1], Huan Zhao[1]*, Yingxue Gao[1], Zhixue Zhao[2], Kehan Wang[1], Xupeng Zha[1], Zhihua Jiang[3]*

[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]Department of Computer Science, University of Sheffield, UK
[3]Department of Computer Science, Jinan University, China
hzhao@hnu.edu.cn, tjiangzhh@jnu.edu.cn

Abstract

Large language model (LLM)-based visual dialogue (VD) systems have made response generation for image-grounded conversations more correct and coherent. However, user engagement - the extent to which a user is *interested*, *emotionally involved*, and *willing* to continue the conversation - remains a challenge. To fully explore engaging VD, we propose: (i) a new task named **A**udio-enhanced **VD** (**AVD**), which introduces additional audio dialogue contexts that can more vividly convey the speaker's emotions as input, with the aim of generating correct but more engaging dialogue responses. Specifically, we employ a text-to-speech model as the modality translator to generate the paired acoustic utterances from the inputting textual utterances; (ii) an accompanying approach named **V**isually-grounded and **I**nterleaved **T**ext-**A**udio **D**ialogue **M**odeling (**VITA-DM**), which utilizes both image-grounded information and interleaved text-audio utterances for visual dialogue modeling, differentiating from previous multi-modal LLM (MLLM)-based methods that normally model text and audio modalities separately. We also present three pre-training tasks to better learn multi-modal interactions across language, vision, and audio; (iii) a novel metric named **M**ulti-**M**odal **E**ngagement (**MME**), which fills the gap of engagement estimation in VD and can provide a fine-grained assessment along emotional, attentional, and reply engagement dimensions (EE, AE, RE). We experiment on two popular datasets and provide extensive evaluations (automatic, engagement-specific, and human), supporting the validity of our approach. Furthermore, based on empirical results that reveal that emotions contribute the most to engagement, we justify our emphasis on the emotional aspect throughout the definition, solution, and evaluation of our task.

Introduction

Visual dialogue (VD) (Feng et al. 2023; Yoon et al. 2024; Abdessaied et al. 2025; Cai, Han, and Wang 2025) involves "understanding" the dialog history (what has been discussed previously), in addition to grounding information in a given image, to generate a correct and fluent response. Existing methods focus on exploring the bi-modal interactions between text and images. For example, VLAW-MDM (Lee et al. 2023) generated captions for images as input and relied on the underlying BlenderBot (400M) (Roller et al. 2021). BI-MDRG (Yoon et al. 2024) considered response generation paths to enhance both relevance and consistency and used OpenFlamingo (4B) (Awadalla et al. 2023) as response generator. However, as exemplified in Fig.1 (left), although the responses generated by these methods can be fluent and relevant to the content of the image, they struggle to be engaging, which can hinder the speakers from being interesting for opening a next dialogue turn (Palmer et al. 2025).

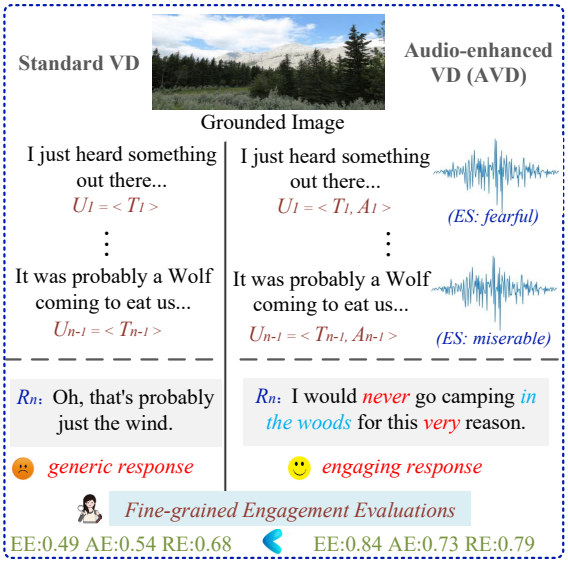

Figure 1: Task comparison. Standard VD incorporates a background image and a multi-turn textual dialogue, while our AVD task first generates synthetic audio and then rebuilds an interleaved text-audio dialogue. We also introduce a MME metric, providing multimodal engaging evaluations. As the example shows, the right response generated by our VITA-DM method is more engaging than the left response.

Engaging dialogue is a key aspect of a chatbot's social intelligence, often gauged by its ability to hold a coherent and entertaining conversation (Ferron et al. 2023; Li et al. 2024a; Palmer et al. 2025). However, there is no standard definition of engagement due to the subjectivity of human judgments. In summary, a reply is engaging when it gets the user's at-

*Corresponding authors.

tention, interests the user, or incites participation from the user (Zhang et al. 2024). For pure-text dialogue, there have been a variety of studies on generating or measuring engaging responses, e.g., WeSEE (Jiang, Vakulenko, and de Rijke 2023). However, for VD, it is underexplored to generate engaging responses, due to the abstract and extensive definitions of engagement measurement and the complexity of understanding the multi-modal content (Kumar et al. 2024).

Emotional engagement has been shown to play a leading role in all kinds of engagement measurement (Xu et al. 2022; Sun, Li, and Peng 2024; Chang and Ko 2025; Xu, Gan, and Jin 2025). In VD, as a first study, the Image-Chat benchmark (Shuster et al. 2020) collected engaging human-human conversations, where speakers are asked to play roles, given a pre-defined emotional style (e.g., "Fearful" and "Miserable" as shown in Fig.1). However, existing VD methods such as VLAW-MDM and BI-MDRG treat these emotional labels as plain text and simply append them at the end of the corresponding dialogue context input, underestimating the impact of the emotional aspect on response generation, particularly in terms of engagement evaluation. *This motivates us to represent emotions more explicitly and feasibly for multi-modal dialogue modeling* so that a visual dialogue system can better capture subtle emotional interactions in conversation turns and thus generate emotionally richer responses. Furthermore, recent studies (Li et al. 2024b; Zha, Zhao, and Zhang 2024; Xu, Jia, and Zhou 2025) show that the absence of audio makes it difficult to identify real emotions, resulting in inappropriate responses in multi-modal scenarios. *This further inspires us to explicitly represent emotions as new modalities such as speech*, which contains non-verbal factors such as tone, speed, and volume that can more vividly convey the speaker's emotions.

As Fig.1 (right) shows, we propose a new task named **A**udio-enhanced **VD** (**AVD**), which aims to generate correct but more engaging responses for VD. We construct new AVD datasets based on existing VD datasets. Specifically, we first employ a sentiment-aware text-to-speech (TTS) model, Parler-TTS (Lyth and King 2024), to explicitly generate acoustic contexts, and then align them with textual contexts in an interleaved form, which we call *cross-modal dialogue turns*. Next, we propose an accompanying approach named **V**isually-grounded and **I**nterleaved **T**ext-**A**udio **D**ialogue **M**odeling (**VITA-DM**), to model the resulting tri-modal dialogue sequence (*a grounding image, plus the interleaved textual and acoustic contexts*). Particularly, we introduce three pre-training tasks: image and caption alignment (ICA), cross-modal contrastive learning (CCL), and next utterance prediction (NUP), to better learn multi-modal interactions. We implement VITA-DM using Llama-2-7B (Touvron et al. 2023) as the backbone. Subsequently, we propose a metric named **M**ulti-**M**odal **E**ngagement (**MME**) to assess the engagement of VD models. By fine-tuning the multi-modal Llama-3.2 (11B), we introduce three sub-dimensions: *emotional engagement (EE), attentional engagement (AE), and reply engagement (RE)*, to provide fine-grained engagement evaluations. We perform extensive experiments sourced from two popular VD benchmarks: Image-Chat (Shuster et al. 2020) and Photo-Chat (Zang et al. 2021). We compare our method with recent state-of-the-art VD models and strong multi-modal LLMs (MLLMs). Both automatic and human evaluations demonstrate the effectiveness and superiority of our method.

Our work addresses an underexplored yet impactful aspect of VD systems: *user engagement*. We pose comprehensive contributions across task definition, dataset construction, modeling, evaluation for the multimedia community.

- We introduce a novel task, AVD, along with audio-enhanced visual dialogue datasets, which offers a new perspective for multi-modal dialogue systems to generate engaging responses while performing well in grammar.
- We propose an accompanying approach, VITA-DM, which utilizes image-grounded information and interleaved text-audio utterances for tri-model dialogue modeling, for guiding better interplay between text and audio.
- We build a novel metric, MME, which can appropriately evaluate the response's engagement in terms of diverse aspects by interacting with MLLM as the evaluator.
- We conduct detailed and solid experiments to evaluate all different levels of contribution of additional audio modalities, showing performance advantages of our approach over prior top methods and strong MLLMs.

The AVD Task

Formally, the VD task aims to generate an appropriate textual response $\mathbf{R}_n = \{w_{n,1}, \ldots, w_{n,k}\}$ conditioned on multi-modal content including a background image $\mathbf{V}$ and a dialogue history $\mathbf{H}_{<n} = \{\mathbf{T}_1^U, \mathbf{T}_2^S, \ldots, \mathbf{T}_{n-1}^U\}$, where $\mathbf{U}$ and $\mathbf{S}$ denote different dialogue roles such as user and system. The target function can be defined to maximize:

$$T_{VD} = \prod_{j=1}^{k} p(w_{n,j}|\mathbf{V}, \mathbf{T}_{<n}^U, \mathbf{T}_{<n}^S, w_{n,<j}; \theta) \quad (1)$$

where θ denotes the set of model parameters.

Audio Generation and Dataset Construction: Existing tri-modal datasets involving language, vision, and audio have been proposed for multi-modal emotion recognition (MER) or sentiment analysis (MSA) (Huang et al. 2024a,b). Their target is to predict a correct sentiment label rather than generating a proper response. In a conversation sample, there are often multiple images (e.g., video frames) for MER/MSA, while there is only one grounding image for our task. Thus, such datasets are fundamentally ill-suited for our task. We implement tri-modal construction based on existing VD datasets. Specifically, we adopt Parler-TTS, a reproduction of the work of (Lyth and King 2024), to generate a paired acoustic utterance $\mathbf{A}$ from each textual utterance $\mathbf{T}$. Parler-TTS is a lightweight TTS model that can generate high-quality, natural sounding speech in the style of a given speaker (gender, pitch, speaking style, etc.). Regarding emotion labels $\mathbf{E}$, there are two cases: (1) If the original dataset (e.g., Image-Chat) has provided $\mathbf{E}$, we use them directly; (2) Otherwise (e.g., Photo-Chat), we call GPT-4 to perform sentiment analysis on $\mathbf{T}$ to predict $\mathbf{E}$. Both $\mathbf{T}$ and $\mathbf{E}$ (e.g., "Fearful") are then fed to Parler-TTS which can generate $\mathbf{A}$ (we

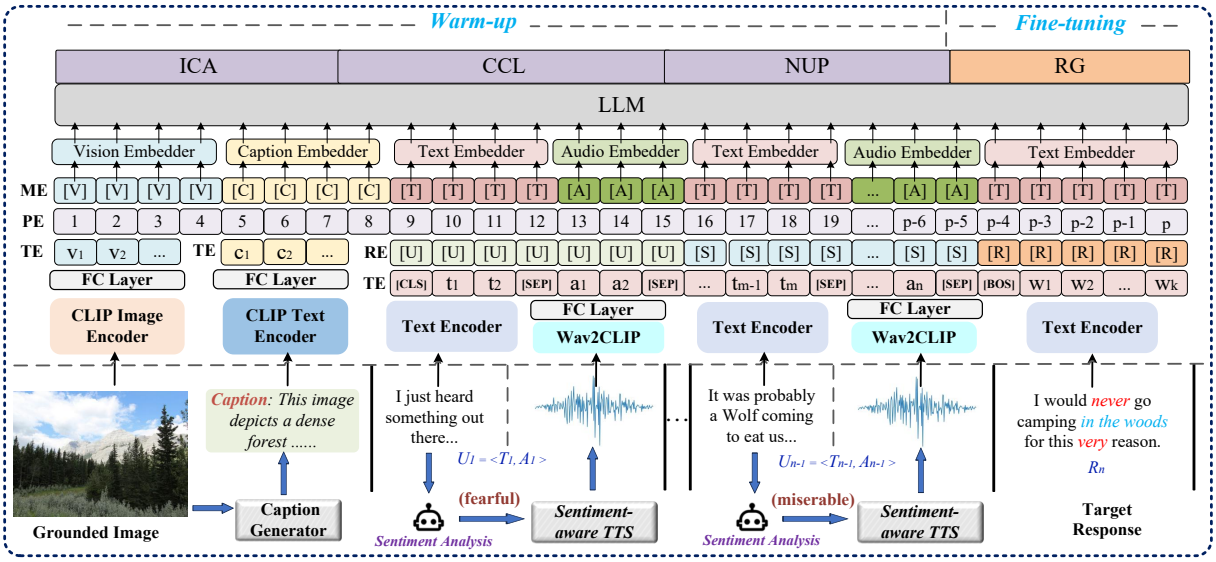

Figure 2: The architecture of VITA-DM. Different from existing methods which encode text and audio organized in a single modality (e.g., [T][T][A][A]), our method encodes the paired text-audio sequence using cross-modal turns (e.g., [T][A][T][A]).

use a TTS prompt such as "The speaker is currently [*fearful*] and the recording is close with little background noise.") :

$$\mathbf{A} = TTS(\mathbf{T}, \mathbf{E}) \tag{2}$$

Subsequently, we align $\mathbf{A}$ with $\mathbf{T}$ following the original temporal dialogue sequence, thus generating our desired datasets which have unique cross-model dialogue turns.

Task Definition: We add a paired audio for each text in the dialogue history $\mathbf{H}_{<n}$. Thus, for each $\mathbf{U}_i$ and $\mathbf{S}_i$, we denote $\mathbf{U}_i = < \mathbf{T}_i^U, \mathbf{A}_i^U >$ and $\mathbf{S}_i = < \mathbf{T}_i^S, \mathbf{A}_i^S >$ after introducing aligned text-audio pairs. We also provide $\mathbf{E}$ as an alternative of $\mathbf{A}$ for some methods that cannot address the audio input directly. Thus, our proposed task is defined to maximize:

$$T_{AVD} = \prod_{j=1}^{k} p(w_{n,j} | \mathbf{V}, \mathbf{E}, \mathbf{T}_{<n}^U, \mathbf{A}_{<n}^U, \mathbf{T}_{<n}^S, \mathbf{A}_{<n}^S, w_{n,<j}; \theta) \tag{3}$$

The VITA-DM Method

Fig.2 exemplifies the architecture of VITA-DM which uses modality-specific encoders (frozen) and aligns them with the LLM through projection layers (learnable). *The entire process of VITA-DM* is as follows: (1) We first generate descriptive captions about images, which offers an effective semantic supplement to visual content; (2) We then extract features of each modality data in the input; (3) Next, we pre-train LLM (i.e., the backbone model) with three proposed warm-up tasks which address distinct modality interactions; (4) Last, we fine-tune the pre-trained model for better response generation. We elaborate each of these steps below.

(1) Image Captioning Image captioning, which can capture crucial information about the image, has been shown to be effective as an enhancement method (Lee et al. 2023; Liu et al. 2024). Specifically, we employ a large vision-language model, BLIP-2 OPT (6.7B) (Li et al. 2023), to generate textual caption $\mathbf{C} = \{c_1, c_2, \ldots, c_m\}$ about $\mathbf{V}$.

(2) Multi-modal Feature Extraction We first generate *token embedding* (TE) via a uni-modal encoder (frozen). Subsequently, we use *modality embedding* (ME) to distinguish between different modalities of the input and *role embedding* (RE) to distinguish between the role of text/audio utterances. Finally, we employ *positional embedding* (PE) to utilize the sequence's ordering. Embedder adopts the *sum* of TE, ME, RE, and PE to represent each token.

(i) Textual Modality: Each $\mathbf{T} = \{t_1, t_2, \ldots, t_{|T|}\}$ is tokenized with the byte-pair encoding (BPE). Then, we can obtain the embedding sequence $E^T = \{E_1^T, E_2^T, \ldots, E_{|T|}^T\} \in \mathbb{R}^{|T| \times d_t}$, where d_t is the dimension of the text embedding. Finally, E^T is fed into the Text Embedder for integrating with other embeddings including E^M, E^P, and E^R.

$$\mathcal{E}^T = Embedder_{text}(E^M, E^P, E^T, E^R) \tag{4}$$

(ii) Visual Modality: We adopt the CLIP Image Encoder (Radford et al. 2021) to extract d_v-dimensional visual features as $F^V = \{F_1^V, F_2^V, \ldots, F_{|F^V|}^V\} \in \mathbb{R}^{|F^V| \times d_v}$ from $\mathbf{V}$. The visual features obtained have dimensions different from the textual representation E^T, so an additional fully connected (FC) layer is used to put the visual features as input for projection. Then, we can obtain the embedding sequence $E^V = \{E_1^V, E_2^V, \ldots, E_{|F^V|}^V\} \in \mathbb{R}^{|F^V| \times d_t}$. Finally, E^V is fed into Visual Embedder to integrate other embeddings:

$$\mathcal{E}^V = Embedder_{vision}(E^M, E^P, E^V) \tag{5}$$

For a generated caption $\mathbf{C} = \{c_1, c_2, \ldots, c_{|C|}\}$, we obtain $E^C = \{E_1^C, E_2^C, \ldots, E_{|C|}^C\} \in \mathbb{R}^{|C| \times d_t}$ using the CLIP Text Encoder. Then, E^C is fused with E^M and E^P:

$$\mathcal{E}^C = Embedder_{text}(E^M, E^P, E^C) \tag{6}$$

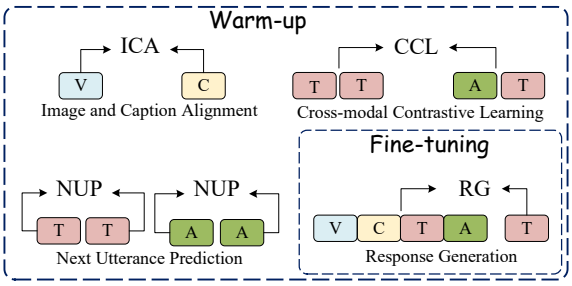

Figure 3: Warm-up tasks in VITA-DM.

(iii) Acoustic Modality: We employ Wav2CLIP (Radford et al. 2021) to extract d_a-dimensional audio features as $F^A = \{F_1^A, F_2^A, \ldots, F_{|F^A|}^A\} \in \mathbb{R}^{|F^A| \times d_a}$. We obtain $E^A = \{E_1^A, E_2^A, \ldots, E_{|F^A|}^A\} \in \mathbb{R}^{|F^A| \times d_t}$ via feeding F^A into an FC layer for projection. Then, $\mathcal{E}^A$ is derived by:

$$\mathcal{E}^A = Embedder_{audio}(E^M, E^P, E^A, E^R) \quad (7)$$

(3) Tri-modal Dialogue Modeling and Pretraining Following the turns of the temporal dialogue sequence (a grounding image plus paired text-audio utterances), we concatenate ($\oplus$) all embeddings and feed the result to LLM.

$$\mathcal{E}^V \oplus \mathcal{E}^C \oplus \mathcal{E}_1^T \oplus \mathcal{E}_1^A \oplus \mathcal{E}_2^T \oplus \mathcal{E}_2^A \oplus \ldots \oplus \mathcal{E}_{n-1}^T \oplus \mathcal{E}_{n-1}^A \oplus \mathcal{E}_n^R \quad (8)$$

We use three (warm-up) pre-training tasks to train the model, as illustrated in Fig.3.

(i) Image and Caption Alignment (ICA): We train the model to generate the target caption $\mathbf{C} = \{c_1, c_2, \ldots, c_m\}$ based on $\mathbf{V}$ by minimizing the negative log-likelihood loss:

$$\mathcal{L}_{ICA} = -\sum_{i=1}^{m} \log p(c_i | \mathbf{V}, c_{<i}; \theta) \quad (9)$$

(ii) Cross-modal Contrastive Learning (CCL): We employ contrastive learning (CL) to guide the better interplay between text and audio. We view the text and audio in the same utterances as positive pairs, and those in the different utterances as negative pairs. The CCL loss is defined as:

$$\mathcal{L}_{CCL} = -\sum_{i=1}^{B} \log \frac{f(\mathbf{T}_i, \mathbf{A}_i)}{\sum_{j=1}^{B} f(\mathbf{T}_i, \mathbf{A}_j)} \quad (10)$$

where B is the batch size. The function f calculates the correlation between text and audio as follows:

$$f(\mathbf{T}, \mathbf{A}) = \exp((\mathcal{E}^T \cdot \mathcal{E}^A)/\tau) \quad (11)$$

where the inner product of $\mathcal{E}^T$ (Eq.4) and $\mathcal{E}^A$ (Eq.7) is calculated and τ is the temperature parameter.

(iii) Next Utterance Prediction (NUP): We propose NUP, assuming that two consecutive textual or acoustic utterances exhibit a certain degree of coherence. Specifically, when choosing U_i and U_j as a training example, 50% of the time U_j is the actual next utterance that follows U_i, and 50% of the time it is a random sampled utterance from the training corpus. Then, we train two binary classifiers $\text{NUP}_t(\mathbf{T}_i, \mathbf{T}_j)$ and $\text{NUP}_a(\mathbf{A}_i, \mathbf{A}_j)$ using the Binary Cross-Entropy (BCE) loss. Provided that the textual/acoustic NUP loss is $\mathcal{L}_{NUP-t}/\mathcal{L}_{NUP-a}$, the final loss $\mathcal{L}_{NUP}$ is obtained:

$$\mathcal{L}_{NUP} = \frac{1}{2}(\mathcal{L}_{NUP-t} + \mathcal{L}_{NUP-a}) \quad (12)$$

Finally, the overall multitask pre-training loss ($\mathcal{L}$) can be formulated as follows:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{ICA} + \lambda_2 \cdot \mathcal{L}_{CCL} + \lambda_3 \cdot \mathcal{L}_{NUP} \quad (13)$$

where λ_i (i=1,2,3) is a configurable hyper-parameter.

(4) Fine-tuning for Response Generation During training, we feed the target response as input and employ the auto-regressive mechanism. After warm-ups, we fine-tune the resulting model for RG. During inference, the [BOS] is used to generate target words. The fine-tuning loss ($\mathcal{L}'$) is:

$$\mathcal{L}' = \sum_{j=1}^{K} -\log p(w_j | \mathbf{V}, \mathbf{C}, \mathbf{T}_{<n}^U, \mathbf{A}_{<n}^U, \mathbf{T}_{<n}^S, \mathbf{A}_{<n}^S, w_{<j}; \theta) \quad (14)$$

The MME Metric

We propose the MME metric that measures the engagement of a generated response conditioned on the given multi-modal content (i.e., in turn-level). An engagement evaluation in dialog-level (i.e., the evaluation object is a complete interactive dialogue process) can be simulated by taking the average over all turns. *MME enjoys unique advantages such as mitigating modal misalignment issues*, which makes it differentiate from previous engagement metrics. The input to MME includes the grounded image, emotion labels, the textual dialogue history, and the response to be assessed. For implementation, we fine-tune the multi-modal Llama-3.2 (11B) model on the Reddit-based Engagement Dataset (RED) (Xu et al. 2022) with 80k labeled single-turn conversations. Each sample in RED is labeled with engaging or non-engaging in terms of different aspects including EE, AE, and RE. Finally, MME predicts an overall engagement score (OE), plus three sub-dimension (EE, AE, RE) scores:

- *Emotional Engagement (EE)*: whether the response is emotionally resonant with the multi-modal content?
- *Attentional Engagement (AE)*: whether the response itself is interesting, diverse, and informative?
- *Reply Engagement (RE)*: whether the response is coherent and relevant with a dialogue context?
- *Overall Engagement (OE)*: whether the response is engaging from an overall perspective?

We will explore the rationality of MME from *weight analysis* and *correlation analysis*. The experimental results (see Tables 7 and 8) can justify our emphasis on the emotional aspect, showing that our definition of MME is reliable.

Experiment Settings

Datasets. We generate two new datasets Image-Chat$_{AVD}$ and Photo-Chat$_{AVD}$ sourced from Image-Chat (Shuster et al. 2020) and Photo-Chat (Zang et al. 2021), respectively. We train VITA-DM on training sets and evaluate the resulting model on testing sets. Image-Chat collects 202k engaging human-human conversations, where speakers are asked to play roles given an emotional style (215 totally). Photo-Chat contains 12k dialogues, each of which is paired with

a photo. Unlike Image-Chat, Photo-Chat does not provide emotional labels. Table 1 lists the key statistics.

Datasets	Attributes	Train	Valid	Test	Total
Image-Chat$_{AVD}$	#Image/Dialogue	186,782	5,000	9,997	201,779
	#T/A-utterance	355,862	15,000	29,991	400,853
Photo-Chat$_{AVD}$	#Image/Dialogue	10,286	1,000	1,000	12,286
	#T/A-utterance	130,546	12,701	12,852	156,099

Table 1: Statistics of Image-Chat$_{AVD}$ and Photo-Chat$_{AVD}$.

Settings. Experimental details are as follows. For $\mathcal{L}$ (Eq. 13), we set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ to treat tasks equally. Regarding the dimensions of the features, we set d_v, d_a, d_t to 512, 512, 4096. We employ Adam (Loshchilov and Hutter 2019) to optimize the model's parameters. For pre-training, the learning rate, batch size and epoch are set as 2e-4, 64 and 20. For fine-tuning, the above three parameters are set as 1e-4, 32 and 30. During inference, we use the beam search with a width of 3 to generate target responses. All experiments were conducted on 8 NVIDIA 3090 24GB GPUs.

Compared Methods. Compared methods can be categorized into three groups: (1) Base dialogue systems that have been trained to generate fluent responses, including DialoGPT (Zhang et al. 2020), BlenderBot (Roller et al. 2021), and MM-BlenderBot (Shuster et al. 2021); (2) Open-source MLLMs that improve the capabilities of LLMs by integrating data from multiple modalities, including Qwen-VL-Chat (9B) (Bai et al. 2023), mPLUG-Owl (8B) (Ye et al. 2023), LLaVA-1.5 (8B) (Liu et al. 2024), and Emotion-LLaMA (7B) (Cheng et al. 2024); (3) SOTA VD methods that can generate better responses conditioned on both the image content and the dialogue history, including Divter (Sun et al. 2022), VLAW-MDM (Lee et al. 2023), and BI-MDRG (Yoon et al. 2024). We did not use closed-source MLLMs such as GPT-4V, due to expensive APIs.

Evaluation Metrics. We perform a comprehensive set of evaluations. (i) *Automatic evaluation.* Following (Lee et al. 2023; Yoon et al. 2024), we adopt three rules-based grammar metrics: BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and CIDEr (Vedantam, Zitnick, and Parikh 2015), which are all based on the word-level overlap ratio between the hypotheses and the ground truths. (ii) *Engagement-specific evaluation.* We train a Llama-3.2-based metric model, MME, to predict the level of engagement in multimodal scenarios. (iii) *Human evaluation.* We also provide human evaluations in terms of *engaging, diverse, and correct*, as a complement to automatic evaluations.

Results and Analysis

Approach Instantiation We implement VITA-DM using three distinct backbone models: GPT-2-Medium (345M) (Radford et al. 2019), GPT-2-XL (1.5B) (Radford et al. 2019), and Llama-2 (7B) (Touvron et al. 2023). In Table 2, the results validate the universality of VITA-DM, which can work in different-size pre-training models and significantly improve the performance of backbones. We also observe that the improvement will sustain as the model size grows. We consider *VITA-DM with Llama-2-7B as our default model*.

(a) The Image-Chat$_{AVD}$ dataset

Models	BLEU-1	BLEU-4	ROUGE-1	ROUGE-L	CIDEr
GPT-2-M-345M	8.01	0.72	9.24	9.25	0.85
GPT-2-M-345M (♠)	**10.34**$^{\uparrow 2.33}$	**1.03**$^{\uparrow 0.31}$	**10.97**$^{\uparrow 1.73}$	**11.26**$^{\uparrow 2.01}$	**0.94**$^{\uparrow 0.09}$
GPT-2-XL-1.5B	8.28	0.77	9.79	9.26	0.93
GPT-2-XL-1.5B(♠)	**11.51**$^{\uparrow 3.23}$	**1.22**$^{\uparrow 0.45}$	**12.36**$^{\uparrow 2.57}$	**12.84**$^{\uparrow 3.58}$	**1.12**$^{\uparrow 0.19}$
Llama-2-7B	10.23	1.09	10.94	11.05	1.05
Llama-2-7B(♠)	**11.87**$^{\uparrow 1.64}$	**1.31**$^{\uparrow 0.22}$	**14.05**$^{\uparrow 3.11}$	**13.96**$^{\uparrow 2.91}$	**1.28**$^{\uparrow 0.23}$

(b) The Photo-Chat$_{AVD}$ dataset

Models	BLEU-1	BLEU-4	ROUGE-1	ROUGE-L	CIDEr
GPT-2-M-345M	10.01	0.79	10.08	10.18	0.90
GPT-2-M-345M (♠)	**11.79**$^{\uparrow 1.78}$	**1.04**$^{\uparrow 0.25}$	**11.49**$^{\uparrow 1.41}$	**11.02**$^{\uparrow 0.84}$	**1.06**$^{\uparrow 0.16}$
GPT-2-XL-1.5B	10.12	0.91	10.43	10.29	1.03
GPT-2-XL-1.5B(♠)	**13.09**$^{\uparrow 2.97}$	**1.35**$^{\uparrow 0.44}$	**12.94**$^{\uparrow 2.49}$	**11.82**$^{\uparrow 1.53}$	**1.15**$^{\uparrow 0.12}$
Llama-2-7B	11.82	1.03	11.87	11.45	1.08
Llama-2-7B(♠)	**13.52**$^{\uparrow 1.70}$	**1.54**$^{\uparrow 0.51}$	**13.47**$^{\uparrow 1.60}$	**12.36**$^{\uparrow 0.91}$	**1.24**$^{\uparrow 0.16}$

Table 2: Automatic evaluation results (%). "♠" indicates that the model is trained using VITA-DM. The **improvement** is statistically significant (t-test with p-value<0.01).

Approach Superiority We compare VITA-DM with three categories of methods, including base dialogue models, strong MLLMs, and SOTA VD methods. We reproduce these methods by using their released codes.

(a) The Image-Chat$_{AVD}$ dataset

Method Categories	BLEU-1↑	BLEU-4↑	ROUGE-1↑	ROUGE-L↑	CIDEr↑
Base Dialogue Systems					
DialoGPT	7.92	0.77	9.41	9.15	0.83
BlenderBot	8.20	0.75	9.40	9.18	0.85
MM-BlenderBot	8.27	0.75	9.53	9.42	0.86
Open-source MLLMs					
Qwen-VL-Chat	11.21	1.15	12.28	12.55	1.07
MPLUG-Owl	10.87	1.05	11.68	11.91	0.98
LLaVA-1.5	10.76	0.94	11.35	11.23	1.01
Emotion-LLaMA	11.23	1.14	12.29	12.50	1.04
Top VD Methods					
Divter	8.31	0.78	9.76	9.35	0.90
VLAW-MDM	9.40	0.89	10.82	12.50	0.96
BI-MDRG	10.83	1.07	11.74	10.98	1.05
***Our Models** (using GPT-2-XL-1.5B or Llama-2-7B)*					
VITA-DM (1.5B)	11.51	1.22	12.36	12.84	1.12
VITA-DM (7B)	**11.87**	**1.31**	**14.05**	**13.96**	**1.28**

(b) The Photo-Chat$_{AVD}$ dataset

Method Categories	BLEU-1↑	BLEU-4↑	ROUGE-1↑	ROUGE-L↑	CIDEr↑
Base Dialogue Systems					
DialoGPT	10.37	0.91	10.58	10.50	0.88
BlenderBot	10.54	0.89	10.59	10.36	0.83
MM-BlenderBot	11.06	0.96	11.05	10.78	0.90
Open-source MLLMs					
Qwen-VL-Chat	12.94	1.20	12.56	11.37	1.13
MPLUG-Owl	12.42	1.12	11.87	10.98	1.10
LLaVA-1.5	12.29	1.09	12.07	10.95	1.03
Emotion-LLaMA	13.01	1.22	12.77	11.43	1.10
Top VD Methods					
Divter	11.24	0.95	10.96	10.84	0.94
VLAW-MDM	11.98	1.09	11.43	10.90	1.06
BI-MDRG	12.43	1.18	12.10	11.18	1.12
***Our Models** (using GPT-2-XL-1.5B or Llama-2-7B)*					
VITA-DM (1.5B)	13.09	1.35	12.94	11.82	1.15
VITA-DM (7B)	**13.52**	**1.54**	**13.47**	**12.36**	**1.24**

Table 3: Main comparisons. Best results regarding each metric (each column) are **in bold** and best results regarding each method category (each block) are underlined, respectively.

Notably, some LLM-based methods cannot directly address the audio input. Thus, we introduce AVD', i.e., replacing **A** with **E**, as shown in Eq.15. In addition, we use BLIP-2 to generate image captions as an alternative to the input image for language-only models such as DialoGPT.

$$T_{AVD'} = \prod_{j=1}^{k} p(w_{n,j}|\mathbf{V}, \mathbf{T}^U_{<n}, \mathbf{E}^U_{<n}, \mathbf{T}^S_{<n}, \mathbf{E}^S_{<n}, w_{n,<j}; \theta) \quad (15)$$

The results in Table 3 show that: (i) Base systems achieve poor results compared to the rest of methods; (ii) Emotion-LLaMA (7B) performs best in the MLLM group, due to integrating audio, visual, and textual inputs through emotion-specific encoders; (iii) BI-MDRG outperforms other VD methods; (iv) *Our models (1.5B and 7B) consistently outperform all compared methods on two datasets.* Our 1.5B version can surpass Emotion-LLaMA (7B) in all aspects.

Engagement Assessment We use MME to evaluate the responses generated by four VD methods on the test sets of both Image-Chat$_{AVD}$ and Photo-Chat$_{AVD}$. Specifically, given the multi-modal content (image **I**, emotion **E**, context **C**), MME can output an overall score OE, plus three sub-dimension scores EE, AE, and RE, for the response **R**. We report the normalized scores in the range [0,1] (the higher, the better) in Table 4. We observe that *our responses are better than the others in terms of all engagement aspects (EE, AE, RE, OE)*. In particular, *our responses achieve a significantly higher score on EE*, showing the benefits of adding sentiment-aware speech signals. Additionally, *our responses also improve both AE and RE*, indicating the effectiveness of our encoding method. In summary, the idea of integrating text and audio utterances in pairs and following the original dialogue turns facilitates all kinds of cross-modal interactions. The evaluation results (in turn-level) in Table 4 validate *the generalization of MME across different datasets*.

(a) The Image-Chat$_{AVD}$ dataset				
VD methods	EE↑	AE↑	RE↑	OE↑
Divter	0.55	0.53	0.44	0.47
VLAW-MDM	0.62	0.52	0.50	0.49
BI-MDRG	0.59	0.55	0.58	0.53
VITA-DM (7B) (ours)	**0.81**	**0.66**	**0.65**	**0.66**
(b) The Photo-Chat$_{AVD}$ dataset				
VD methods	EE↑	AE↑	RE↑	OE↑
Divter	0.48	0.50	0.46	0.47
VLAW-MDM	0.59	0.56	0.53	0.54
BI-MDRG	0.60	0.61	0.52	0.58
VITA-DM (7B) (ours)	**0.76**	**0.67**	**0.65**	**0.70**

Table 4: Engagement scores generated by the MME metric.

Human Evaluations We employ five trained annotators (Ph.D. students and NLP experts) to assess the quality of the generated responses on the Image-Chat$_{AVD}$ test set. Totally, the responses of 200 random samples per each model are rated by each participant, using the 5-Point Likert scale (the higher, the better). We consider three key dimensions: (i) *Emotionally Engaging*: whether the response conveys the right emotion with regards to conversation roles? (ii) *Diverse*: whether the response is diverse or specific? (3) *Correct*: whether the response is correct conditioned on the grounded image and previous conversations? For each response, we average the scores from five participants in terms of a specific dimension. Then, we average all the responses to express the model's performance. As shown in Fig.4, *human evaluation results further confirm that our approach VITA-DM can generate emotionally more appropriate responses compared to other VD methods*, while maintaining superiority in terms of diversity and correctness.

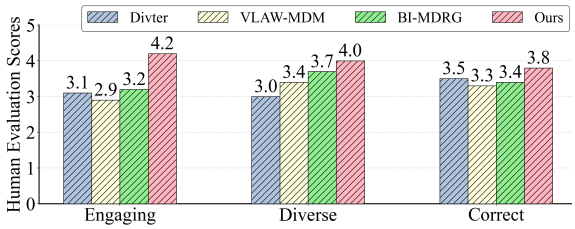

Figure 4: Human evaluation on generated responses.

Ablation Studies We perform ablation studies on the Image-Chat$_{AVD}$ test set. The experiments encompass two parts: (a) ablation on modalities and (b) ablation on pre-training tasks. Specifically, in the w/o A setting, we use ICA and only NUP$_t$; in the w/o T setting, we use ICA and only NUP$_a$; in the w/o C setting, we use CCL and both NUP$_t$ and NUP$_a$. In Table 5, we find that: (i) The model's performance can drop if we remove any modality (T, A, C) or any task (ICA, CCL, NUP), indicating that *each component has a positive contribution*; (ii) The ablation variant w/o T degrades most, showing that *text is most important* for a model to generate responses while only image or audio is not sufficient; (iii) The model's performance will decrease if audio is canceled, suggesting that *generating speech from text provides substantial additional information for the model*.

(a) Ablation on modalities (T: Text; A: Audio; C: Caption)					
Model	BLEU-1↑	BLEU-4↑	ROUGE-1↑	ROUGE-L↑	CIDEr↑
VITA-DM (7B)	**11.87**	**1.31**	**14.05**	**13.96**	**1.28**
w/o A	10.88	1.15	11.96	12.20	1.08
w/o C	10.34	1.09	11.61	11.59	1.03
w/o A+C	8.41	0.88	9.90	9.58	0.94
w/o T	7.98	0.80	9.54	9.31	0.86
(b) Ablation on pre-training tasks (in the T+A+C setting)					
Model	BLEU-1↑	BLEU-4↑	ROUGE-1↑	ROUGE-L↑	CIDEr↑
VITA-DM (7B)	**11.87**	**1.31**	**14.05**	**13.96**	**1.28**
w/o ICA	10.35	1.01	11.65	11.93	0.98
w/o CCL	11.10	1.12	11.74	11.98	1.09
w/o NUP	11.45	1.20	12.13	12.64	1.20

Table 5: Results of ablation studies. w/o denotes without.

In-depth Analysis

Comparison of Emotion Audio and Emotion Label We first validate the necessity of incorporating emotion audio. Specifically, we replace audio utterances **A** with emotional labels **E** (Eq.16) in our method. To do this, we remove Wav2CLIP, CCL, NUP$_a$, and train the tailored VITA-DM in the corpus. The results in Table 6 show that VITA-DM (audio) leads to a higher performance than VITA-DM (label). We also compare these two settings when removing all pre-training tasks. We find that VITA-DM (audio) actually performs worse than VITA-DM (label) in this test. This shows that the observed performance gains benefit not only from representing emotion in a more appropriate modality but also from introducing better cross-modal interactions.

$$\mathcal{E}^V \oplus \mathcal{E}^C \oplus \mathcal{E}_1^T \oplus \mathcal{E}_1^E \oplus \ldots \oplus \mathcal{E}_{n-1}^T \oplus \mathcal{E}_{n-1}^E \oplus \mathcal{E}_n^R \qquad (16)$$

Weight Analysis of Engagement Dimensions We explore the rationality of MME from two aspects: *the weight*

(a) The Image-Chat$_{AVD}$ dataset					
Model	BLEU-1↑	BLEU-4↑	ROUGE-1↑	ROUGE-L↑	CIDEr↑
VITA-DM (label)	11.19	1.16	12.95	12.36	1.10
VITA-DM (audio)	**11.87**	**1.31**	**14.05**	**13.96**	**1.28**
(b) The Photo-Chat$_{AVD}$ dataset					
Model	BLEU-1↑	BLEU-4↑	ROUGE-1↑	ROUGE-L↑	CIDEr↑
VITA-DM (label)	12.71	1.23	12.54	11.79	1.09
VITA-DM (audio)	**13.52**	**1.54**	**13.47**	**12.36**	**1.24**

Table 6: Comparison of using emotional label and audio.

analysis of sub-dimensions and *the correlation analysis with human ratings*. Firstly, we hypothesize that the OE can be derived from EE, AE, and RE via linear integration:

$$S_{OE} = w_{EE} \cdot S_{EE} + w_{AE} \cdot S_{AE} + w_{RE} \cdot S_{RE} \quad (17)$$

Then, we curate a small set composed of 500 samples with human annotations from Image-Chat$_{AVD}$. Specifically, we randomly select samples and manually assign four engagement (EE, AE, RE, OE) scores, referring as human scores. Subsequently, we separately use human scores and our MME scores to learn the best weights w_{EE}, w_{AE}, and w_{RE} in Eq.17. For more comparisons, we also repeat this test, using two recent LLM-based evaluators, G-Eval (Liu et al. 2023) which integrates GPT-4, and GPTScore (Fu et al. 2024) which employs GPT-3. In Table 7. We observe that: (i) Each dimension has a positive contribution; (ii) Our weights align best with human weights; (iii) For all evaluators, EE has the highest weight, showing that it plays a major role in the evaluations, which provides further support for our work.

Weight	w_{EE}	w_{AE}	w_{RE}
Human	0.65	0.15	0.20
GPTScore	$0.58^{\Delta 0.07}$	$0.26^{\Delta 0.09}$	$0.16^{\Delta 0.04}$
G-Eval	$0.55^{\Delta 0.10}$	$0.27^{\Delta 0.12}$	$0.18^{\Delta 0.02}$
MME (ours)	$0.63^{\Delta 0.02}$	$0.14^{\Delta 0.01}$	$0.23^{\Delta 0.03}$

Table 7: Weight analysis. Numbers (marked with triangles) represent the weight differences between model and human.

Correlation with Human Ratings Secondly, we calculate the correlations with human judgments on the same sample set. Concretely, we adopt two widely-used correlation measures (Zar 2005): the Pearson correlation (r) measures the linear correlation between two sets of data and the Spearman correlation (ρ) assesses the statistical dependence between the rankings of two variables. In Table 8, we find that MME correlates most strongly with human judgments, due to effective multi-modal interactions (i.e., warm-ups).

Correlation (r/ρ) ↑	EE	AE	RE	OE
GPTScore	0.39/0.37	0.33/0.33	0.38/0.39*	0.35/0.34
G-Eval	0.51/0.48	0.44/0.43	0.40/0.41	0.46/0.44
MME (ours)	**0.69/0.66**	**0.53/0.52**	**0.50/0.50**	**0.58/0.57**

Table 8: Correlation with human scores. All numbers are statistically significant (t-test with p-value < 0.01) except for *.

The impact of TTS models To investigate the robustness of our model to audios generated by TTS systems of varying quality, we provide comparative experiments under different audio quality conditions. We use an alternative model, UniCATS (Du et al. 2024), which can generate natural and noise-free speech but not sentiment-aware. Specifically, we generate new audio from text (**A**=UniCATS(**T**)), different from our previous attempts (**A**=Parler-TTS(**T**, **E**)). The results in Table 9 show that using Parler-TTS can make our approach perform better than using UniCATS in terms of all metrics, suggesting performance advantages achieved by using a sentiment-aware TTS model such as Parler-TTS. We further check the fidelity of Parler-TTS when asked to produce audio with the intended emotion. We use SpeechGPT (Zhang et al. 2023) to predict the emotion of generated audio and then calculate the accuracy compared to its emotion label. We perform this test on Image-Chat (test) and find that our audio is scored with a high emotion accuracy of 93%.

(a) The Image-Chat$_{AVD}$ dataset					
TTS Model	BLEU-1↑	BLEU-4↑	ROUGE-1↑	ROUGE-L↑	CIDEr↑
UniCATS	10.89	1.17	12.21	12.31	1.14
Parler-TTS	**11.87**	**1.31**	**14.05**	**13.96**	**1.28**
(b) The Photo-Chat$_{AVD}$ dataset					
TTS Model	BLEU-1↑	BLEU-4↑	ROUGE-1↑	ROUGE-L↑	CIDEr↑
UniCATS	12.73	1.14	12.17	10.98	1.13
Parler-TTS	**13.52**	**1.54**	**13.47**	**12.36**	**1.24**

Table 9: Comparison of using distinct TTS models.

Qualitative Analysis

We present a case study in Fig.5. Concretely, we show the image, dialogue context, ground-truth (GT), and responses generated by four model candidates. We observe that: (i) All responses are contextually relevant (in blue) to the multi-modal input; and (ii) Our response is more informative and emotionally more appropriate (in red) with regards to GT.

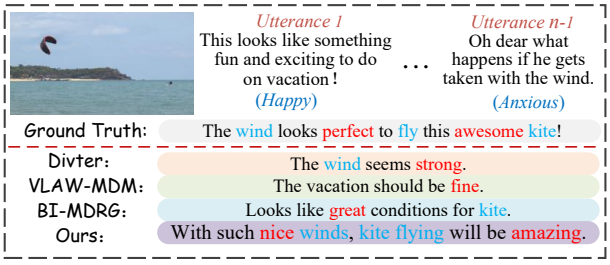

Figure 5: Case studies on generated responses.

Conclusions

Our work presents a valuable step toward building more engaging visual dialogue systems. Specifically, we propose a new task (AVD), a solution method (VITA-DM), and a novel metric (MME). Empirical results, including extensive human and automatic evaluations, support the effectiveness of our method. Future works include: (1) Due to the fact that we spent 3.6/0.3 hours to generate audio from the Image-Chat/Photo-Chat dataset, the computational trade-offs should be further explored; (2) The applicability of our method should be further examined in more challenging scenarios, such as dialogues containing human speech.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC Grant No.62076092).

References

Abdessaied, A.; Rohrbach, A.; Rohrbach, M.; and Bulling, A. 2025. V^2Dial: Unification of Video and Visual Dialog via Multimodal Experts. In *Proc. CVPR*, 8637–8647. Computer Vision Foundation / IEEE.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S. Y.; Sagawa, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *CoRR*, abs/2308.01390.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR*, abs/2308.12966.

Cai, S.; Han, X.; and Wang, S. 2025. Divide-and-Conquer: Tree-structured Strategy with Answer Distribution Estimator for Goal-Oriented Visual Dialogue. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *Proc. AAAI*, 1917–1925. AAAI Press.

Chang, Y.; and Ko, Y. 2025. Soft engagement with pseudo initiatives for multi-party dialogue generation. *Pattern Recognit. Lett.*, 191: 103–109.

Cheng, Z.; Cheng, Z.; He, J.; Wang, K.; Lin, Y.; Lian, Z.; Peng, X.; and Hauptmann, A. G. 2024. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Proc. NeurIPS*.

Du, C.; Guo, Y.; Shen, F.; Liu, Z.; Liang, Z.; Chen, X.; Wang, S.; Zhang, H.; and Yu, K. 2024. UniCATS: A Unified Context-Aware Text-to-Speech Framework with Contextual VQ-Diffusion and Vocoding. In *Proc. AAAI*, 17924–17932.

Feng, J.; Sun, Q.; Xu, C.; Zhao, P.; Yang, Y.; Tao, C.; Zhao, D.; and Lin, Q. 2023. MMDialog: A Large-scale Multi-turn Dialogue Dataset Towards Multi-modal Open-domain Conversation. In *Proc. ACL*, 7348–7363.

Ferron, A.; Shore, A.; Mitra, E.; and Agrawal, A. 2023. MEEP: Is this Engaging? Prompting Large Language Models for Dialogue Evaluation in Multilingual Settings. In *Proc. Findings of EMNLP*, 2078–2100.

Fu, J.; Ng, S.; Jiang, Z.; and Liu, P. 2024. GPTScore: Evaluate as You Desire. In *Proc. NAACL*, 6556–6576.

Huang, J.; Pu, Y.; Zhou, D.; Shi, H.; Zhao, Z.; Xu, D.; and Cao, J. 2024a. Multimodal Sentiment Analysis Based on 3D Stereoscopic Attention. In *Proc. ICASSP*, 11151–11155.

Huang, Q.; Cai, P.; Nie, T.; and Zeng, J. 2024b. CLIP-MSA: Incorporating Inter-Modal Dynamics and Common Knowledge to Multimodal Sentiment Analysis With Clip. In *Proc. ICASSP*, 8145–8149.

Jiang, S.; Vakulenko, S.; and de Rijke, M. 2023. Weakly Supervised Turn-level Engagingness Evaluator for Dialogues. In *Proc. CHIIR*, 258–268.

Kumar, D.; Madan, S.; Singh, P.; Dhall, A.; and Raman, B. 2024. Towards Engagement Prediction: A Cross-Modality Dual-Pipeline Approach using Visual and Audio Features. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proc. MM*, 11383–11389. ACM.

Lee, J.; Park, S.; Park, S.; Kim, H.; and Kim, H. 2023. A Framework for Vision-Language Warm-up Tasks in Multimodal Dialogue Models. In *Proc. EMNLP*, 2789–2799.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proc. ICML*, 19730–19742.

Li, J.; Yu, Y.; Chen, Y.; Zhang, Y.; Jia, P.; Xu, Y.; Li, Z.; Wang, M.; and Hong, R. 2024a. DAT: Dialogue-Aware Transformer with Modality-Group Fusion for Human Engagement Estimation. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proc. MM*, 11397–11403. ACM.

Li, L.; Zhang, D.; Zhu, S.; Li, S.; and Zhou, G. 2024b. Response generation in multi-modal dialogues with split pre-generation and cross-modal contrasting. *Inf. Process. Manag.*, 61(1): 103581.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. In *Proc. CVPR*, 26286–26296.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proc. EMNLP*, 2511–2522.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proc. ICLR*.

Lyth, D.; and King, S. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *CoRR*, abs/2402.01912.

Palmer, D.; Zhu, Y.; Lai, K.; VanderHoeven, H.; Bradford, M.; Khebour, I.; Mabrey, C.; Fitzgerald, J.; Krishnaswamy, N.; Palmer, M.; and Pustejovsky, J. 2025. Speech Is Not Enough: Interpreting Nonverbal Indicators of Common Knowledge and Engagement. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *Proc. AAAI*, 29676–29678. AAAI Press.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*, 311–318. ACL.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. ICML*, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In *Proc. EACL*, 300–325.

Shuster, K.; Humeau, S.; Bordes, A.; and Weston, J. 2020. Image-Chat: Engaging Grounded Conversations. In *Proc. ACL*, 2414–2429.

Shuster, K.; Smith, E. M.; Ju, D.; and Weston, J. 2021. Multi-Modal Open-Domain Dialogue. In *Proc. EMNLP*, 4863–4883.

Sun, J.; Li, Z.; and Peng, X. 2024. EmoEcho: Designing Emotion Mimicry Mechanics for Enhancing Social Engagement in Digital Games. In Mueller, F. F.; Kyburz, P.; Williamson, J. R.; and Sas, C., eds., *Proc. CHI*, 120:1–120:6. ACM.

Sun, Q.; Wang, Y.; Xu, C.; Zheng, K.; Yang, Y.; Hu, H.; Xu, F.; Zhang, J.; Geng, X.; and Jiang, D. 2022. Multimodal Dialogue Response Generation. In *Proc. ACL*, 2854–2866.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *Proc. CVPR*, 4566–4575. IEEE Computer Society.

Xu, F.; Jia, F.; and Zhou, W. 2025. Multimodal Prompt Learning for Audio Visual Scene-Aware Dialog. In *Proc. MMM*, 87–100.

Xu, G.; Liu, R.; Harel-Canada, F.; Chandra, N. R.; and Peng, N. 2022. EnDex: Evaluation of Dialogue Engagingness at Scale. In *Proc. Findings of EMNLP*, 4884–4893.

Xu, L.; Gan, Y.; and Jin, Y. 2025. Class Activation Regularization-Based Facial Emotion Recognition Network and its Application in Students' Emotional Engagement Assessment. *IEEE Trans. Affect. Comput.*, 16(2): 1044–1055.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qi, Q.; Zhang, J.; and Huang, F. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *CoRR*, abs/2304.14178.

Yoon, H. S.; Yoon, E.; Tee, J. T. J.; Zhang, K.; Heo, Y.-J.; Chang, D.-S.; and Yoo, C. D. 2024. BI-MDRG: Bridging Image History in Multimodal Dialogue Response Generation. *CoRR*, abs/2408.05926.

Zang, X.; Liu, L.; Wang, M.; Song, Y.; Zhang, H.; and Chen, J. 2021. PhotoChat: A Human-Human Dialogue Dataset With Photo Sharing Behavior For Joint Image-Text Modeling. In *Proc. ACL/IJCNLP*, 6142–6152.

Zar, J. H. 2005. Spearman Rank Correlation. *Encyclopedia of biostatistics*, 7.

Zha, X.; Zhao, H.; and Zhang, Z. 2024. Esihgnn: Event-State Interactions Infused Heterogeneous Graph Neural Network for Conversational Emotion Recognition. In *Proc. ICASSP*, 11136–11140.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *Proc. Findings of EMNLP*, 15757–15773.

Zhang, S.; Lu, Y.; Liu, J.; Yu, J.; Qiu, H.; Yan, Y.; and Lan, Z. 2024. Unveiling the Secrets of Engaging Conversations: Factors that Keep Users Hooked on Role-Playing Dialog Agents. *CoRR*, abs/2402.11522.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proc. ACL (demo)*, 270–278.