
ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst

Zijia Zhao^{1,3}, Longteng Guo², Tongtian Yue^{1,3},
 Sihan Chen^{1,3}, Shuai Shao², Xinxin Zhu^{1,3}, Zehuan Yuan², Jing Liu^{1,3}
¹Institute of Automation, Chinese Academy of Sciences ²Bytedance Inc.
³School of Artificial Intelligence, University of Chinese Academy of Sciences
<https://iva-chatbridge.github.io>

Abstract

Building general-purpose models that can perceive diverse real-world modalities and solve various tasks is an appealing target in artificial intelligence. In this paper, we present ChatBridge, a novel multimodal language model that leverages the expressive capabilities of language as the catalyst to bridge the gap between various modalities. We show that only language-paired two-modality data is sufficient to connect all modalities. ChatBridge leverages recent large language models (LLM) and extends their zero-shot capabilities to incorporate diverse multimodal inputs. ChatBridge undergoes a two-stage training. The first stage aligns each modality with language, which brings emergent multimodal correlation and collaboration abilities. The second stage instruction-finetunes ChatBridge to align it with user intent with our newly proposed multimodal instruction tuning dataset, named MULTIS, which covers a wide range of 16 multimodal tasks of text, image, video, and audio modalities. We show strong quantitative and qualitative results on zero-shot multimodal tasks covering text, image, video, and audio modalities. All codes, data, and models of ChatBridge will be open-sourced.

1 Introduction

Humans interact with the world through multiple modalities — we see objects, hear sounds, feel textures, smell odors, speak words, and so on. By leveraging complementary information from each modality, we obtain a comprehensive understanding of our surroundings.

In order for Artificial Intelligence to complete various real-world tasks in the wild, it needs to be able to interpret, relate, and reason about information from multiple modalities. Significant processes have been made in multimodal learning applications, including vision and language learning [13, 33, 4, 32], video understanding [12, 39], audio-visual speech recognition [36], autonomous driving *etc.* However, current paradigms in multimodal learning often still require acquiring all types and combinations of paired data, and their capabilities are often limited to solving specific tasks with model tuning, *e.g.*, visual question answering, sentiment analysis *etc.*

In this paper, we present *ChatBridge*, a unified multimodal model that harnesses the power of advanced large language model (LLM) as a language catalyst to interpret, correlate, and reason about various modalities, and can perform zero-shot tasks of human instructions through multi-round dialogues. Large language models, such as ChatGPT [40], GPT-4 [41], and LLAMA [54], have demonstrated exceptional proficiency in understanding and generating human-like text. They show that language can act as a universal interface for a general-purpose assistant, where various tasks can be explicitly represented and responded to in language. By extending LLMs' capabilities to incorporate diverse multimodal inputs, we devise a multimodal language model that can perceive

real-world modalities, as well as follow instructions, think, and interact with humans in natural language.

Our method doesn't require datasets where all modalities co-occur with each other. Instead, we leverage language as the catalyst to bridge modalities – we only require easy-acquired, language-paired two-modality data (*e.g.*, image-text pairs, video-text pairs, audio-text pairs, *etc.*). Such strategy leads to an emergent multimodal correlation and collaboration across all of the modalities, enabling zero-shot perception capabilities on multimodal inputs without explicitly paired training data (*e.g.* the rare video-audio-text triples data where the text description describes both the video and audio contents).

Specifically, ChatBridge integrates multiple modality-specific encoders and an LLM, Vicuna [14], which is built upon LLaMA [54], with learnable perceiver modules in between to project embeddings from different modalities into the semantic space of LLM. ChatBridge undergoes a two-stage training on large-scale language-paired two-modality data and self-built multimodal instruction-following data. In the first stage, we pretrain ChatBridge to align each modality with language, which brings emergent multimodal correlation and collaboration abilities with LLM as a language catalyst. In the second stage, we instruction-finetune ChatBridge to align the model with user intent on our newly collected MULTimodal InSTRUCTION tuning dataset (MULTIS), enabling more effective zero-shot generalization on multimodal tasks. MULTIS covers a wide range of 16 multimodal task categories and 15 source datasets involving image, video, and audio content. It consists of both standardized task-specific data and open-ended multimodal chat data.

Our paper makes the following contributions:

- We propose ChatBridge, an approach to learning a unified multimodal model to interpret, correlate, and reason about various modalities without relying on all combinations of paired data.
- We collect MULTIS, a multimodal instruction tuning dataset that consists of 16 diverse multimodal tasks covering text, image, video, and audio modalities.
- We quantitatively and qualitatively evaluate ChatBridge on a wide range of zero-shot multimodal tasks covering text, image, video, and audio modalities, and demonstrate that ChatBridge achieves strong zero-shot performance on these tasks.
- We will open-source the codebase, the MULTIS data, the model checkpoint, and a multimodal chat demo to facilitate future research toward building a general-purpose multimodal assistant.

2 Related works

2.1 Multimodal Learning

Multimodal learning aims to acquire knowledge from diverse forms of data and apply it to various tasks involving multiple modalities. Some approaches typically employ joint learning strategies and generate a merged embedding as the complementary multimodal representation across two modalities, including vision-language [33, 13, 32, 4], audio-visual [64, 44, 30], RGB-D [61, 68, 9], and speech-text [71], among others. In contrast, some other methods focus on learning separate representations for each modality and then aligning them in a shared latent space with metric learning. The seminal work CLIP [46] initially maps visual and textual embeddings into the same space, achieving remarkable performance in retrieval and classification tasks. Subsequent methods [23, 39] have extended this paradigm to incorporate additional modalities. More recently, ImageBind [21] has proposed a multimodal method for aligning different modalities with images, following a similar paradigm.

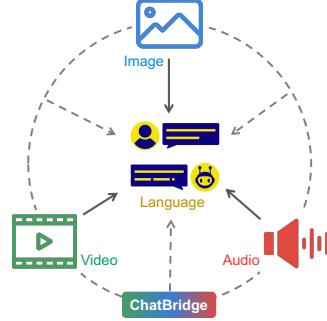


Figure 1: ChatBridge overview. We harness the power of advanced LLM as the catalyst to bridge modalities with easy-acquired, language-paired two-modality data (*e.g.*, image-text, video-text, and audio-text), resulting in a multimodal LLM with emergent multimodal correlation and collaboration abilities across all of the modalities.

ChatBridge projects embeddings from different modalities into the semantic space of LLM and facilitates modality collaboration within LLM. This integration harnesses the strengths of both multimodal learning methods, resulting in a more comprehensive and effective approach.

2.2 LLM and Multimodal LLM

In recent years, there has been remarkable progress in the development of LLMs, particularly in the context of GPT-3[6]. As a result, numerous LLMs have been introduced, such as OPT[67], BLOOM[47], PaLM[15], GLM[66], and LLaMA[54]. The success of the general purpose assistant ChatGPT[40] has inspired researchers to explore methods for emulating its proficiency by employing instruction tuning techniques on language models [43, 58, 53, 14, 16] and vision-language models [35, 62, 73].

Despite the successful applications of LLMs in processing language, their ability to comprehend modalities beyond text, such as vision and audio, remains a challenge. Recently, researchers have made efforts to extend language models to incorporate visual inputs, employing two distinct paradigms: systematic collaboration and end-to-end trained approaches. Systematic collaboration approaches, exemplified by Visual ChatGPT [59], MM-REACT [63], HuggingGPT [50] and ChatVideo [56], leverage various vision experts or tools to express visual information through textual descriptions. In these methods, LLMs, such as ChatGPT [40], act as agents and are prompted to select the appropriate experts and tools for visual comprehension. On the other hand, end-to-end trained approaches utilize LLMs to construct unified image-based multimodal models. Flamingo [3] freezes the pretrained vision encoder and LLM, integrating visual and language modalities using gated cross-attention, which exhibits impressive few-shot capabilities. BLIP-2 [31] employs Q-Former to align visual features from the frozen visual encoder and LLMs. Additionally, PaLM-E [18] directly incorporates features from sensor modalities into PaLM [15]. There are also several methods built on open-sources LLM LLaMA [54] and its variations Alpaca [53] and Vicuna [14], including LLaVA [35], MiniGPT4 [73] and mPLUG-Owl [65].

3 Methods

ChatBridge is a multimodal language model capable of perceiving real-world multimodal information, as well as following instructions, thinking, and interacting with humans in natural language. In this paper, we consider the modalities of image, video, and audio, while deferring the integration of additional modalities such as sketch and point cloud to future works.

3.1 Architecture Overview

As illustrated in Figure 2, ChatBridge consists of multiple modal-specific encoders and perceiver modules, and a transformer-decoder-based LLM. Inspired by Flamingo [3] and BLIP-2 [31], we introduce perceiver modules to bridge the encoders and the LLM. The perceiver summarizes the variable-length embeddings from each encoder’s outputs within a given number of learnable query tokens. It thereby produces outputs of the same shape for all modalities. Also, as the number of query tokens is much smaller than the size of encoder features, it significantly reduces the computation cost in LLM. We instantiate the perceiver as a transformer decoder with learnable query tokens and the encoder embeddings as the input.

The information \mathbf{X}_i from the i -th modality is first fed into the encoder h_i to extract its features. And then each perceiver ρ_i with learnable queries \mathbf{H}_i transforms them into a shared latent space:

$$\mathbf{Z}_i = \rho_i(Q = \mathbf{H}_i, K, V = h_i(\mathbf{X}_i)) \quad (1)$$

Given the all multimodal inputs and human instruction $\mathbf{X}_{instruction}$ as inputs, the LLM f generates the final response text sequence \mathbf{Y} by:

$$\mathbf{Y} = f(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n, \mathbf{X}_{instruction}) \quad (2)$$

Specifically, we choose open-sourced Vicuna-13B [14] as the LLM, which is built upon LLaMA, and reported to achieve 90% of ChatGPT’s quality as per GPT-4’s evaluation. As for the modal-specific encoders, we choose ViT-G [51] as the vision encoder to encode images and videos, and BEAT [11] as the audio encoder to encode audios. We sample 4 frames from each video and concatenate

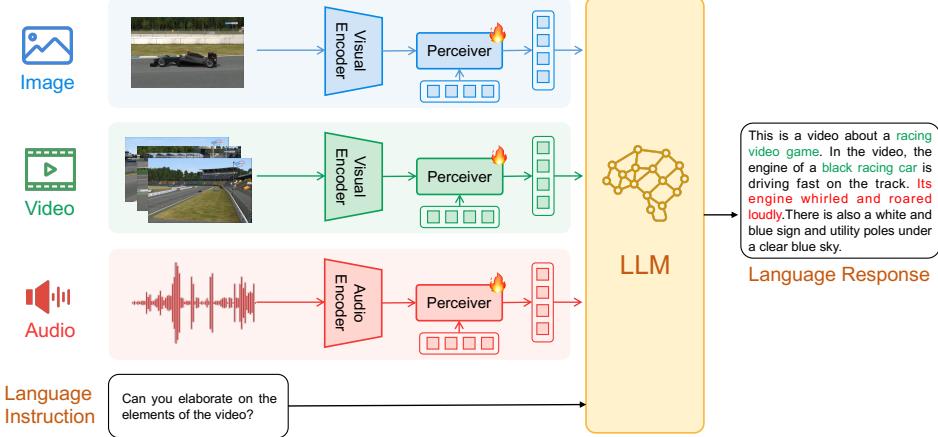


Figure 2: Model architecture of ChatBridge. It consists of multiple modal-specific encoders and perceiver modules and an LLM.

their respective frame features to form the video features, which are inputs of the video perceiver. Similarly, for each audio, we divide it into clips of 10-second intervals and concatenate the clip features to create the audio features. We use a shared perceiver for all modalities while each modality has its independent learnable query tokens. Due to limited computation resources, we only train the perceivers and their learnable query tokens while keeping the encoders and LLM frozen during the whole training process.

3.2 Two-stage Training

Motivated by ChatGPT [40] that is built upon the pretrained GPT-3.5, ChatBridge also undergoes a two-stage training on large-scale language-paired two-modality data, and self-built multimodal instruction-following data.

Stage1: Multimodal Alignment In the first stage, we pretrain ChatBridge to align each modality with language, which brings emergent multimodal correlation and collaboration abilities with LLM as a language catalyst. We leverage large-scale language-paired two-modality data for multimodal alignment training, including image-text, video-text, and audio-text pairs. Specifically, the training data consists of publicly available datasets of image-text pairs (including MS-COCO [34], SBU Captions [42], Conceptual Captions [49, 7], LAION-115M [48]), video-text pairs of Webvid10M [5], and audio-text pairs of WavCaps [38]. The raw unimodal data (*i.e.* images, videos, audios) is sequentially fed into the modality-specific encoder and perceiver to get unimodal embeddings. The input format of LLM in this training stage is:

$$\text{"<unimodal input><text>"}$$

where *<unimodal input>* is the sequence of unimodal embeddings from the perceiver, which can be regarded as soft prompts. The LLM directly tasks *<unimodal input>* as input and is trained to output the corresponding text *<text>* in the training samples. We train for 150k steps in this training stage with a batch size of 256 on 8 A100 GPUs.

Stage2: Multimodal Instruction Tuning After aligning unimodal data with LLM, our model already has the basic ability to understand information from various modalities. However, the model still needs to improve its ability for processing different modalities and following human instructions. Some previous methods [43, 58, 53, 14] have proved that tuning the large model with instructions can help it to understand the intent of human beings. Inspired by these methods, in the second training stage, we further instruction-finetune ChatBridge to align the model with user intent on a wide range of multimodal tasks, enabling more effective zero-shot generalization on multimodal tasks. To this end, we carefully collect a multimodal instruction tuning dataset to finetune our model, where the instructions are multimodal containing text, image, video, and audio, while the responses are text only. A specific introduction to the data collection process will be provided in Section 3.3. We organized all the samples in a standardized format as below:

“###Human: <multimodal input prompt>, <instruction>###Assistant: <response>”

where <multimodal input prompt> is some human-craft template prompt that combines multiple sequences of unimodal embeddings from the perceivers. The LLM ingests the whole sequence and is trained to output the correct response <response>. This training stage costs 10k steps with a batch size of 4k tokens on 8 A100 GPUs.

3.3 Multimodal Instruction Tuning Dataset — MULTIS

We have developed a diverse dataset for multimodal instruction-tuning, named MULTIS, to instruction-finetune ChatBridge model. MULTIS consists of two distinct parts: task-specific data and multimodal chat data. The former presents standardized tasks that require concise responses, while the latter simulates real-world problem-solving scenarios by featuring open-ended dialogue between a human and a multimodal assistant. As shown in Figure 3, the whole collection of MULTIS covers 16 multimodal task categories and 15 source datasets. We hold out 6 datasets for model evaluation purposes.

3.3.1 Task-Specific Data

We collect a vast array of publicly available human-annotated multimodal datasets and transform them into a unified instruction tuning format. Specifically, a plethora of common Question-Answering (QA) and captioning datasets that contain image-text, video-text, and audio-text pairs are assembled, encompassing VQAv2[22], VG-QA[28], COCO Caption[34], MSRVTTQA[52], MSRVTT Caption[52], AudioCaps[26]. For each task, we employ ChatGPT [40] to derive 10~15 unique instruction templates, which are then manually filtered and refined to ensure rationales and diversity are optimal. As the public datasets inherently favor shorter responses, we craft instruction template modifiers to specify the desired response style, such as *short* and *brief* for short-answer data, and *a sentence* and *single sentence* for caption data.

3.3.2 Multimodal Chat Data

While task-specific data empowers the model towards completing standardized tasks, multimodal chat data offers real-world, open-ended dialogues demanding more sophisticated intent comprehension and contextual reasoning abilities, as well as providing more diverse, helpful, human-like responses. Despite the image-to-text chat dataset generated by LLaVA-Instruct-150K [35], chat data across other modalities remains limited. To this end, we have constructed a multimodal chat dataset that comprises both unimodal and multimodal inputs of image, video, and audio modalities.

We adopt LLaVA-Instruct-150K [35] as the image chat data. To incorporate additional modalities, namely video, audio, and video-audio content, we design a pipeline with the assistance of offline open-source models[27, 10, 60, 31, 69] and ChatGPT[40], as is shown in Figure 4. Following [35], we generate three types of instruction-following data including conversation, detailed description, and

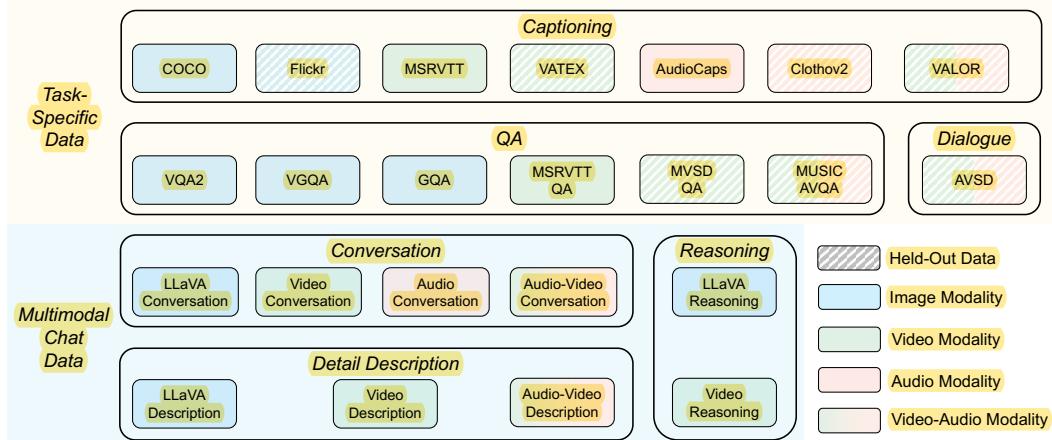


Figure 3: Tasks groups included in MULTIS multimodal instruction tuning dataset.

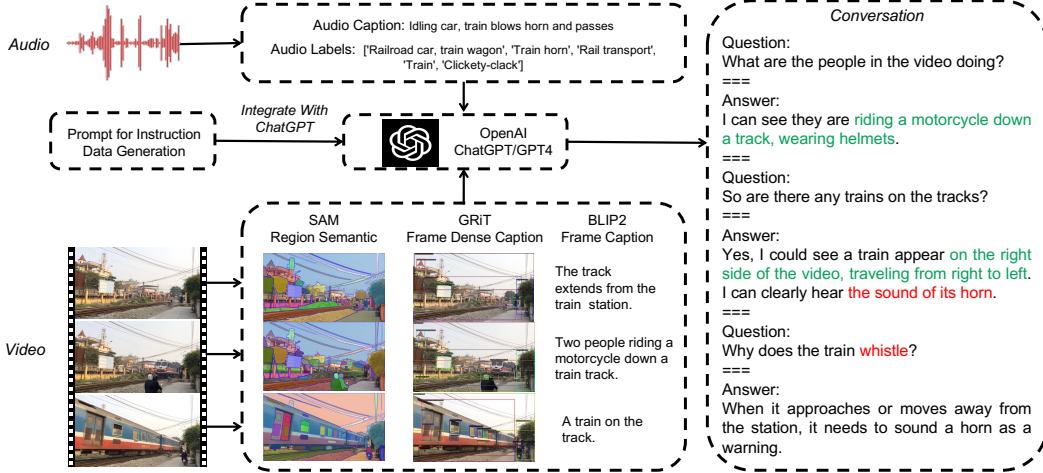


Figure 4: Illustration of the pipeline to collect the multimodal chat data for MULTIS.

complex reasoning. We obtain our video from the MSRVTT [52] dataset, audio from AudioCaps [26], and video-audio from VALOR [12]. In order to prompt a text-only ChatGPT or GPT-4, we convert those non-textual modalities into textual descriptions. Specifically, for each video, we extract three frames and employ Semantic-SAM[10], GRIT[60], and BLIP-2 [31] to develop annotations of the frames at the region semantic, region caption, and frame caption levels. We then concatenate these frame annotations in temporal order with the human-annotated video captions utilizing hand-crafted prompts. For each audio, we use its audio captions and labels from the original dataset. We combine those collections of fine-grained and global descriptions with manually designed seed examples to query ChatGPT or GPT-4 in an in-context-learning manner. With the above pipeline, we collect video, audio, and video-audio multimodal chat samples comprising 24k in conversations, 18K in detailed descriptions, and 9k in complex reasoning.

Overall, MULTIS contains 4.4M task-specific samples and 209k multimodal chat samples.

4 Experiments

4.1 Zero-shot Task Evaluation

We evaluate ChatBridge’s zero-shot ability on the held-out datasets of MULTIS’s task-specific data. The model is instructed with *unimodal* and *multimodal* inputs along with text instructions to generate the corresponding answers.¹

Unimodal Input Tasks Unimodal input tasks comprise of question answering (QA) and captioning tasks on image-text (OKVQA [37], GQA [24], Flickr Captioning [45], nocaps [1]), video-text (MSVDQA [52], VATEX [57]), and audio-text (clothoV2 [19]) datasets. QA tasks require the model to predict a short answer about the unimodal input, while captioning tasks require outputting a sentence description. As shown in Table 1, ChatBridge exhibits remarkable performance on unimodal input tasks, indicating successful alignment of unimodal input and language. On image-text datasets, our method achieves comparable performance as advanced image-based methods, Flamingo [3] and BLIP-2 [31], and achieves new zero-shot state-of-the-art (SoTA) on Flickr30k and VATEX captioning tasks.

Multimodal Input Tasks Multimodal input tasks necessitate the ability in interpreting, correlating, and reasoning about cross-modal information. We evaluate on multimodal input tasks encompassing audio-visual question answering (MUSIC-AVQA [30]), audio-visual dialogue (AVSD [2]), and audio-visual captioning (VALOR [12]) tasks. These tasks analyze videos containing both visual and auditory content. Since our model represents a pioneering approach in handling multi-modal inputs, we perform ablation on the input modalities, as shown in Table 2. Our model achieves

¹Please note that here “unimodal input” and “multimodal input” refer to the use of text and either a single or multiple modalities in image, video, and audio.

Table 1: Zero-shot evaluation of SoTA methods on unimodal input tasks. We report the accuracy for QA tasks and the CIDEr [55] score for captioning tasks.

Methods	Image-Text Tasks				Video-Text Tasks			Audio-Text Tasks	
	OKVQA QA	GQA QA	Flickr30k Caption	NoCaps Caption	MSVD QA	VATEX Caption	Clothov2 Caption		
Finetuned SoTA	66.1 [18]	65.1 [70]	67.4 [72]	121.6 [31]	60.0 [12]	95.8 [12]	48.8 [38]		
Flamingo-9B [3]	44.7	-	61.5	-	30.2	39.5	-		
Flamingo-80B [3]	50.6	-	67.2	-	35.6	46.7	-		
BLIP-2 (FlanT5-XXL) [17]	-	42.4	73.7	98.4	34.4	-	-		
BLIP-2 (Vicuna-13B) [17]	-	32.3	71.6	103.9	20.3	-	-		
ChatBridge w/o MULTIS	41.4	37.4	77.7	107.5	23.5	47.7	22.4		
ChatBridge	45.2	41.8	82.5	115.7	45.3	48.9	26.2		

Table 2: Zero-shot evaluation of the effect of multimodal inputs on multimodal input tasks.

Input Modality	AVSD Dialogue		VALOR Captioning		MUSIC-AVQA
	BLEU-4	CIDEr	BLEU-4	CIDEr	Acc.
Finetuned SoTA	40.0 [44]	108.5 [44]	9.6 [12]	61.5 [12]	78.9 [12]
Video	28.3	73.1	2.8	22.3	33.1
Audio	20.2	46.2	0.3	5.2	28.9
Video+Audio	29.8	75.4	4.2	24.7	43.0

better performance across all three tasks when incorporating both video and audio for solving these tasks, validating its capability to correlate and cooperate different modalities. We observe that video information has a greater influence on video-audio tasks, resulting in relatively higher performance for the video-only input modality as opposed to the audio-only counterpart. Nonetheless, amalgamating audio and video details enhances performance across all tasks to varying extents. Consequently, there is potential for further exploration in constructing a modality-balanced multimodal evaluation benchmark.

Effect of Instruction Tuning with MULTIS Data We also conduct a comparison of the performance in zero-shot task evaluation before and after applying our multimodal instruction tuning training stage. As depicted in Table 1, after instruction tuning, the model demonstrates varying degrees of performance improvement across different downstream tasks. Specifically, we observe a 21.8% boost in accuracy in MSVDQA, a 3.8% improvement in OKVQA, and a 3.6% improvement in GQA. Moreover, our approach also yields advancements in captioning tasks concerning Cider Score. The experimental results demonstrate that the utilization of multimodal instruction tuning data within MULTIS can facilitate the model’s abilities in integrating diverse modalities and generalizing effectively to unseen tasks.

4.2 Multimodal Chat Evaluation

Chat with Unimodal Input We conduct a comparative analysis of our model with three image-based LLMs: BLIP-2[31], LLaVA[35], and MiniGPT-4[73]. The evaluation is based on GPT-4 generated image-text chat data provided by LLaVA[35], comprising 90 samples. We follow the evaluation protocol proposed by LLaVA[35], where GPT-4 is used to evaluate the helpfulness, relevance, accuracy, and level of detail of each model’s responses. The responses are scored on a scale of 1 to 10, where a higher score indicates superior overall performance. As depicted in Table 3a, our method surpasses other approaches that also employ a frozen LLM, *i.e.* BLIP-2 and MiniGPT-4. Nevertheless, our approach still falls short of the performance demonstrated by LLaVA, which fine-tuned the LLM end-to-end. As an avenue for future research, enhancing the model’s capability to process unimodal input with efficacy could be considered.

Chat with Multimodal Input We also examine the ability of ChatBridge to perform human-assistant chat on multimodal inputs. Firstly, we construct a high-quality test set of multimodal chat data by following a similar pipeline as our MULTIS. Specifically, we employ GPT-4 [41] to produce a pool of candidate samples, where we manually choose 90 high-quality samples based on the following criteria: diversity of instructions, correctness of responses, and reliance on different modalities. The selected samples are categorized into three types: reasoning (solving problems through logical thinking and analysis), understanding (comprehending information and interpreting its meaning), and knowledge-based (applying specific knowledge to answer questions accurately).

Table 3: Multimodal chat evaluation results. Response qualities of different methods are assessed by GPT-4 (text-only).

(a) Multimodal chat evaluation on **image-text** chat data.

Methods	Reason-ing	Description	Conver-sation	Overall
LLaVA [35]	8.87	7.07	7.63	7.86
BLIP-2 [31]	5.80	6.00	7.03	6.28
MiniGPT-4 [73]	7.53	7.27	5.63	6.14
ChatBridge	7.17	6.23	7.23	6.88

(b) Multimodal chat evaluation on **video-audio-text** chat data.

Input Modality	Under-standing	Reason-ing	Know-ledge	Overall
Video	5.86	5.27	7.70	6.15
Audio	2.43	3.77	8.09	4.24
Video+Audio	6.10	6.73	8.43	6.87

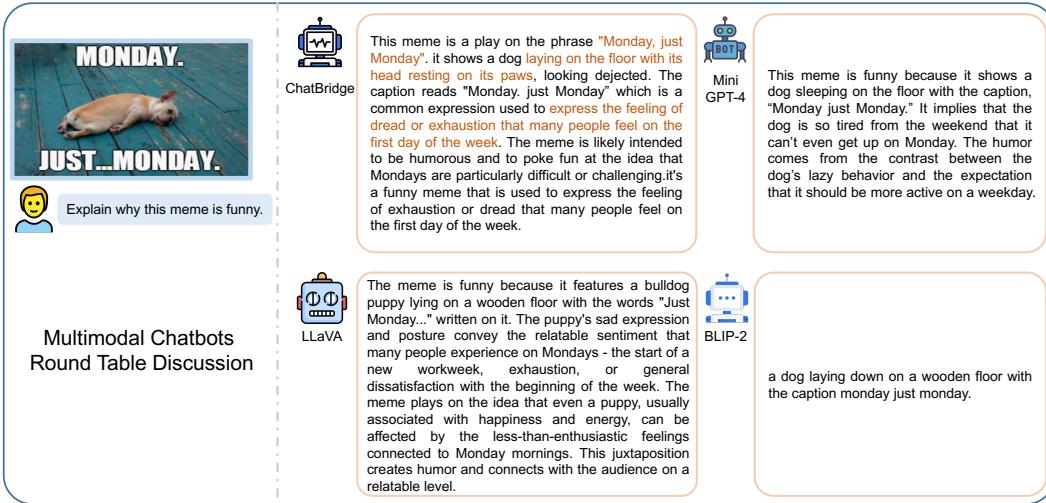


Figure 5: A comparison of SoTA models’ responses to an image-based reasoning task.

To illustrate the effectiveness of our model in processing multimodal inputs, we primarily compare the performance of our model when exposed to different modality inputs. Similar to the above evaluation, we use GPT-4 to score the responses of each method. According to the evaluation results reported in Table 3b, the model incorporating both audio and video inputs outperforms the model relying solely on unimodal inputs on all three sub-tasks. This finding underscores the model’s ability to effectively harness the complementary information inherent in each modality, thereby facilitating a more comprehensive comprehension of the data.

Qualitative Analysis Furthermore, Figure 5 depicts ChatBridge’s ability to generate more comprehensive responses when both video and audio modalities are present, providing evidence for its competence in co-relating and cooperating different modalities. Finally, the multi-round conversation cases illustrated in Figure 6 and Figure 8 demonstrate ChatBridge’s efficacy in following human instructions and comprehending both unimodal and multimodal contents.

In Figure 5, we show an image-base reasoning case to SoTA image-based LLMs, and compare their responses. Our ChatBridge accurately recognizes the text in the image and almost perfectly understands the humor of the image. Furthermore, Figure 5 depicts ChatBridge’s ability to generate more comprehensive responses when both video and audio modalities are present, providing evidence for its competence in co-relating and cooperating different modalities. Finally, the multi-round conversation cases illustrated in Figure 6 and Figure 8 demonstrate ChatBridge’s efficacy in following human instructions and comprehending both unimodal and multimodal contents.

5 Discussions

This paper proposes ChatBridge, a multimodal language model capable of interpreting, correlating, and reasoning about various modalities through two-modality data paired with language. We introduce MULTIS, a multimodal instruction tuning dataset, to instruction-tune our model, which covers

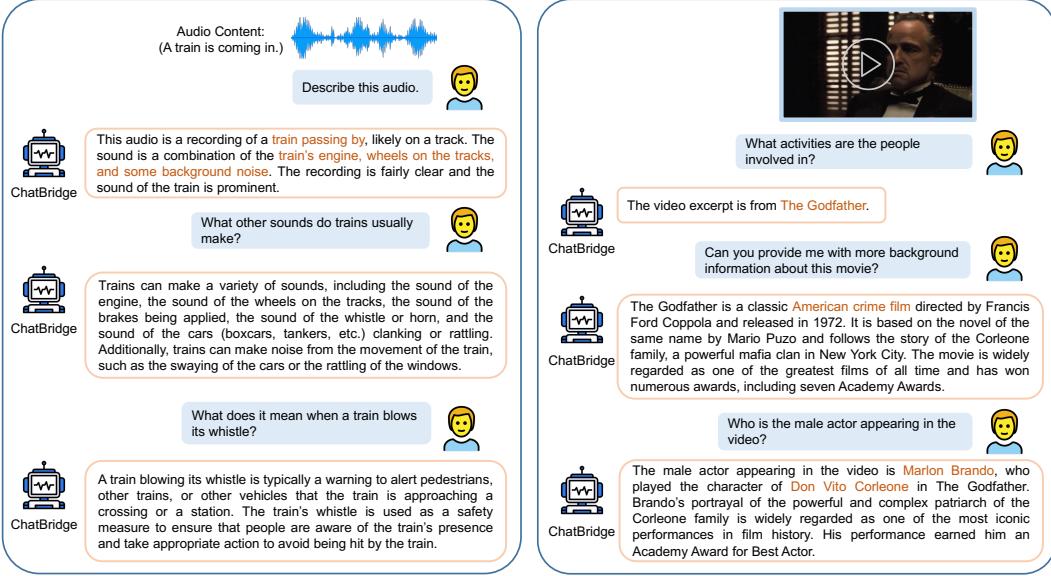


Figure 6: Multi-round conversation cases with audio or video inputs.



Figure 7: A comparison of ChatBridge’s responses to different input modalities with a video-audio input.

a wide range of multimodal tasks in text, image, video, and audio modalities. Consequently, our model exhibits strong and noteworthy quantitative results on zero-shot multimodal tasks. Moreover, qualitative examples demonstrate ChatBridge’s diverse capabilities in following multimodal instructions, such as complex multimodal reasoning, knowledge-grounded multimodal understanding, and multi-turn conversations. These contributions and findings may pave the way for future research into building general-purpose multimodal assistants.

Limitations Here, we describe some limitations of our model as well as opportunities for further improving our model. Specifically, we note the following: 1) We observe that our model exhibits weaknesses in understanding and grounding long-range videos and audios, necessitating a more precise temporal modeling approach. 2) Our framework can be extended to incorporate additional modalities, such as sketch and point cloud. 3) Though the frozen modules in our framework alleviate computational burden, they may lead to insufficient performance and also introduce prior biases from pretrained models.

References

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.
- [2] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019.
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.
- [5] Max Bain, Arsha Nagrani, Gülgün Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.
- [8] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.
- [9] Huai Chen, Yuxiao Qi, Yong Yin, Tengxiang Li, Xiaoqing Liu, Xiuli Li, Guanzhong Gong, and Lisheng Wang. Mmfnet: A multi-modality mri fusion network for segmentation of nasopharyngeal carcinoma. *Neurocomputing*, 394:27–40, 2020.
- [10] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. <https://github.com/fudan-zvg/Semantic-Segment-Anything>, 2023.
- [11] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.
- [12] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.
- [13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.
- [14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- [15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [16] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [17] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

- [18] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- [19] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020.
- [20] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 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. IEEE, 2017.
- [21] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. *arXiv preprint arXiv:2305.05665*, 2023.
- [22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- [23] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.
- [24] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- [25] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [26] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.
- [27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
- [28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
- [29] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.
- [30] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022.
- [31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [32] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.
- [33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

- [34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [36] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021.
- [37] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.
- [38] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuxian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*, 2023.
- [39] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 1–18. Springer, 2022.
- [40] OpenAI. Chatgpt. Technical report, 2023. <https://openai.com/blog/chatgpt>.
- [41] OpenAI. Gpt-4 technical report. Technical report, 2023. <https://arxiv.org/abs/2303.08774>.
- [42] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.
- [43] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [44] Hoang-Anh Pham, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. Video dialog as conversation about objects living in space-time. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 710–726. Springer, 2022.
- [45] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [47] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- [48] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [49] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [50] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.

- [51] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.
- [52] Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. Learning to discretely compose reasoning module networks for video captioning. *arXiv preprint arXiv:2007.09049*, 2020.
- [53] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [55] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [56] Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*, 2023.
- [57] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.
- [58] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [59] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.
- [60] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.
- [61] Zongwei WU, Zhuyun ZHOU, Guillaume Allibert, Christophe Stoltz, Cédric Demonceaux, and Chao Ma. Transformer fusion for indoor rgb-d semantic segmentation. Available at SSRN 4251286, 2022.
- [62] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2022.
- [63] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.
- [64] Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. i-code: An integrative and composable multimodal learning framework. *arXiv preprint arXiv:2205.01818*, 2022.
- [65] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [66] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.
- [67] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [68] Yang Zhang, Yang Yang, Chenyun Xiong, Guodong Sun, and Yanwen Guo. Attention-based dual supervised decoder for rgbd semantic segmentation. *arXiv preprint arXiv:2201.01427*, 2022.
- [69] Hengyuan Zhao. Image2paragraph. <https://github.com/showlab/Image2Paragraph>, 2023.

- [70] Zelin Zhao, Karan Samel, Binghong Chen, et al. Proto: Program-guided transformer for program-guided tasks. *Advances in Neural Information Processing Systems*, 34:17021–17036, 2021.
- [71] Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *International Conference on Machine Learning*, pages 12736–12746. PMLR, 2021.
- [72] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.
- [73] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A More Details about MULTIS Dataset

Modality	Dataset	Type	Held-Out	Description
Image	COCO Caption [34]	Caption	✗	We use Karpathy [25] split, which divides caption data into 82k/5k/5k images for train/val/test sets, with each image corresponding to 5 captions.
	Flickr30k Caption [45]	Caption	✓ (test)	We use test split in Flickr30k dataset, which contain 1k images, with each image corresponding to 5 captions.
	NoCaps [1]	Caption	✓ (val)	NoCaps dataset consists of 166,100 human-generated captions describing 15,100 images from the Open Images [29] validation and test sets. We use the val set in NoCaps dataset.
	VQAv2 [22]	QA	✗	VQAv2 dataset contains 265,016 images, with 5.4 questions on average per image. Each question has 10 ground truth answers.
	QA [28]	QA	✗	Visual Genome consists of 101,174 images from MSCOCO [34] with 1.7 million QA pairs, with 17 questions per image on average.
	GQA [24]	QA	✓ (test-dev)	GQA dataset is a visual question answering dataset with real images from the Visual Genome [28] dataset and balanced question-answer pairs. We use the balanced test-dev split in GQA dataset.
Video	OK-VQA [37]	QA	✓ (test)	OK-VQA is a dataset for visual question answering that requires methods that can draw upon outside knowledge to answer questions. OK-VQA dataset contains 14,055 open-ended questions and with 5 ground truth answers per question. We use the test split in OK-VQA dataset.
	MSR-VTT Caption [52]	Caption	✗	MSR-VTT dataset consists of 10,000 video clips from 20 categories, and each clip is annotated with 20 sentences. The standard splits use 6,513 clips for training, 497 clips for validation, and 2,990 clips for testing.
	VATEX Caption [57]	Caption	✓ (test)	We use the test split in VATEX dataset, which contains 6000 videos with 10 captions per video.
Audio	MSR-VTT QA [52]	QA	✗	MSR-VTT QA dataset consists of about 158k video QA pairs.
	MSVD QA [8]	QA	✓ (test)	We use test split in MSVD QA dataset, which consists 13k video QA pairs.
	AudioCaps [26]	Caption	✗	AudioCaps dataset consists of about 46K audio clips to human-written text pairs collected via crowd-sourcing on the AudioSet [20] dataset.
Video-Audio	ClothoV2 [19]	Caption	✓ (test)	Clotho consists of 6974 audio samples, and each audio sample has 5 captions. We use the val split in Clotho datasets.
	VALOR32K [12]	Caption	✓ (test)	VALOR-32K is an audio-visual captioning dataset. In this dataset, each video corresponds to a caption consisting of both audio and visual contents. VALOR-32K is split into 25K/3.5K/3.5K videos for training, validation, and testing. We use the test split for evaluation.
	MUSIC AVQA [30]	QA	✓ (test)	MUSIC AVQA is an audio-visual question answering (AVQA) dataset, which aims to answer questions regarding different visual objects, sounds, and their associations in videos. It contains 9.3K videos. We use the test split for evaluation.
	AVSD [2]	Dialogue	✓ (DSTC7 test)	AVSD is a audio-visual dataset for dialogue understanding, which aims to generate responses in a dialog about a video, given the dialog history and audio-visual content of the video. We use the test split in DSTC7 Track for evaluation.

Table 4: Compositions of task-specific data in our multimodal instruction dataset MULTIS.

Modality	Source Dataset	Type	#Samples	Description
Image	MSCOCO [34]	Detailed Description	26k	A rich and comprehensive description for an image
		Conversation	58k	Multi-turn conversation between the assistant and a person asking questions about the image
		Reasoning	77k	In-depth reasoning questions about image contents which requires a step-by-step reasoning process to answer
Video	MSRVTT [52]	Detailed Description	10k	A rich and comprehensive description for a video, including more motion information
		Conversation	10k	Multi-turn conversation between the assistant and a person asking questions about the video
		Reasoning	10k	In-depth reasoning questions about video contents which requires a step-by-step reasoning process to answer
Audio	AudioCaps [26]	Conversation	9k	Multi-turn conversation between the assistant and a person asking questions about the audio
Video-Audio	VALOR [12]	Detailed Description	9k	A rich and comprehensive description for a video together with its background audio
		Conversation	9k	Multi-turn conversation between the assistant and a person asking questions about the video and its background audio

Table 5: Compositions of multimodal chat data in MULTIS. We use the image instruction dataset provided by LLaVA [35] as the chat data for image modality.

Modality	Type	Examples of Instruction Templates
Image	Caption	Generate a brief sentence to describe the content of the image. Write a single sentence that conveys what the image depicts. Summarize the image content in a single sentence.
	QA	Use the visual aid to respond to the question briefly: <QUESTION> Analyze the picture and provide a brief answer to <QUESTION> Use the information presented in the image to shortly answer <QUESTION>
	Caption	Generate a concise description for this video. Give a brief overview of the information presented in this video. Write a short summary that effectively conveys the main message of this video.
Video	QA	Based on the information presented in the video, provide a short answer to question <QUESTION> Analyze the video and provide a one-word answer to question <QUESTION> With the aid of the given video, what is your simple answer to <QUESTION>
	Caption	Listen to this audio and summarize its content in one sentence. Write a succinct summary of the key takeaways from this audio. After listening to the audio, generate a one-sentence overview of its main ideas.
	QA	Combining the audiovisual information of this video, generate a sentence to describe its content. Synthesize the audio and visual data in this video to create a sentence that encapsulates its meaning. Describe the content of this video by integrating the audio and visual elements into a single sentence.
Video-Audio	Caption	Based on the video and audio, could you provide a short answer to question: <QUESTION> Utilizing the video and audio content, briefly respond to <QUESTION> Analyze the video and audio and give me a short answer about <QUESTION>
	Dialogue	Based on the video and audio, answer my following questions. Synthesize the audio and visual data in this video and answer my questions. Given this video together with its background audio, answer my next questions.

Table 6: Examples of instruction templates for constructing task-specific data in MULTIS. We show 3 examples for each task. We use prompts: "Give image: <image input>.", "Give audio: <audio input>.", "Give video: <video input>.", and "Give video: <video input> and its background audio: <audio input>." to deal with different modality inputs.

B Example Samples of MULTIS Dataset

Audio Conversation Sample1:



Audio Content: (Idling car, train blows horn and passes.)

Human	Where is the audio recording location?
Assistant	It should be near the train station, because I hear the sound of the train whistle.
Human	Do you hear any other voices?
Assistant	Yes, I hear the car engine idles.

Audio Conversation Sample2:



Audio Content: (A dog is barking and howling.)

Human	Can you hear the sounds of animals?
Assistant	Yes, I heard a dog's voice.
Human	Do you have further detail about it?
Assistant	This dog seems very angry because it is barking and howling.

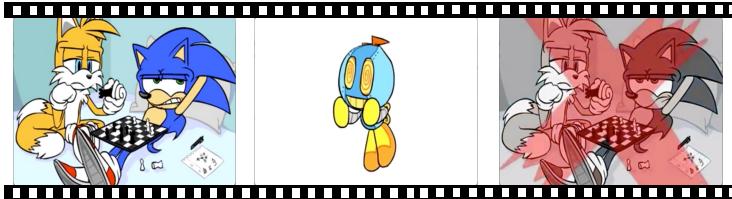
Table 7: Examples of audio-input samples in MULTIS. We use red color to highlight audio-related content in generated data.

Video Conversation Sample:



Human	What is the primary action happening in the video?
Assistant	In this video, a man is making a cup of noodles in the kitchen .
Human	What kind of cup is he using?
Assistant	He's using a styrofoam cup to make the noodles.
Human	Is the man adding any other ingredients to the cup?
Assistant	Yes, the man is also adding onions to the cup of noodles.
Human	What is the likely purpose of this video?
Assistant	The purpose of this video is to demonstrate a simple, quick meal preparation process using a styrofoam cup and noodles, possibly to serve as a tutorial for others who would like to give it a try.

Video Detailed Description Sample:



Human	What are the key elements in this video?
Assistant	The video showcases various animated characters and scenes in different settings. It starts with Sonic and Tails playing a game of chess while surrounded by a graffiti-covered wall . Then, a scene shifts to another cartoon character wearing a blue and yellow hat , flying a kite with a smiling face , and an orange flag atop a pole . The video transitions to a black and white scene featuring a red cross . Sonic the Hedgehog is featured in the video too while advertising a new game.

Video Reasoning Sample:



Human	What can be inferred about the man's profession and the purpose of his discussion?
Assistant	Based on the video, the man is dressed in a suit and tie sitting at a table surrounded by papers . He is talking about the Kentucky Derby and horse racing . It can be inferred that he is a professional sports analyst or an expert in horse racing. The purpose of his discussion is to inform the viewers about the race, share his insights on the competition, handicap potential outcomes, and provide interesting background information on the event.

Table 8: Examples of video-input samples in MULTIS. We use **green** color to highlight video-related content in generated data.

Video-Audio Conversation Sample1:



Audio Content: (A man speaks earnestly.)



Human	What is the man at the podium doing?
Assistant	The man at the podium is speaking earnestly , probably giving a speech or addressing an audience .
Human	Please describe the man's outfit.
Assistant	The man is wearing a suit and tie , and the tie is red and grey striped .
Human	What else can be seen in the room?
Assistant	In the room, there is a man standing, a brown seat back , a table with a stack of papers and a leather chair .

Video-Audio Conversation Sample2:



Audio Content: (Large group of people singing and clapping.)



Human	What type of event is taking place in the video?
Assistant	A choir is performing at a church event.
Human	Can you describe some of the people in the choir?
Assistant	Certainly! Some members of the choir include a person wearing a pink shirt , a woman wearing a pink jacket , a person in a plaid shirt , and a person in a red shirt.
Human	What is the audience's reaction to the choir's performance?
Assistant	The audience sounds very appreciative, as I can hear a large group of people clapping .

Video-Audio Detailed Description Sample:



Audio Content: (A busy road with traffic going by and then a vehicle honks a horn.)



Human	Write a detailed description of the given video.
Assistant	In this video, a busy city street is filled with the sounds of traffic and a honking horn . People are walking on the sidewalk with some wearing jackets. A white and blue bus passes by, and there are various objects along the sidewalk, such as a gray trash can and a concrete planter . The street is surrounded by buildings.

Table 9: Examples of video-audio input samples in MULTIS. We use **green** and **red** colors to highlight video- and audio-related content in generated data.

C More Case Studies of ChatBridge



What is the name of the food in the picture?

 ChatBridge The food in the picture is called "Chinese Fried Chicken".

 ChatBridge How can I make it?

To make Chinese Fried Chicken, you will need the following **ingredients**: chicken pieces, flour, eggs, vegetable oil, salt, pepper, and any desired seasonings. First, **coat the chicken pieces in flour**, then **dip them in beaten eggs**, and finally **coat them in a mixture of flour and seasonings**. Heat up some vegetable oil in a pan and fry the chicken until it is golden brown. Serve it hot with your favorite dipping sauce.



Can you introduce me some information about this picture?

 ChatBridge This painting is a famous work by **Vincent van Gogh**, titled "**The Starry Night**". It depicts a night sky with stars and a cypress tree in the foreground. The painting is known for its **vibrant colors** and swirling **brushstrokes**, which convey a sense of movement and energy. The painting is considered one of van Gogh's most famous works, and it is widely recognized as a masterpiece of modern art.

Figure 9: Providing advice and step-by-step instructions about an image.

Figure 10: Introducing the background knowledge about an image.



Why is this character loved by children?

 ChatBridge This character is a **cartoon pig** who is a **mascot for a candy company**. The character is dressed in a **red and yellow suit**, and is **holding a lollipop**. The character is also known for its cheerful and playful personality, which makes it a popular choice for children.

 ChatBridge Please write a few slogans for him to highlight its characteristics.

1."Sweet and playful, just like you!"
2."Piggy Pops, the candy that's always there for you!"
3."Piggy Pops, the tasty treat that brings joy to your day!"
4."Piggy Pops, the candy that's always there to make you smile!"
5."Piggy Pops, the candy that's as sweet as you are!"



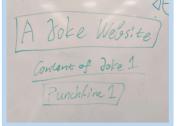
THE FOURTH OF MAY
BE WITH YOU

Can you explain this meme?

 ChatBridge This meme is a **Star Trek themed joke**. The image shows a man in a Star Trek uniform, with a caption that reads "**The fourth of May be with you.**" This is a play on words, as the phrase "**May the force be with you**" is a famous quote from the Star Wars franchise. The joke is that the man is dressed as a Star Trek character, but the caption is a reference to Star Wars.

Figure 11: Reasoning and creating with an image.

Figure 12: Explaining the amusing point in an image.



Write a brief HTML to turn this mock-up into a real website.



Sure, here's a brief HTML code to turn this mock-up into a real website:

```
<!DOCTYPE html>
<html>
<head>
<title>A Joke Website</title>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0">
</head>
<body>
<header>
<h1>A Joke Website</h1>
<nav>
<ul>
<li><a href="#">Home</a></li>
<li><a href="#">About Us</a></li>
<li><a href="#">Contact Us</a></li>
</ul>
</nav>
</header>
<main>
<section>
<h2>Welcome to Our Joke Website</h2>
<p>Our team is dedicated to providing you with the best jokes and humor. Our team of professional comedians and writers work hard to bring you the latest and greatest jokes. So, whether you're looking for a good laugh or just want to brighten up your day, you've come to the right place.</p>
</section>
<section>
<h2>Our Joke Categories</h2>
<ul>
<li>Political Jokes</li>
<li>Relationship Jokes</li>
<li>Workplace Jokes</li>
<li>Dad Jokes</li>
</ul>
</section>
<section>
<h2>Our Joke of the Day</h2>
<p>Check out our daily joke of the day and get a good laugh every day. Our team of comedians and writers work hard to bring you the best jokes every day. So, whether you're looking for a good laugh or just want to brighten up your day, you've come to the right place.</p>
</section>
</main>
<footer>
<p>Copyright © 2023 A Joke Website</p>
</footer>
</body>
</html>
```

This HTML code creates a basic website structure with a header, main content, and footer.

The header includes a title and navigation menu with links to the home, about us, and contact us pages. The main content includes a welcome message, an introduction to the joke categories, and a section for the joke of the day.

Please add a real joke in "Our Joke of the Day".



Sure, here's an example of a joke that could be added to the "Our Joke of the Day" section:

```
<section>
<h2>Our Joke of the Day</h2>
<p>Why did the tomato turn red? Because it saw the salad dressing!</p>
</section>
```

This joke is a classic example of a pun, which is a type of joke that relies on wordplay to create humor. The joke plays on the word "red" being used to describe both the tomato and the salad dressing, creating a humorous contrast.

A Joke Website

- [Home](#)
- [About Us](#)
- [Contact Us](#)

Welcome to Our Joke Website

We are a website dedicated to providing you with the best jokes and humor. Our team of professional comedians and writers work hard to bring you the latest and greatest jokes. So, whether you're looking for a good laugh or just want to brighten up your day, you've come to the right place.

Our Joke Categories

- Political Jokes
- Relationship Jokes
- Workplace Jokes
- Dad Jokes

Our Joke of the Day

Check out our daily joke of the day and get a good laugh every day. Our team of comedians and writers work hard to bring you the best jokes every day. So, whether you're looking for a good laugh or just want to brighten up your day, you've come to the right place.

Copyright © 2023 A Joke Website

A Joke Website

- [Home](#)
- [About Us](#)
- [Contact Us](#)

Welcome to Our Joke Website

We are a website dedicated to providing you with the best jokes and humor. Our team of professional comedians and writers work hard to bring you the latest and greatest jokes. So, whether you're looking for a good laugh or just want to brighten up your day, you've come to the right place.

Our Joke Categories

- Political Jokes
- Relationship Jokes
- Workplace Jokes
- Dad Jokes

Our Joke of the Day

Why did the tomato turn red? Because it saw the salad dressing!

Copyright © 2023 A Joke Website

Figure 13: Generating HTML code based on a hand-drawn draft and modifying the code according to human opinions. We compile the HTML code in the conversation and show the website in the right column.

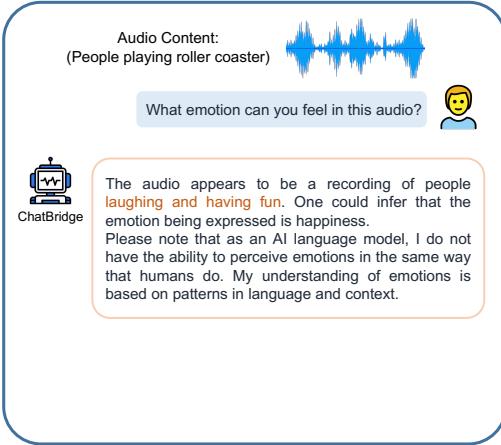


Figure 14: Inferring the emotions behind an audio.



Figure 15: Writing lyrics based on audio.

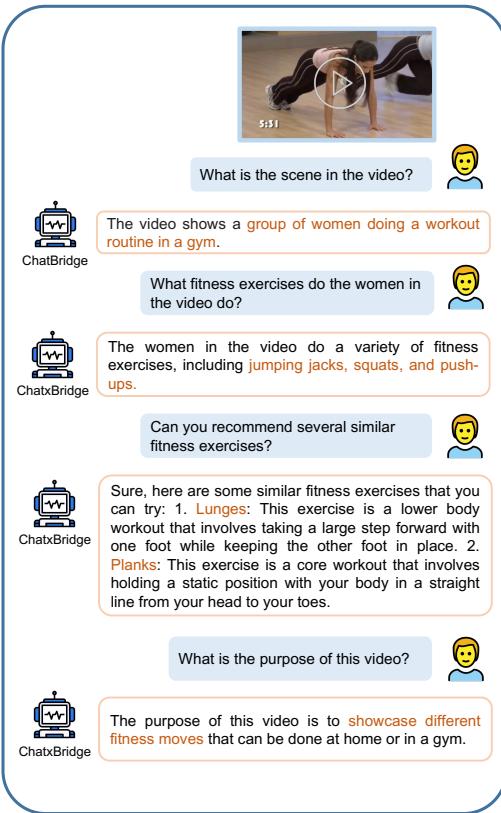


Figure 16: Engaging in informative, multi-turn conversations about a video.

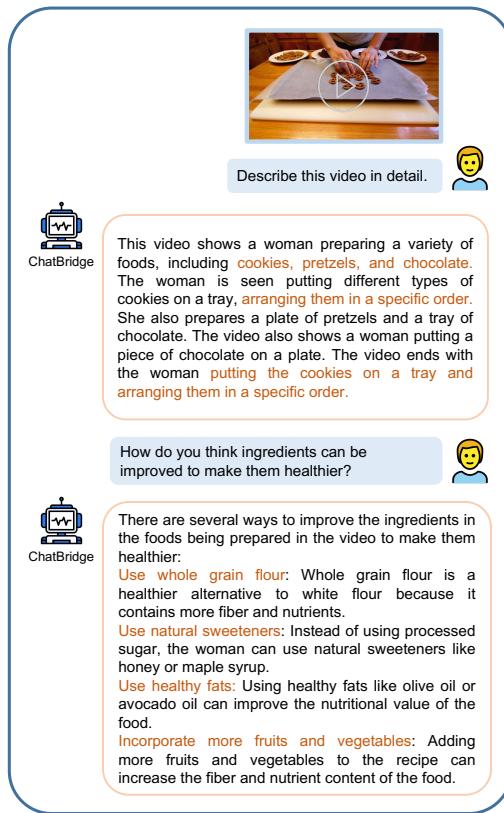


Figure 17: Summarizing and extending the video content.

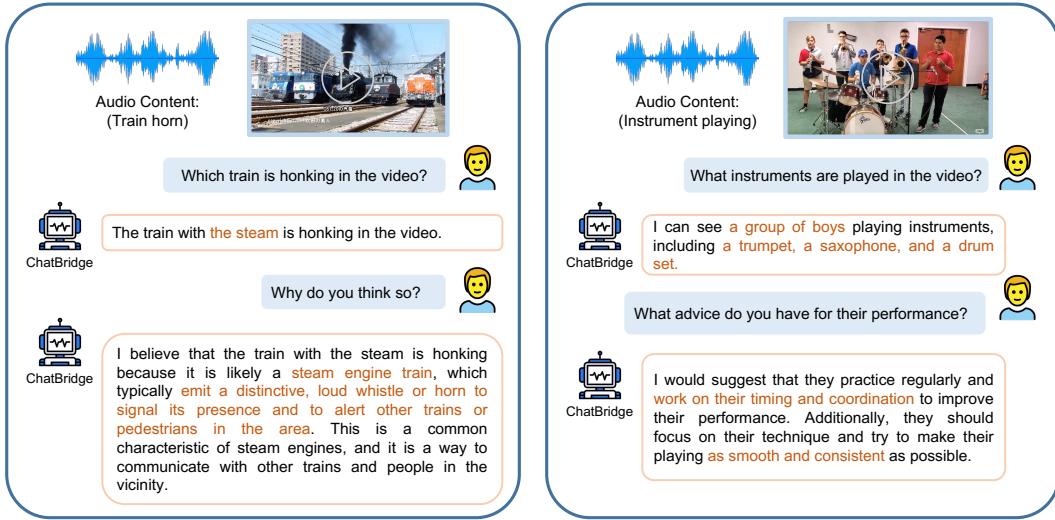


Figure 18: Understanding video with its background audio and chatting with human.

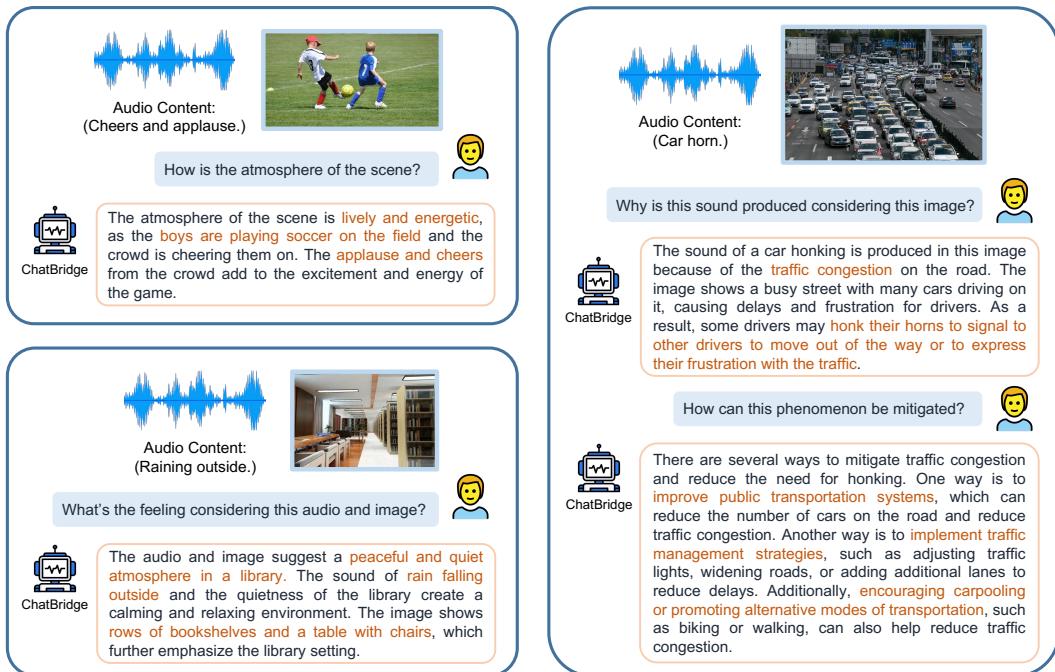


Figure 19: Understanding and reasoning about independent image and audio.



Audio Content:
(Cheers and applause)






Write a story based on them all.



ChatBridge

The boy was **sitting on the floor in front of the tv, watching a soccer match**. suddenly, the ball **came out of the tv and rolled towards him**. he tried to kick it, but missed. the ball rolled towards the tv again, and this time the boy was ready. he kicked the ball with all his might, and it went straight into the goal! the boy **cheered and jumped up and down**, feeling very happy with his achievement.



Audio Content:
(siren)






Write a short news report based on them all.



ChatBridge

A red fire truck is parked on the side of the road, its **siren blaring**. the flames from the fire it is carrying are **shooting high into the sky, engulfing the trees in the surrounding area**. the firefighters inside the truck are ready to tackle any blaze that comes their way.

Figure 20: After completing the first-stage training, which involves multimodal alignment, the ChatBridge model is already capable of comprehending and analyzing multimodal input where all image, video, and audio modalities are present.



What's the similar element between them?

 ChatBridge

The similar element between the two videos is that they both show people preparing food in a kitchen.



What's the difference between them?

 ChatBridge

The difference between the two videos is that the first video is a **real video** of a car driving on a road, while the second video is an **animation** of a car driving on a road. the first video shows the **actual** movement of the car, while the second video is a **computer-generated** representation of the movement of the car. In the first video, the car is driving on a road with **trees and buildings in the background**, while in the second video, the car is driving on a road with **neon lights in the background**. the first video is shot from the perspective of a person in the car, while the second video is shot from a bird's eye view.



What's the same character between them?

 ChatBridge

Both the image of the **red umbrella** in the rain and the image of the **green field** with a bench in the foreground **have a common element of nature**. Both images feature **natural elements** such as rain, clouds, and greenery. the red umbrella in the rain image also has a man-made element in the form of the bench in the background. Both images also **have a sense of stillness and calmness**, with the rain and the green field creating a peaceful atmosphere. The red umbrella in the rain image has a sense of movement as the rain falls around it, while the green field with the bench has a sense of tranquility as the clouds move slowly across the sky. Overall, both images **share a common theme of nature and stillness**, with the red umbrella in the rain image adding a sense of movement and the green field with the bench adding a sense of tranquility.



What's the difference between two images?

 ChatBridge

The first image appears to be a **bird perched on a wooden bench**, while the second image appears to be a **bird perched on a branch**. the bird in the first image appears to **have a brown and white pattern on its feathers**, while the bird in the second image appears to **have a brown and black pattern on its feathers**. additionally, the bird in the second image appears to **have a longer tail than the bird in the first image**.



Write a short story about these images.

 ChatBridge

As the sun began to set on the **baseball field**, the players **took to the field** for the final game of the season. **the crowd was filled with excited fans**, eager to see their favorite players in action. In the top of the first inning, the visiting team **scored a quick run**, but the home team was not about to give up that easily. in the bottom of the inning, the home team's starting pitcher took the mound, ready to show the crowd what he was made of. The game was a close one, with both teams exchanging runs throughout the night. in the bottom of the ninth, the home team was down by one run with two outs. **the crowd held its breath as the batter stepped up to the plate**, hoping for a miracle. To the delight of the crowd, **the batter hit a double, scoring the tying run**. the game went into extra innings, with both teams giving it their all. in the twelfth inning, the home team finally scored the winning run, sending the crowd into a frenzy. As the players left the field, **the crowd cheered and applauded**, grateful for the exciting game they had just witnessed. it was a night that would be remembered for years to come, a true testament to the power of baseball.

Figure 21: ChatBridge model after the first-stage training, *i.e.* multimodal alignment, already has the intrinsic ability to deal with input of multiple images or videos. It can write a story with images, and find the similarities and differences among multiple inputs.