OmniDialog: A Multimodal Benchmark for Generalization Across Text, Visual, and Audio Modalities

Anton Razzhigaev[1,2], **Maxim Kurkin**[1,2], **Elizaveta Goncharova**[2], **Irina Abdullaeva**[2], **Anastasia Lysenko**, **Alexander Panchenko**[1,2], **Denis Dimitrov**[2,3], **Andrey Kuznetsov**[2,3]
[1]Skoltech, [2]AIRI, [3]Sber AI, **Correspondence:** razzhigaev@airi.net

Abstract

We introduce *OmniDialog* — the first tri-modal comprehensive benchmark grounded in a knowledge graph (Wikidata) to evaluate the generalization of Large Multimodal Models (LMMs) across three modalities. Our benchmark consists of more than 4,000 dialogues, each averaging 10 turns, all annotated and cross-validated by human experts. The dialogues in our dataset are designed to prevent shortcut learning by incorporating various formats and misleading or irrelevant multimodal cues. We also evaluate both multimodal and unimodal models to gain insights into how they process modality inputs introduced in the conversation.

1 Introduction

Multimodal dialogue systems became a focal point in research, drawing significant attention of both academia and industry. This surge of interest stems from their potential to contribute to more natural and nuanced human-computer interactions by seamlessly integrating text, audio, and visual cues (Zhu et al., 2023; Liu et al., 2023b; Koh et al., 2023b,a). Yet, the complexity of these systems has led to challenges in their evaluation. Existing benchmarks, in many instances, fall short in capturing the intricacies of the real-world interactions, lacking the necessary depth and diversity to evaluate the true capabilities of multimodal dialogue systems (Huang et al., 2024).

In response, we introduce the *OmniDialog* benchmark, a multimodal, multi-turn benchmark designed to evaluate the generalization abilities of Large Multimodal Models (LMMs). Specifically, our benchmark assesses their capability to support multi-turn conversations, process modality injections at random points within the dialogue, and operate with three modalities simultaneously (text, visual, and audio). It stands out by grounding on Wikidata knowledge graph and encompasses a vast array of more than 4,000 dialogues, each with an average of 10 turns. To ensure the highest quality, our human annotators designed these dialogues from scratch and then cross-validated them to ensure accuracy and consistency. The uniqueness of *OmniDialog* lies in its design: it requires deep understanding of three modalities – text, visual, and audio. Moreover, to ensure that systems truly understand the context rather than exploit shortcuts, we present dialogues in various formats.

Our contributions are as follows:

- We introduce *OmniDialog* — the first comprehensive benchmark for evaluating multimodal dialogue models, where questions are based on Wikidata KG facts and incorporates three modalities: text, visual, and audio. This offers a robust, diverse, and challenging platform for assessment.
- We provide comprehensive evaluation of the existing multimodal dialogue systems against this new benchmark.

The primary data for *OmniDialog* is sourced from Wikipedia[1] and Wikidata[2], ensuring both the authenticity and generalization ability of the dialogues. The datasets and evaluation code will be released under an open source licence at https://github.com/ai-forever/OmniDialog.

2 Related Work

In this section, we provide a brief description of popular multimodal datasets and state-of-the-art multimodal transformer architectures.

[1] https://www.wikipedia.org/
[2] https://www.wikidata.org

Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP, pages 183–195
November 16, 2024 ©2024 Association for Computational Linguistics

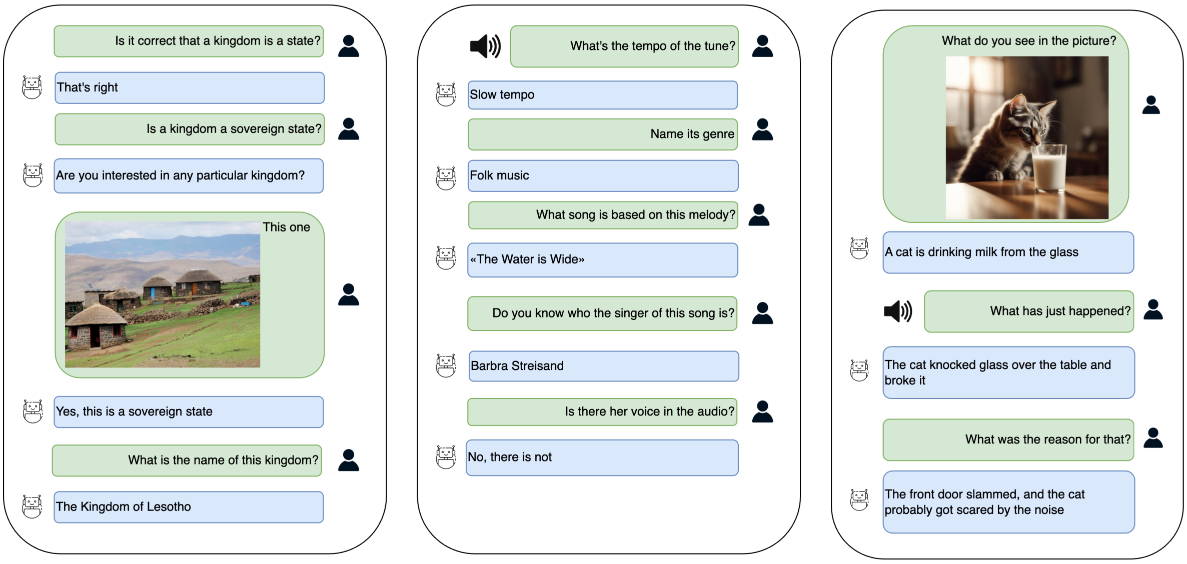

Is it correct that a kingdom is a state?
That's right
Is a kingdom a sovereign state?
Are you interested in any particular kingdom?
This one
Yes, this is a sovereign state
What is the name of this kingdom?
The Kingdom of Lesotho
What's the tempo of the tune?
Slow tempo
Name its genre
Folk music
What song is based on this melody?
«The Water is Wide»
Do you know who the singer of this song is?
Barbra Streisand
Is there her voice in the audio?
No, there is not
What do you see in the picture?
A cat is drinking milk from the glass
What has just happened?
The cat knocked glass over the table and broke it
What was the reason for that?
The front door slammed, and the cat probably got scared by the noise

Figure 1: Examples of dialogues from the OmniDialog dataset.

2.1 Multimodal Dialogue Datasets

Dialogue datasets that merge various modalities play a crucial role in training and evaluating multimodal systems. There are two key aspects that such datasets need to consider: first, the strength and the robustness of relationships among different modalities, and second, the ability to identify user's query and to follow instructions.

Significant progress in cross-modal benchmarking has been achieved recently. Specifically, combinations like language and vision have tackled challenges in image captioning, visual question answering (VQA), and visual reasoning (Young et al., 2014; Chen et al., 2015; Krishna et al., 2016; Goyal et al., 2016). Language and audio studies have emphasised audio captioning and classification (Kim et al., 2019; Drossos et al., 2019; Gemmeke et al., 2017), while language and video research has concentrated on video series description and visual grounding (Li et al., 2021; Chen et al., 2023; Sigurdsson et al., 2016) among other tasks. The vast majority of these datasets are focused on one specific domain - a natural language description of the modality. In contrast, substantially fewer works have addressed a relatively new field of multimodal dialogue systems benchmarking.

Early works in this area considered evaluating multimodal dialogue systems via QA approach on vision-language based tasks (Goyal et al., 2016; Johnson et al., 2016; Gurari et al., 2018). Other studies further aim to estimate the ability of visual instruction following (Dai et al., 2023; Liu et al., 2023b; Xu et al., 2022), visual grounding (Chen et al., 2022; Kazemzadeh et al., 2014). Several benchmarks emphasise the incorporation of common knowledge bases in unstructured form and retrieval techniques into the general VQA setup (Yu et al., 2023; Marino et al., 2019; Schwenk et al., 2022). Besides the visual-text dialogues type, there are also datasets for the joint evaluation of text and audio in both Spoken QA (SQA) (Lee et al., 2018b,a) and audio captioning tasks (Kim et al., 2019; Drossos et al., 2019; Zhao et al., 2023). However, all of these benchmarks suffer from a single drawback: due to the question-answering problem setting, they put very little attention to the model's ability to maintain the long context of the dialogue in order to further rely on it for later response generation and do not focus on interaction of several modalities (e.g. image + audio).

Only in recent time due to the significant success of OpenAI GPT4 and GPT4-V (OpenAI, 2023a,b), along with open-source LLMs (Li et al., 2023a; Awadalla et al., 2023; Dai et al., 2023; Shuster et al., 2020) to carry on complete visually-augmented conversations, there were taken steps towards advanced multimodal dialogue datasets design. These datasets bring together the visual conditioning with the instructionally formulated questions, with reliance upon dialogue context and requirement of extensive domain and world knowledge (Das et al., 2016; de Vries et al., 2016; Mostafazadeh et al., 2017; Johnson et al., 2016; Shuster et al., 2018; Meng et al., 2020; Huang et al., 2023b; Liu et al., 2024). However, the scope of their application is constrained since they focus on using only two modalities — visual and text ones. Other modalities, therefore, remain relatively unexplored in the dialogue setting.

In contrast to the above mentioned works, our benchmark fuses the data from three modalities: text, visual, and audio, and enables building complex relationships on their basis. Furthermore, to the best of our knowledge, *OmniDialog* is the first benchmark to merge all three modalities together in a single dialogue setup. Our benchmark demands multiple knowledge forms, such as basic factual world knowledge and scientific knowledge in historical, physical, and biological domains. The underlying factual evidence in the *OmniDialog* benchmark is derived from the WikiData knowledge graph, therefore, it is precise and reliable. Dialogs are constructed based on a random subset of entities and images from Wikidata.

2.2 Visual-Audio-Language Models

One straightforward approach to embed the ability to understand other modalities into pre-trained LLMs is to use specialised out-of-the-box visual, audio, etc. based models as external tools (Schick et al., 2023; Yang et al., 2023; Li et al., 2023b). This means that the language model serves as a skills orchestrator, invoking "expert" models of particular modality via language calls in order to complete certain tasks when necessary. However, these methods suffer from weak connectivity and limited interaction between modalities, resulting in a loss of significant cross-modal information.

More recently, end-to-end multimodal language models have gained considerable interest. Some of the early studies embedded visual data understanding into LMs via additional parameters augmentation and further joint cross-modal training (Alayrac et al., 2022; Wang et al., 2022; Gong et al., 2023).

As opposed to training from scratch, follow-up research has focused on integration of pre-trained visual and language models. The dominant approach was to implement a trainable projection layer between the pre-trained modality feature extractor and the LLM. This setup leads to the injection of high-quality modality embeddings into the language context, which is perceived as a "foreign language" by the language model. Moreover, keeping the number of tunable parameters small, improves the computational efficiency of the cross-modal training. So far, a variety of different network architectures and learning strategies have been proposed to fuse different vision and language models in a single multimodal system (Liu et al., 2023b; Koh et al., 2023b; Zhang et al., 2023; Gao et al., 2023).

However, these approaches are limited to using mostly image content as input. Only a handful of works have attempted to broaden the model's input feature space by incorporating other modalities (Huang et al., 2023a; Girdhar et al., 2023; Wang et al., 2023; Zhao et al., 2023).

3 OmniDialog

In this section, we describe *OmniDialog* — a benchmark for evaluating multimodal dialogue systems in English. Our dataset is distinguished by its diversity in dialogue types, human annotation, and strict evaluation metrics. It is grounded to knowledge graphs, which means that the discussed facts, images, and audios in dialogs are taken from Wikidata. Our benchmark comprised of more than 4,000 dialogues, each averaging 10 turns, with data and facts sourced primarily from Wikipedia and Wikidata.

3.1 Dialog Types

OmniDialog consists of four main types of dialogues: text-text, visual-text, audio-text, and trimodal dialogues. Each of these is designed to test the system's ability to generalise across different modalities and comprehend information retaining general knowledge from an LLM.

3.1.1 Text Dialogues

Most contemporary generative pre-trained models come with a conversational counterpart. Even though many multimodal conversational systems possess a robust linguistic foundation, understanding how multimodal tuning impacts unimodal capabilities is crucial. Consequently, *OmniDialog* incorporates a strictly textual segment.

These dialogues aim to gauge solely language-based, in-context comprehension. Hence, models must rely exclusively on linguistic understanding to navigate the dialogue. For a cohesive integration between textual and multimodal dialogues, we employ Wikidata facts as our primary dialogue question source, emphasizing factual discourse over creative content.

For constructing textual dialogues, we selected topics and extracted corresponding random Wikidata entities, including films, writers, actors, animals and food. Human annotators then crafted dialogues based on the relationships and facts (Wikidata triples) associated with these entities. Recog-

Dataset	Multi-turn	Interleaved	#Dialogs	#Turns	#Images	#Audio	KG	Annotation
CLEVR-Dialog (Johnson et al., 2016)	Yes	No	425.0 k.	4250.0 k.	85.0 k.	No	No	Synthetic
OpenViDial (Meng et al., 2020)	No	No	1100.0 k.	1100.0 k.	1100.0 k.	No	No	Synthetic
DialogCC (Lee et al., 2022)	Yes	Yes	92.9 k.	930.0 k.	651.0 k.	No	No	Synthetic
SparklesEval (Huang et al., 2023b)	Yes	No	6.5 k.	26.0 k.	10.9 k.	No	No	Synthetic
MPCHAT (Ahn et al., 2023)	Yes	Yes	15.0 k.	42.5 k.	153.0 k.	No	No	Synthetic
LLaVA (Liu et al., 2023b)	Yes	No	56.7 k.	514.0 k.	56.7 k.	No	No	Synthetic
PhotoBook (Haber et al., 2019)	Yes	No	2.5 k.	164.6 k.	0.4 k.	No	No	Human
VisDial (Das et al., 2016)	Yes	No	133.0 k.	1200.0 k.	133.0 k.	No	No	Human
GuessWhat?! (de Vries et al., 2016)	Yes	No	155.0 k.	821.0 k.	66.0 k.	No	No	Human
IGC (Mostafazadeh et al., 2017)	Yes	No	4.2 k.	25.3 k.	4.2 k.	No	No	Human
Image-Chat (Shuster et al., 2018)	No	No	201.8 k.	401.0 k.	201.8 k	No	No	Human
MMD (Saha et al., 2017)	Yes	Yes	151.6 k.	6400.0 k.	4200.0 k.	No	No	Human
PhotoChat (Zang et al., 2021)	Yes	Yes	12.3 k.	156.0 k.	10.9 k.	No	No	Human
MMDialog (Feng et al., 2022)	Yes	Yes	1800.0 k.	4920.0 k.	1530.0 k.	No	No	Human
VDialogUE (Li et al., 2023c)	Yes	Yes	1080.0 k.	4900.0 k.	1530.0k.	No	No	Human
MMDU Benchmark (Liu et al., 2024)	Yes	Yes	110	1.6 k.	421	No	No	Human
OmniDialog (Ours)	Yes	Yes	4.0 k.	27.0 k.	2.4 k.	1.0 k.	Yes	Human

Table 1: Comparison of OmniDialog with existing multi-modal English-language dialogue datasets.

Modality	Dialogue Type	# Dialogues	# Questions (KG-based)	# Questions (General)
Text	General dialogues	1 455	2,000	6 492
Visual	Single Image	1 794	6 506	5 552
	Clarifying Image	400	1 349	1 420
	Misleading Image	220	290	396
Audio	Single Audio	283	2 366	1 334
	Clarifying Audio	500	2 109	1 471
	Dual Audio	100	499	203
Trimodal	General dialogues	165	0	2 000

Table 2: Statistics of dialogues in the OmniDialog, categorized by modality and dialogue type.

nizing that Wikidata might occasionally offer limited information, annotators were advised to supplement dialogue content using relevant Wikipedia articles.

3.1.2 Visual Dialogues

Visual dialogues in OmniDialog are designed to assess the model's capacity to integrate visual processing with natural language understanding. In each dialogue, a single image is employed (not necessarily in the initial dialog turn), and at least four facts from WikiData are utilized. The visual dialogues are divided into three categories, each with its unique structure and purpose:

1. **Single Image Dialogues:** In this format, the user introduces a single image and poses questions related to it. These questions encompass both intricate queries oriented towards facts from WikiData and straightforward inquiries regarding the content of the image. A sample dialogue is illustrated in Figure 1.

2. **Clarifying Image Dialogues:** This dialogue structure begins with the user posing a question that cannot be answered without additional clarifying information. The user then provides an image to supplement the dialogue and to facilitate further discussion.

3. **Misleading Image Dialogues:** In this scenario, the user poses a question along with an image that, while thematically related, is irrelevant. The model must identify the image's irrelevance and respond accurately, followed by a discussion about the image. Some baseline multimodal LLMs tend to shift focus on the image, ignoring its irrelevance to the query. This dialogue type is designed to address such tendencies, encouraging models to balance attention between visual and textual inputs mitigating such shortcut behaviors.

3.1.3 Audio Dialogues

Audio and textual modalities have been fused within dialogues in *OmniDialog* in a way that allows a better understanding of the model's sound comprehension ability's contribution to its multimodal dialogue performance.

The audio-text OmniDialog dialogues were used to evaluate the role of the model's sound comprehension in its multimodal conversation skills. To ensure the diversity of different sound types and the balance between the factual reliability and the realism of the discussion, we defined the following rules: 1) the same audio should be used in dialogues only once; 2) each dialogue should contain at least 2 and no more than 4 facts; 3) non-factual knowledge questions should be as simple that a 5-year-old child would be able to answer.

Based on the collected sound files and Wikidata textual facts, we also identified three categories within the structure of the dialogues:

1. **Single Audio Dialogues**: In this split, first turn of conversation contains an audio recording and a question about its content. In the further dialogue progression, both evidence-based questions referring to the WikiData entity and sound based questions are used.

2. **Clarifying Audio Dialogues**: Within this dialogue type, the user sends an audio recording along with an accompanying question in the middle of the discussion. In this case, the audio serves as a clarification to one of the evidence-based questions. The subsequent dialogue is built around the audio content, with a variety of relevant questions about it.

3. **Dual Audio Dialogues**: Two audio recordings are used simultaneously in these dialogues, setting them apart from the other types. Conversations include both questions about content of audios and evidence-based ones that pertain to the related entities recognized on audio. The dialogues should emphasise the connection between the sounds, whether by questioning their characteristics or comparing the entities they are associated with. The purpose of this dialogue type is to examine the model's ability to differentiate and remember various audio information throughout the conversation.

3.1.4 Trimodal Dialogues

Trimodal dialogues in OmniDialog aim to gauge the model's capability to process and combine information from three modalities: text, visual, and audio ones. Within these dialogues, models are expected to integrate and act on the information derived from all sources to provide accurate answers. We have curated 100 high-quality trimodal dialogues, categorized into four distinct scenarios:

1. **Image-sound Matching:** The model is asked to match an object's sound from the audio with its visual representation in an image to identify the subject of discussion.

2. **Multimodal Navigation:** The audio clarifies the object, concept, or event depicted in the image. Subsequent questions focus on this audio-visual correspondence.

3. **Audio-based Continuation:** These dialogues start with an image showing a certain situation. The task is to understand how this situation might change considering the given audio.

4. **Misleading Dialogues:** The dialogues contain unrelated audio and visual prompts. Models must adeptly shift attention between these different sources of information to respond accurately.

3.2 Human Annotation

Multimodal dialogue creation is a nuanced task demanding attention to the details. To ensure the integrity of the data, we implemented a rigorous protocol:

Annotation Protocol:

- Develop comprehensive guidelines with illustrative examples.
- Host training sessions to resolve annotators' doubts.
- Enforce a rigorous verification process for all dialogues.

Media Criteria:

- Ensure audio and images align with articles and provide multiple facts.
- Maintain a minimum duration of 4 seconds for audio clips.

Dialogue Rules:

- Keep dialogues between 4 and 20 messages.
- Base multimodal questions on Wikidata facts or clear links between articles.

- Root text dialogues in Wikipedia data.

- Limit user questions to 2-15 words and answers to 1-10 words.

Synonyms: Include synonyms in dialogues when alternate answers exist. Ensure they fully address the user's questions. They are essential for text and audio dialogues.

Negative examples: Negative examples are required for multiple choice evaluation setup to be passed as answer options along with ground truth answer. So they have been added to all questions, including names, dates, numbers, yes/no, and others. Along with negative examples, neutral options "can't answer", "not enough information" were always added.

Challenges: We faced issues like repetitive facts and overly concise answers. Solutions included rephrasing, adding comparative elements, and varying sentence structures.

Team Workflow: Our six-member team produced 20-25 dialogues each day. Validation was quicker, taking about 40% of the dialogue creation time.

3.3 Evaluation and Metrics

Evaluating generative models is inherently challenging. The evaluation method can alter both numerical results and leaderboard rankings drastically. While *OmniDialog* was initially designed for open-ended generative responses, this design posed evaluation complexities. The variability in training can result in models producing diverse or brief responses, making it essential for ground truth options to account for such variability.

Given our benchmark's focus on factual questions, we aim to prioritize answer correctness over stylistic variations. To this end, we convert dialogues into multiple-choice questions and determine accuracy based on the selected response.

The given answer is recorded, but for consequential questions the *correct* answer to the current question is used in the previous context to enforce model capabilities and fair comparison between multimodal and text-only models. For evaluation details, please refer to Appendix A.2.

The final score is reported as the mean accuracy of model answers. **Total** stands for mean of accuracies reported on benchmark subsets.

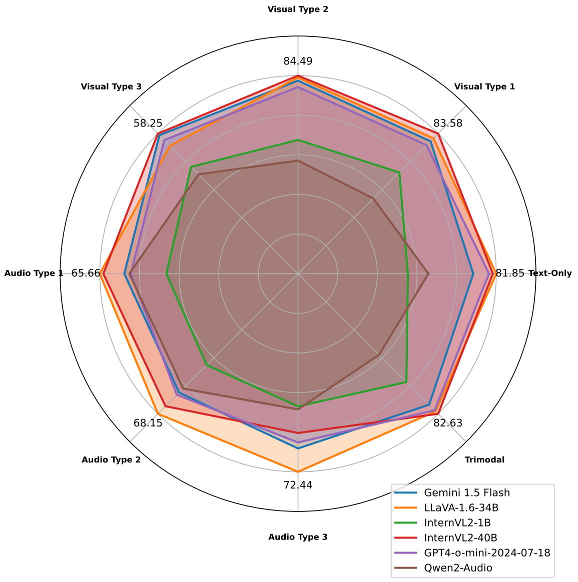


Figure 2: Comparison of models: Radar plot of baseline evaluation against different dialog types.

4 Baselines

Given that the most progress has been achieved for the combination of visual and textual modalities, we have adapted several recently introduced models to process the dialogue input. The baselines include not only the trimodality models but also bimodal (visual-text) and unimodal models (language-only) models. The latter are evaluated using the corresponding part of OmniDialog.

Trimodal. Gemini (Team, 2024) serves as a baseline model supporting all three modalities.

Visual-Language. Several (near) state-of-the-art models are assessed on the language-only and visual-language OmniDialog parts: series of LLaVA models (Liu et al., 2023a), series of InternVL 2.0 (Chen et al., 2024) chat models with the LLM backbones of various size (from 1B to 40B parameters) and strong vision encoder (InternViT-6B and its distilled version InternViT-300M), and Idefics2 8B (Laurençon et al., 2024) vision-language model trained with the interleaved data.

These models ignore audio data during the evaluation process and only encode images with proposed visual adapter architectures.

Language-only. GPT-4-mini is used as a language-only baseline reference for OmniDialog. No encoding of visual or audio data is performed during the evaluation and model solely reasons from previously answered questions.

Model	Text-Only	Visual			Audio			Trimodal	Total
		Type 1	Type 2	Type 3	Type 1	Type 2	Type 3		
Random Guessing	16.41	26.68	26.18	26.54	17.85	18.90	15.56	40.08	22.16
Trimodal Models									
Gemini 1.5 Flash (Team, 2024)	72.36	79.01	82.34	57.58	57.55	57.83	**63.90**	77.27	68.48
Visual-Language Models									
LLaVA-v1.5-7B (Liu et al., 2023a)	68.45	69.23	64.15	44.11	54.58	59.25	59.23	69.26	61.03
LLaVA-v1.5-13B	69.74	73.22	72.39	46.46	55.26	60.11	63.01	75.05	64.40
LLaVA-v1.6-Mistral-7B	69.47	73.92	73.44	48.15	59.59	62.90	62.45	75.05	65.62
LLaVA-1.6-34B	81.85	80.77	83.70	53.20	65.66	68.15	72.44	81.16	73.37
Idefics2-8B (Laurençon et al., 2024)	66.45	65.86	69.85	54.21	59.16	62.67	63.58	71.47	64.16
InternVL2-1B	45.36	60.42	57.06	44.44	43.63	44.44	48.51	63.89	50.97
InternVL2-2B	50.30	65.79	63.45	46.56	49.13	47.66	51.33	69.58	55.47
InternVL2-8B	70.02	78.63	78.44	56.90	60.71	61.51	65.75	78.00	68.74
InternVL2-26B	71.45	79.90	79.40	53.20	56.25	56.96	53.20	78.42	66.1
InternVL2-40B	**80.43**	**83.58**	**84.49**	**58.25**	**64.48**	**64.46**	58.25	**82.63**	**72.07**
GPT4-o-mini-2024-07-18 (OpenAI, 2023a,b)	78.85	76.67	79.61	55.56	55.40	58.92	61.72	80.56	68.41
Audio-Language Models									
Qwen2-Audio (Chu et al., 2024)	53.84	44.90	48.25	41.33	55.87	55.87	49.63	47.79	49.69
Language-only Models									
GPT4-o-mini-2024-07-18	-	26.61	30.14	34.68	-	-	-	42.55	-

Table 3: Performance evaluation results across different categories of models on our OmniDialog benchmark. Best in each dialogue category is highlighted id **bold**.

5 Discussion

We discuss result of OmniDialog benchmark evaluation presented in Table 3.

1. Influence of the LLM Backbone on Performance. There is a strong correlation between the evaluation results and the performance of the model's LLM backbone. Models with more powerful backbones consistently achieve higher results across all image types. For instance, there is a significant performance gap of over 20% between the InternVL2 40B and InternVL2 1B models across all dialogue types. Similarly, the difference between LLaVA-1.6 34B and 7B models averages nearly 8%. The dialogs in OmniDialog are constructed using Wikipedia and WikiData entities. Hence, the LLM's knowledge of Wikipedia content helps multimodal models based on these backbones generalize better across multimodal information.

2. Challenges in Visual Type-3 Dialogues. The lowest performance in visual-language models is observed in Type 3 visual dialogues, where misleading images are introduced into the dialogue context. Although models can answer textual questions correctly, they struggle with questions related to the image modality, especially when the question is distant from the image. This challenge may arise from the typical training process, where models are used to encountering the image and accompanying question in a strict sequence. LLaVA-based models experience an average performance drop of 20% in visual Type 3 dialogues compared to visual Types 1 and 2.

All benchmarked models do not generalize well on questions not matching context of the distracting image. We show the example of Type 3 visual dialogue in Section A Figure 7.

3. Challenges in Audio Dialogues. The most challenging modality type of dialogue in the benchmark is audio dialogues. Visual-language models struggle to guess the correct answer to questions, even with the teacher-forcing approach, leading to lower performance compared to other dialogue types. It is evident that models adapted to the audio domain, such as Qwen-Audio 7B, show lower metrics on audio datasets compared to stronger baselines that do not process audio.

4. High Results on Multimodal Inputs. The teacher-forcing of context introduced in 4 reduces the influence of modality input, enabling models to provide correct answers even when they struggle to process a specific modality. Reinforcing past context with correct answers leads to a significant performance boost in tasks with added modalities. Thus, strong visual-language models without audio perception (such as LLaVA-v1.6 34B) perform well on audio-based dialogues.

If we switch to a mode where specific model answers are added as the continuation of the dialogue for further assessment, the results might differ. We leave this type of evaluation for further research.

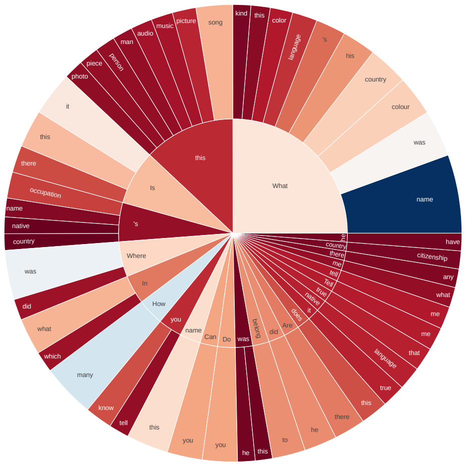


Figure 3: Bigram word pairs in user questions.

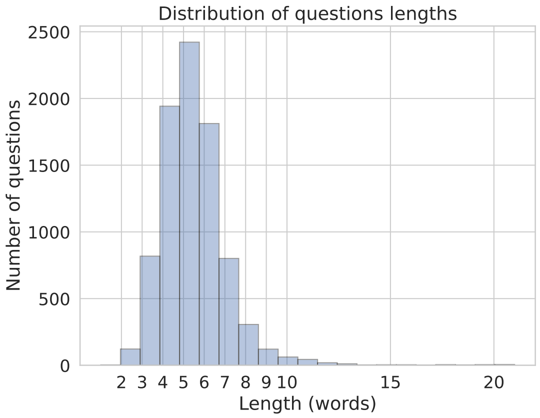


Figure 4: Distribution of word lengths in user questions with an average of 6 words.

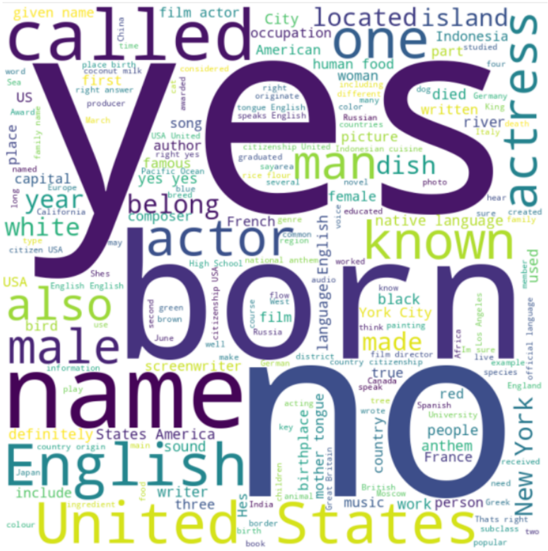


Figure 5: Wordcloud of assistant answers.

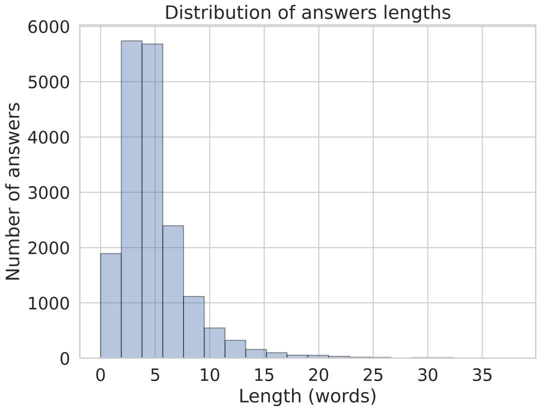


Figure 6: Distribution of word lengths in assistant answers with an average of 4.5 words

6 Benchmark Statistics

Figures 3 - 4 show characteristics of the OmniDialog dataset in terms of textual analysis of the replicas. We explore key elements such as the word bigram distribution in user questions, the word cloud of required assistant responses as well as the length distributions for both of them.

7 Conclusion

We introduced *OmniDialog*, a diverse and comprehensive benchmark for multimodal dialogue systems, comprising more than 4,000 dialogues based on entities, facts, and media from Wikidata knowledge graph. These dialogues, designed to prevent shortcut learning, offer a unique challenge for multimodal systems with various formats and misleading cues. The knowledge graph foundation of OmniDialog makes it a valuable resource for comparing multimodal models based on factual information.

We also analyze several baselines, both open-source and proprietary, with respect to their generalization across various modalities and types of dialogues introduced with OmniDialog. Following the teacher-forcing evaluation, we compared models operating with different modalities. We believe that the creation of multimodal benchmarks will motivate the community to develop dialogue assistants further and enhance their evaluation. We also aim to build upon this work by providing various types of assessments using LLM oracles, both with and without teacher forcing, which may result in a fairer comparison of the quality of multimodal models.

Limitations

While Wikidata is generally considered a reliable source of information, it carries an inherent risk of bias, as it may contain its own biases, errors, and inconsistencies. Since OmniDialog is focused on English, using this source limits our ability to fairly evaluate the multilingual generalization of models. Additionally, the potential for human errors during annotation and editing cannot be overlooked, which might introduce further discrepancies or inaccuracies into the dataset.

Ethical Statement

We acknowledge the importance of diversity and representation in data sources. Our aim with *OmniDialog* was to ensure a broad and diverse representation in dialogues, striving to avoid potential biases where certain cultures might be underrepresented.

References

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. Mpchat: Towards multimodal persona-grounded conversation. *ArXiv*, abs/2305.17388.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv*, abs/2308.01390.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding answers for visual questions asked by visually impaired people. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19076–19085.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023. Valor: Vision-audio-language omni-perception pretraining model and dataset. *ArXiv*, abs/2304.08345.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Harm de Vries, Florian Strub, A. P. Sarath Chandar, Olivier Pietquin, H. Larochelle, and Aaron C. Courville. 2016. Guesswhat?! visual object discovery through multi-modal dialogue. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4466–4475.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: an audio captioning dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Annual Meeting of the Association for Computational Linguistics*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, W. Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Jiao Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *ArXiv*, abs/2304.15010.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set:

An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *CoRR*, abs/2305.04790.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414.

Danna Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and R. Fernández. 2019. The photobook dataset: Building common ground through visually-grounded dialogue. In *Annual Meeting of the Association for Computational Linguistics*.

Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, Jingyang Yuan, Wei Ju, Luchen Liu, Tianyu Liu, Baobao Chang, and Ming Zhang. 2024. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. *Preprint*, arXiv:2407.00468.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023a. Language is not all you need: Aligning perception with language models. *ArXiv*, abs/2302.14045.

Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. 2023b. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *ArXiv*, abs/2308.16463.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

Sahar Kazemzadeh, Vicente Ordonez, Marc andre Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *North American Chapter of the Association for Computational Linguistics*.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023a. Generating images with multimodal language models. *CoRR*, abs/2305.17216.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023b. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32 – 73.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung yi Lee. 2018a. Odsqa: Open-domain spoken question answering dataset. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 949–956.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018b. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Interspeech*.

Young-Jun Lee, ByungSoo Ko, Han-Gyu Kim, and Ho-Jin Choi. 2022. Dialogcc: Large-scale multi-modal dialogue dataset. *ArXiv*, abs/2212.04119.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597.

Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohith Krishnan Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, Tamara L. Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, and Zicheng Liu. 2021. Value: A multi-task benchmark for video-and-language understanding evaluation. *ArXiv*, abs/2106.04632.

Minghao Li, Feifan Song, Yu Bowen, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023b. Api-bank: A benchmark for tool-augmented llms. *ArXiv*, abs/2304.08244.

Yunshui Li, Binyuan Hui, Zhaochao Yin, Wanwei He, Run Luo, Yuxing Long, Min Yang, Fei Huang, and Yongbin Li. 2023c. Vdialogue: A unified evaluation benchmark for visually-grounded dialogue. *ArXiv*, abs/2309.07387.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *Preprint*, arXiv:2406.11833.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *ArXiv*, abs/2012.15015.

N. Mostafazadeh, Chris Brockett, William B. Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *ArXiv*, abs/1701.08251.

OpenAI. 2023a. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2023b. Gpt-4v(ision) system card.

Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2017. Towards building large scale multimodal domain-aware conversation systems. In *AAAI Conference on Artificial Intelligence*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *ArXiv*, abs/2302.04761.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image-chat: Engaging grounded conversations. In *Annual Meeting of the Association for Computational Linguistics*.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2020. Multi-modal open-domain dialogue. In *Conference on Empirical Methods in Natural Language Processing*.

Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, Ivan Laptev, and Abhinav Kumar Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. *ArXiv*, abs/1604.01753.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. 2023. One-peace: Exploring one general representation model toward unlimited modalities. *ArXiv*, abs/2305.11172.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Annual Meeting of the Association for Computational Linguistics*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction. *ArXiv*, abs/2305.18752.

Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv*, abs/2308.02490.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *ArXiv*, abs/2108.01453.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Jiao Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *ArXiv*, abs/2303.16199.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *ArXiv*, abs/2307.08581.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

A Appendix

A.1 Misleading Image example

In Figure 7, we present the evaluation of the visual-language model on a visual dialogue containing a misleading image. Although the model successfully answered the textual questions, it failed to provide the correct answer related to the image in the subsequent conversation.

A.2 Evaluation pipeline

In this section we clarify how teacher forced evaluation is performed. Past context is obtained from ground truth labels for previous questions.

System Prompt

You are an expert whose job is to answer questions from the user possibly using images and audios and previous questions with given answers.
First, reason about the correct answer. Then write the answer in the following format where X is exactly one of English letters (A,B,C,...): "ANSWER: X".
If you are uncertain of the correct answer, guess the most likely one.

Context

User: What is the country of origin of the film "Addicted to Plastic?"
A) Canada.
B) Colombia.
C) Australia.
D) France.
E) China.
F) Russia.
G) Argentina.
Assistant: Answer: A

User: Can you Tell me the director's name?
A) Stanley Kubrick.
B) It's Ian Connacher.
C) Akira Kurosawa.
D) Alfred Hitchcock.
E) John Ford.
F) Ingmar Bergman.
G) Martin Scorsese.
Assistant: Answer: B

User: What genre?
A) Gothic film.
B) Biographical film.
C) Historical film.
D) Gangster film.
E) It's a documentary film.
F) War film.
G) Detective film.
Assistant: Answer: E

User: Who composed the music?
A) Ian Connacher.
B) Richard Wagner.
C) Siegfried Ochs.
D) Antonio Vivaldi.
E) Pyotr Tchaikovsky.
F) Franz Schubert.
G) Frederic Chopin.
Assistant: Answer: A

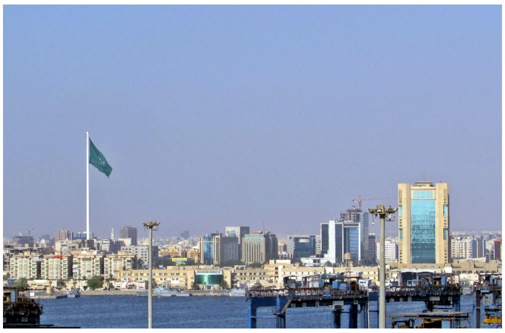
User: What year did it come out?
A) In 2009.
B) In 2007.
C) In 2005.
D) In 2010.
E) In 2001.
F) In 2006.
G) In 2008.
Assistant: Answer: G

User: Answer, what is its duration?
A) 85 minutes.
B) 86 minutes.
C) 99 minutes.
D) 88 minutes.
E) 89 minutes.
F) 87 minutes.
G) 91 minutes.

Model Generation

Assistant: Answer: A

Visual Dialogue (Type-3): Misleading Image



User: In which country is the city of Berlin located?

Choices:

- A. Spain
- B. Italy
- C. France
- D. Germany

Assistant: D. Germany

Ground Truth: D. Germany

Continuation

User: In which country is the city of Berlin located?
Assistant: Germany
User: What time of day is shown in the picture?
Choices:

- A. It is nighttime
- B. It is daylight hours
- C. It is sunset
- D. It is sunrise

Assistant: D. It is sunrise

Ground Truth: B. It is daylight hours

Continuation

User: In which country is the city of Berlin located?
Assistant: Germany
User: What time of day is shown in the picture?
Assistant: It is daylight hours
User: What country is shown in the picture?
Choices:

- A. United States
- B. Saudi Arabia
- C. China
- D. Brazil

Assistant: A. United States

Ground Truth: B. Saudi Arabia

Figure 7: Example of visual Type 3 with the misleading image