MEERKAT: Audio-Visual Large Language Model for Grounding in Space and Time

Sanjoy Chowdhury*[1], Sayan Nag*[2], Subhrajyoti Dasgupta*[3], Jun Chen[4], Mohamed Elhoseiny[4†], Ruohan Gao[1†], and Dinesh Manocha[1†]

[1] University of Maryland, College Park
[2] University of Toronto
[3] Mila and Université de Montréal
[4] King Abdullah University of Science and Technology (KAUST)
https://github.com/schowdhury671/meerkat
{sanjoyc,rhgao,dmanocha}@umd.edu sayan.nag@mail.utoronto.ca
subhrajyoti.dasgupta@umontreal.ca
{jun.chen,mohamed.elhoseiny}@kaust.edu.sa

Abstract. Leveraging Large Language Models' remarkable proficiency in text-based tasks, recent works on Multi-modal LLMs (MLLMs) extend them to other modalities like vision and audio. However, the progress in these directions has been mostly focused on tasks that only require a coarse-grained understanding of the audio-visual semantics. We present *MEERKAT*, an audio-visual LLM equipped with a fine-grained understanding of image and audio both spatially and temporally. With a new modality alignment module based on optimal transport and a cross-attention module that enforces audio-visual consistency, *MEERKAT* can tackle challenging tasks such as audio referred image grounding, image guided audio temporal localization, and audio-visual fact-checking. Moreover, we carefully curate a large dataset *AVFIT* that comprises 3M instruction tuning samples collected from open-source datasets, and introduce *MEERKATBENCH* that unifies *five* challenging audio-visual tasks. We achieve state-of-the-art performance on all these downstream tasks with a relative improvement of up to *37.12%*.

Keywords: Audio-Visual LLM · AV Localization · AVFIT Dataset

1 Introduction

Large Language Models (LLMs) [3, 14, 15, 58, 71] have demonstrated remarkable performance in various natural language processing tasks, achieving human-level accuracies in comprehension and reasoning abilities. Furthermore, powered by

*Equal contribution.
†Equal advising.

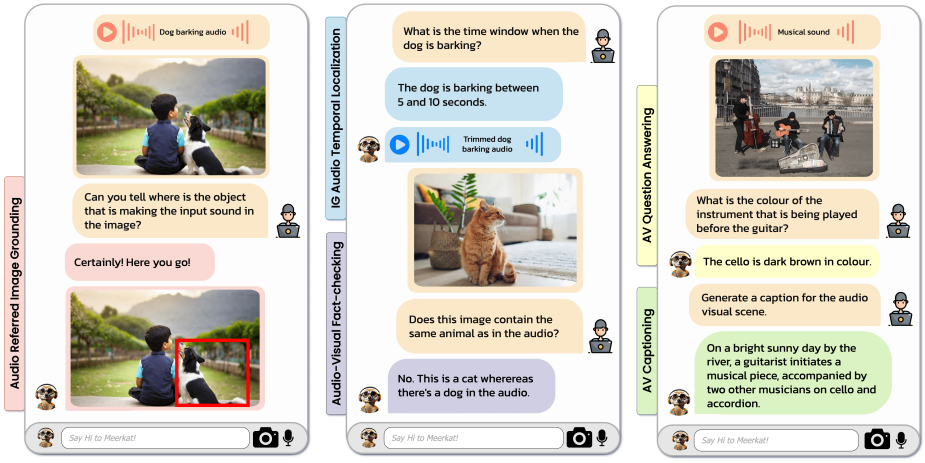

Fig. 1: We present **MEERKAT, an audio-visual LLM that can effectively ground both spatially and temporally in image and audio**. Our model is adept in tasks that require fine-grained understanding such as Audio Referred Image Grounding, Image Guided (IG) Audio Temporal Localization & Audio-Visual (AV) Fact-checking. It can also be extended to perform coarse-grained tasks like AVQA & AV Captioning.

the emergent instruction fine-tuning paradigm [17,49,53], these language models can be equipped to follow open-ended natural language instructions, or even combined with other modalities, especially vision [2,4,24,31,38,42–44,65,82,83,88]. Audio, though often complementary to the associated visual scene, remains largely under-explored in the context of LLMs. Building Multi-modal LLMs (MLLMs) that can *listen* may enable new applications in multimedia content analysis, multi-modal virtual assistants, education and training, etc.

Limited prior works (refer to Tab. 1) have incorporated audio in MLLMs [24,50,63]. However, they mostly focus on coarse-grained tasks such as captioning and question-answering, which is comparatively straightforward to be subsumed into an LLM interface [43, 63, 65, 82]. Although there have been some recent advancements in leveraging MLLMs for grounding [8, 9, 56, 73, 74, 79, 86], they either only focus on the visual modality [8, 9, 30, 56, 79], or struggles to capture fine-grained details occurring within audio-visual events due to insufficient joint modeling of the two modalities [43, 65, 82].

Our goal is to harness the power of LLMs for fine-grained audio-visual understanding. This is challenging mainly because: *(i)* there is a disparity of input and output formats across different tasks (e.g., image grounding from an audio query, image-guided audio temporal localization), *(ii)* no large-scale datasets exist for training audio-visual LLMs with grounding capabilities. Existing audio-visual LLMs [43, 63, 65] are restricted to coarse-grained tasks and do not incorporate cross-modality fusion, which is a crucial component for achieving fine-grained understanding and reasoning capabilities, as shown in [18, 35]. Although there exist individual models capable of handling image grounding (BuboGPT [86])

Model	Audio Types		Output Image Grounding	Output Audio Grounding	End-to-end	Data Features		
	Speech	Open-domain				Convention	GPT-Prompted	Robustness
VideoLlama [82]	✓	✓	✗	✗	✓	✓	✗	✗
Macaw-LLM [43]	✓	✗	✗	✗	✓	✓	✓	✗
PandaGPT [65]	✓	✓	✗	✗	✓	✓	✗	✗
AV LLM [63]	✓	✓	✗	✗	✓	✓	✓	✗
X-InstructBLIP [50]	✗	✓	✓	✗	✓	✓	✗	✗
TimeChat [59]	✓	✗	✗	✓	✓	✓	✓	✗
BuboGPT [86]	✗	✓	✓	✗	✗	✓	✗	✗
Meerkat (ours)	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of Meerkat with recent Audio-Visual LLMs. 'Convention' refers to a collection of publicly available data that has been transformed using templates, 'GPT-Prompted' signifies if the generated instructions are obtained/refined employing GPT, and 'Robustness' is the model's ability to tackle negative samples. We compare our method against these approaches in Sec. 5.

and temporal localization (TimeChat [59]) separately, they are either not suitable for open-domain audio (TimeChat) or are not trained in an end-to-end fashion (BuboGPT) (refer to Tab. 1).

In light of these challenges, we present Meerkat[5] (ref Fig. 1), the first *unified* audio-visual LLM framework that can effectively ground both spatially and temporally in image and audio, respectively. It has two crucial modules that are key to its strong capability in fine-grained understanding: a modality alignment module that learns the cross-modal alignment between image and audio patches in a weakly-supervised manner based on optimal transport, and a cross-modal attention module that is capable of enforcing consistency in the cross-attention heatmaps. Together, these two modules enable learning better joint audio-visual representations that subsequently enhance downstream tasks.

To support Meerkat, we further introduce MeerkatBench that unifies *five* different audio-visual tasks (shown in Tab. 2), including audio referred image grounding, image-guided audio temporal localization, audio-visual fact checking, audio-visual question answering, and audio-visual captioning (see Fig. 1 for examples). To enable the training of these five tasks, we also curate a large dataset AVFIT, which contains 3M instruction tuning samples with various degrees of difficulties for learning fine-grained audio-visual semantics. Extensive experiments on these tasks demonstrate the effectiveness of our proposed model.

In summary, we make the following main contributions:

- We present Meerkat, the first audio-visual LLM equipped with fine-grained spatio-temporal understanding that can ground in image and audio.
- We introduce MeerkatBench that unifies five audio-visual learning tasks, and a new large instruction-tuning dataset AVFIT to enable learning fine-grained audio-visual semantics.
- Evaluating on these five benchmark tasks, we set new state-of-the-art results on all of them with a relative improvement up to *37.12%*.

[5] Meerkats are known for their strong spotting and listening abilities.

Task Granularity	Task Name	Dataset	Train	Test	Spatial Bounding Box	Time Interval	# Samples Train / Test	Metrics
Fine	Audio Referred Image Grounding	Openimages-AudioSet	✓	✗	✓	✗	1.07M / –	–
		Openimages-VGGSound	✓	✗	✓	✗	180K / –	–
		AVSBench†	✓	✓	✓	✗	2.30K / 0.49K	cIOU, AUC
		VGGSS	✗	✓	✓	✗	– / 4.38K	cIOU, AUC
		PASCAL Sound	✗	✓	✓	✗	– / 0.56K	cIOU, AUC
		Flickr-Soundnet	✗	✓	✓	✗	– / 2.78K	cIOU, AUC
	Image Guided Audio Temporal Localization	Openimages-AudioSet Strong	✓	✓	✗	✓	96.5K / 24.1K	F1-score
		LLP	✗	✓	✗	✓	– / 2.32K	F1-score
	Audio-Visual Fact-checking	Openimages-AudioSet	✓	✓	✗	✗	1.18M / 321K	F1-score
Coarse	AV Question Answering	AVQA	✓	✓	✗	✗	40.4K / 16.9K	Accuracy
		Music AVQA	✓	✓	✗	✗	25.7K / 7.36K	Accuracy
	AV Captioning	VALOR	✓	✓	✗	✗	25.0K / 3.50K	B@4, M, R, C

Table 2: Task-wise dataset distribution, dataset details, and metrics. We collect AVFIT, which is a collection of 12 datasets. We denote dataset-wise train/test usage. The visual grounding datasets contain spatial bounding box annotations while the audio temporal localization contains time-interval annotations. We consider audio-visual fact-checking as a fine-grained task as it requires an understanding of spatio-temporal grounding information (refer to Sec. 5.2 for more details). Here B@4: BLUE@4, M: METEOR, R: ROUGE, C: CIDEr. For all our experiments we consider F1@0.5. † We obtain the bounding box from the segmentation maps.

2 Related Works

Multi-modal Large Language Models. Inspired by the success of instruction following capabilities of large language models [14, 49, 67], the community has recently started to leverage LLMs for understanding multi-modal contents. Powered by high-quality multi-modal instructional data, recent methods [2, 4, 9, 31, 38, 54, 65, 88] extend LLMs for multi-modal learning. While some approaches such as MiniGPT4 [88], X-LLM [4], and Video-ChatGPT [44] perform latent alignment between the pre-trained LLM and other modalities via learned visual encoder. Other methods like Otter [31], and LLaMA-Adapter [83] learn cross-attention layers into the LLM to infuse multi-modal information. Prior works in the realm of LLMs predominantly focus on either visual-only inputs [31, 38, 78, 88] or tackle coarse-grained tasks [34, 44] leaving room for fine-grained audio-visual understanding. Unlike prior approaches, in this work, we focus on equipping LLMs with strong audio-visual comprehension abilities.

Fine-grained Multi-modal Understanding. Of late, general-purpose multi-modal large language models have demonstrated their effectiveness in unifying a versatile array of vision-language or video-understanding tasks. These models, powered by LLMs [15, 68, 71, 72, 75, 76, 85] have superior reasoning and understanding capabilities. As a natural extension, MLLMs have been leveraged to unify region-based grounding tasks [8, 9, 30, 54, 73, 74, 79, 84, 86]. Despite significant strides, these models are still limited to fine-grained comprehension within a single modality. In this work, we propose MEERKAT to precisely address this research gap under in-the-wild audio-visual event settings. To this end, we present a novel audio-visual task unification framework which promotes strong multi-modal reasoning and understanding capabilities.

We discuss further related works in the appendix.

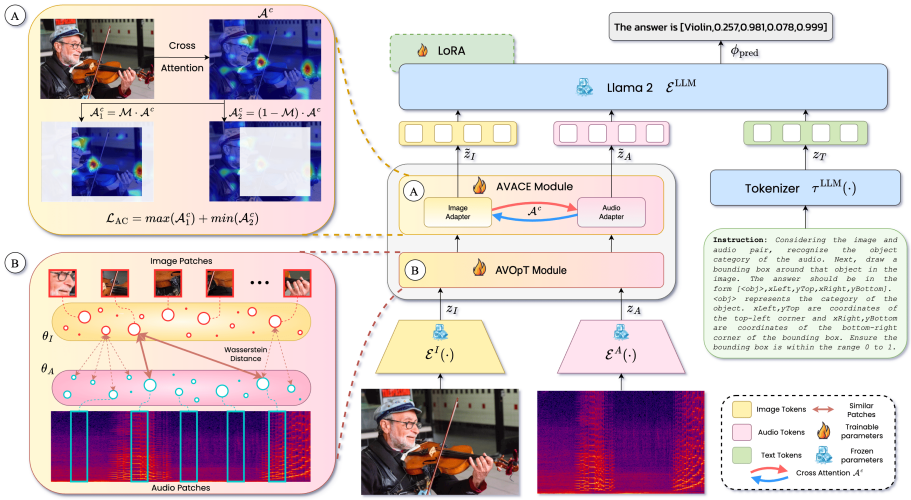

Fig. 2: Overview of Meerkat. Our model is equipped with fine-grained audio-visual comprehension abilities. When fed with image I, audio A pairs, the Audio-Visual Optimal Transport alignment (AVOpT) module (B) learns the patch-wise image-audio association to facilitate weak alignment between the two modalities by minimizing the patch-level Wasserstein distance. Subsequently, the Audio-Visual Attention Consistency Enforcement (AVACE) module (A) maximizes the region-level alignment by confining the cross-modal attention maps around the objects of interest and minimizing the association with the background. After tokenizing the text instruction T, the modality-specific latents $(\tilde{z}_I, \tilde{z}_A, z_T)$ are passed to the instruction tuned Llama 2 model which serves as a unified interface for the downstream tasks. We employ a LoRA-based fine-tuning of the LLM.

3 Methodology

In this section, we introduce Meerkat. Fig. 2 provides an overview of our approach. We first discuss the multi-modal feature extraction in Sec. 3.1. In Sec. 3.2 we introduce our novel audio-visual feature alignment modules. In Sec. 3.3 we add the overall training objective followed by Sec. 3.4 where we elaborate the numerical representations of the visual bounding box and time intervals.

3.1 Multi-modal Feature Extraction

Image Encoder. Given a batch of k input images $\mathbf{I} = \{I_i\}_{i=1}^k : I_i \in \mathbb{R}^{H \times W \times C}$ where H, W, C represent the height, width and channels respectively, we employ a pretrained CLIP-ViT-B/16 [57] encoder $\mathcal{E}^I(\cdot)$ to extract the image embeddings. Where i^{th} image embedding can be represented as $z_I \in \mathbb{R}^{\mathcal{S}_I \times \mathcal{D}_I}$, where $\mathcal{S}_I$ and $\mathcal{D}_I$ denote the number of image tokens and hidden dimension respectively.

Audio Encoder. The audio encoder transforms the raw audio input into an audio embedding. We use the audio transformer backbone from CLAP [19] as

our audio encoder due to its success in diverse audio tasks owing to its superior multi-modal alignment. We leverage this powerful pre-trained encoder ($\mathcal{E}^A(\cdot)$) to extract meaningful audio representations. For a batch of k processed audio inputs $\mathbf{A} = \{A_i\}_{i=1}^{k}$: $A_i \in \mathbb{R}^{F \times T}$ where F is the number of spectral components (e.g. Mel bins) and T is the number of time bins. Each i^{th} audio embedding is denoted as $z_A \in \mathbb{R}^{\mathcal{S}_A \times \mathcal{D}_A}$, $\mathcal{S}_A$ and $\mathcal{D}_A$ are the number of audio tokens and hidden dimension respectively.

LLM. MEERKAT adopts the open sourced Llama 2-Chat (7B) [71] as the large language model backbone. Pre-trained LLMs tokenizer projects the text sequence T into embeddings $z_T \in \mathbb{R}^{\mathcal{S}_T \times \mathcal{D}_T}$, where $\mathcal{S}_T$ and $\mathcal{D}_T$ refer to token length and hidden dimension respectively. Before passing the image and audio embeddings into the LLM, they undergo transformations via additional linear layers to ensure the embedding dimensions across different modalities remain consistent. Since the LLM serve as the unified interface for audio-visual inputs, we rely on the language tokens to carry out the individual tasks.

3.2 Audio-Visual Feature Alignment

Inspired by the success of recent pre-training frameworks in grounding tasks [8, 18, 35], we equip our model with two different levels of supervision: weak supervision through modality alignment module (AVOpT) and strong supervision through audio-visual consistency enforcement module (AVACE). We follow a single-stage training strategy and empirically show our method achieves similar performance compared to two-stage training (more details in the appendix).

Audio-Visual Optimal Transport Alignment Module (AVOpT). Weak supervision as a precursor to fine-grained supervision has been proven to be an effective training strategy in various tasks [18, 33]. Earth Mover Distance based algorithms [81] involving Optimal Transport (OT) methods [10] have been recently leveraged for patch-level alignment between the query and the support images in a siamese network [81]. Furthermore, in the context of vision-language models, OT-based algorithms have been employed for patch-word alignment [13]. Recently proposed VLAP [51] achieves alignment between vision-language by predicting assignments via linearly projecting one modality into the other. As the image (CLIP) and audio (CLAP) encoders are trained separately their learned embeddings are in a different semantic space. Our intuition is that such a patch-level alignment can improve vision and audio semantic consistency [23]. We experimentally demonstrate that this patch-level weak guidance is superior to contrastive loss-based [25, 48] global supervision (more details in appendix).

From a given image I and audio A pair, we obtain patch-level (local) feature embeddings z_I and z_A where, $z_I = \mathcal{E}^I(I); z_A = \mathcal{E}^A(A)$. For modeling cross-modal relations by utilizing the inherent rich semantic structures in these feature representations, we generate two discrete distributions, represented by $\theta_I \in \mathbf{P}(\mathbb{Z}_I)$ and $\theta_A \in \mathbf{P}(\mathbb{Z}_A)$, for image and audio respectively:

$$\theta_I = \sum_{k=1}^{M} u_I(k)\delta_{z_I}(k); \theta_A = \sum_{l=1}^{N} u_A(l)\delta_{z_A}(l) \tag{1}$$

where, $\sum_{k=1}^{M} u_I(k) = \sum_{l=1}^{N} u_A(l) = 1$, u_I and u_A being the respective weight vectors for the probability distributions θ_I and θ_A. δ_z is the Dirac delta function placed at support point z in the embedding space [5]. The goal is to discern the *optimal* transport plan while matching these two distributions. Therefore, we compute the Wasserstein Distance (WD) between these probability distributions θ_I and θ_A while preserving the topological information during the cross-domain alignment process, mathematically given as follows:

$$\mathcal{L}_{\text{OT}} = \mathcal{D}_{\text{Wasserstein}}(\theta_I, \theta_A) = \min_{\mathbf{\Omega} \in \Psi(u_I, u_A)} \sum_k \sum_l \mathbf{\Omega}_{kl} \cdot \phi(z_I(k), z_A(l)) \quad (2)$$

Here, $\Psi(u_I, u_A) = \{\mathbf{\Omega} \in \mathbb{R}_+^{M \times N} | \mathbf{\Omega} \mathbf{1}_N = u_I, \mathbf{\Omega}^\top \mathbf{1}_M = u_A\}$, $\phi(z_I(k), z_A(l))$ is the function computing the cosine distance between the cross-modal embedding pair, and $\mathbf{\Omega}$ is the transport plan, imitating the amount of *mass* shifted from the distribution θ_I to the distribution θ_A. An exact solution to the above expression leads to a sparse representation of the transport plan $\mathbf{\Omega}$ which at most $(2 \cdot \max(M, N) - 1)$ non-zero elements, ensuing an explainable and robust cross-modal alignment. We defer additional details to the appendix.

Audio-Visual Attention Consistency Enforcement Module (AVACE). Cross-modal interaction is essential for aligning the audio and visual modalities. Moreover, region-level supervision can encourage efficient localization. Inspired by the success of recent methods [16,18,62], we employ an adapter-based cross-attention strategy for efficient sound source localization. The modality-specific features in AVOpT lack awareness [28] of information from alternative modalities which can be infused through cross-modal attention. Therefore, to enable the audio-visual cross-modal reciprocity, we propose the AVACE module.

Although in a multi-modal context, feature fusion through a cross-attention scheme is effective in attending to relevant objects in the image, inconsistencies may arise such as attended regions being dispersed throughout the image including background objects. The reasons can be attributed to the quality of interplay between the feature embeddings. Considering CLAP audio encoder pre-trained with examples such as *'a man playing the violin'* (refer Fig. 2) paired with audio of a violin, the cross-modal knowledge of audio representations encourages it to focus on both the man and the violin in the image. Therefore, to ensure superior region-level alignment we confine the cross-modality attention map ($\mathcal{A}^c$) within the boundaries of the object of interest, denoted by the ground-truth bounding box. Considering a bounding box represented as $[x_{\text{Left}}, y_{\text{Top}}, x_{\text{Right}}, y_{\text{Bottom}}]$, we define a mask $\mathcal{M}$ such that $\mathcal{M}(y_{\text{Top}} : y_{\text{Bottom}}, x_{\text{Left}} : x_{\text{Right}}) = 1$, otherwise 0. Our goal is to maximize the attention within this bounding box and minimize it elsewhere. Therefore, we mathematically formulate the attention consistency objective $\mathcal{L}_{\text{AC}}$ as follows:

$$\mathcal{L}_{\text{AC}} = \lambda_1 \left(1 - \frac{\sum_{i,j} \mathcal{M}(i,j) \mathcal{A}^c(i,j)}{\sum_{i,j} \mathcal{M}(i,j) + \epsilon_1}\right) + \lambda_2 \left(\frac{\sum_{i,j} (1 - \mathcal{M}(i,j)) \mathcal{A}^c(i,j)}{\sum_{i,j} (1 - \mathcal{M}(i,j)) + \epsilon_2}\right) \quad (3)$$

Here, $\mathcal{A}^c$ denotes the audio-visual cross-modality attention, (i, j) represents the pixel location, λ_1, λ_2 are the loss hyper-parameters (we keep $\lambda_1 = \lambda_2 = 0.5$), and

Algorithm 1 MEERKAT: Training

Input: Image: I; Audio: A; Textual Instruction: T; Pre-trained LLM: $\mathcal{E}^{\text{LLM}}(\cdot)$; LLM Tokenizer: $\tau^{\text{LLM}}(\cdot)$; Pre-trained Image Encoder: $\mathcal{E}^{I}(\cdot)$; Pre-trained Audio Encoder: $\mathcal{E}^{A}(\cdot)$; AVACE Module: AVACE$(\cdot,\cdot)$; Masks from GT Bounding-Boxes: $\mathcal{M}$; Loss Hyperparameters: $\lambda_{\text{OT}}, \lambda_{\text{AC}}$; GT Tokens: ϕ_{GT}.

Output: Fine-tuned LLM: $\mathcal{E}^{T}(\cdot)$; Trained AVACE Module: AVACE$(\cdot,\cdot)$; Predicted Tokens: ϕ_{pred}.

$z_I \leftarrow \mathcal{E}^{I}(I); z_A \leftarrow \mathcal{E}^{A}(A)$ ▷ *Obtain Visual and Audio Embeddings.*

$z_T \leftarrow \tau^{\text{LLM}}(T)$ ▷ *Tokenize and Obtain Textual Encodings.*

$\tilde{z}_I, \tilde{z}_A, \mathcal{A}^c \leftarrow \text{AVACE}(z_I, z_A)$ ▷ *Obtain Audio-Visual Projections, Cross-Attn Map.*

$z_{AVT} \leftarrow (\tilde{z}_I \parallel \tilde{z}_A \parallel z_T)$ ▷ *Concatenate Embeddings.*

$\phi_{\text{pred}} \leftarrow \mathcal{E}^{\text{LLM}}(z_{AVT})$ ▷ *LLM Output.*

$\mathcal{L}_{\text{MEERKAT}} \leftarrow \mathcal{L}_{\text{CE}}(\phi_{\text{pred}}, \phi_{\text{GT}}) + \lambda_{\text{OT}} \cdot \mathcal{L}_{\text{OT}}(z_I, z_A) + \lambda_{\text{AC}} \cdot \mathcal{L}_{\text{AC}}(\mathcal{A}^c, \mathcal{M})$

Optimize model parameters to reduce $\mathcal{L}_{\text{MEERKAT}}$ until convergence.

return $\mathcal{E}^{T}(\cdot)$, AVACE$(\cdot,\cdot)$, ϕ_{pred}.

ϵ_1, ϵ_2 are the stability factors respectively. In Sec. 5.3, we demonstrate that $\mathcal{L}_{\text{AC}}$ encourages efficient localization and audio-visual alignment of the cross-attention maps, eventually leading to improved fine-grained cross-modal representations for downstream tasks.

3.3 Overall training objective

Our overall training objective comprises a combination of three sub-objectives: cross-entropy loss ($\mathcal{L}_{\text{CE}}$), weak AV alignment loss ($\mathcal{L}_{\text{OT}}$), and attention consistency loss ($\mathcal{L}_{\text{AC}}$). These losses are added together to obtain the final training loss for MEERKAT given as:

$$\mathcal{L}_{\text{MEERKAT}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{OT}} \cdot \mathcal{L}_{\text{OT}} + \lambda_{\text{AC}} \cdot \mathcal{L}_{\text{AC}} \tag{4}$$

Here, λ_{OT} and λ_{AC} are the loss weighting factors. We provide Algorithm 1 outlining the overall training procedure.

3.4 Numerical Representation of Box Location and Time Segment

Representation of Box Location. We embed the location of bounding boxes with numerical values in the natural language sequence. A box is represented intuitively by its top-left and bottom-right corners, i.e., [x_{Left}, y_{Top}, x_{Right}, y_{Bottom}]. Notably, these values are normalized whose factors are determined by the size of the respective image to which the bbox belongs. These coordinates may appear in either the input or the output sequences depending on the task. For instance, in *Audio Referred Image Grounding* task, MEERKAT predicts the bounding box of the object of interest, whereas, for *Audio-Visual Fact-checking* task, the text input to MEERKAT might contain the box coordinates.

Representation of Time Segment. We embed the time interval information using numerical figures in the natural language expression. A time segment is

intuitively represented by its start and end times, i.e., [`tStart`, `tEnd`], designating the onset of an event or an activity. Similar to boxes, these representations may appear in either the input or the output sequences depending on the task. For instance, in *Image Guided Audio Temporal Localization* task, the model predicts the time interval within which the query might have occurred, while for *Audio-Visual Fact-checking*, the input sequence might contain a reference time window. We add more details on the instruction preparation formats in the appendix.

4 MeerkatBench: A Unified Benchmark Suite for Fine-grained Audio-Visual Understanding

4.1 Task Overview

One of our primary contributions is to introduce a novel audio-visual fine-grained task unification benchmark. To this end, we present MeerkatBench comprising three fine-grained tasks: *(i)* audio referred image grounding, *(ii)* image guided audio temporal localization, *(iii)* audio-visual fact-checking, and two coarse-grained tasks: *(iv)* audio-visual question answering, *(v)* audio-visual captioning.

4.2 AVFIT-3M: Audio Visual Finegrained Instruction Tuning Dataset

In this section, we present AVFIT, an AV instruction tuning dataset comprising 3M multi-modal dialogues for model training. AVFIT consists of samples collected in the following ways: *(i)* suitable adaptation of public datasets and *(ii)* instruction-tuning data generation via prompting GPT-3.5 [3]. Next, we discuss the data curation procedure:

Adaptation of Public Datasets. Depending on the task and availability of datasets, we either collect the image-audio pairs directly from the publicly available datasets (VGG-SS [6], AVSBench [87], Flickr-SoundNet [61], LLP [69], AVQA [77], MUSIC-AVQA [32], VALOR [11]) or follow a semi-automated strategy to prepare the pairs by forming matching image-audio pairs from large-scale datasets having visual grounding annotation such as Openimages [29], PASCAL [20] and audio event datasets like AudioSet/AudioSet Strong [22], VGG-Sound [7]. We retain the original category labels (*'Existential', 'Temporal', etc.*) from the MUSIC-AVQA. To get similar insights in the AVQA dataset, we categorise every sample into one of the *'Existential', 'Temporal', 'Localisation', 'Count'* and *'World Knowledge'* categories. During the direct collection of pairs, we augment the audio snippet with a carefully chosen representative frame from the associated video. On the other hand, while forming pairs ourselves, we refer to a lookup table which we prepare beforehand by matching the corresponding class labels from the image and the audio datasets (more details in the appendix). We associate each image sample with its counterpart from the audio dataset. Finally, we supplement the image-audio pairs with the generated instructions as explained next. Details on the task-wise dataset details can be found in Tab. 2.

GPT-Assisted Instruction Generation. Instruction tuning datasets [26,38, 41, 55] have primarily focused on coarse-grained details like global image descriptions in the form of captioning or question answering without explicitly capturing fine-grained details. In this work, we aim to bridge this gap by introducing AVFIT that promotes region-level and time-sensitive understanding in the following ways: *(i)* AVFIT includes spatial coordinates of objects of interest (bounding box) along with corresponding audio snippets which leverage the synergy between audio-visual data. *(ii)* The designed dialogues audio time intervals either in input or output or both. *(iii)* To generate high-quality instructions we manually write a few example descriptions of each task and resort to GPT-3.5 [3] to create different variations. For further refinement of the generated dialogues we re-prompt GPT-4 [1] to ensure quality by reducing its context size. During training, we randomly pick one instruction for each sample. Fig. 2 illustrates a sample instruction from MeerkatBench. We use special tokens `<image>`, `<audio>`, `<obj>` which we later replace with instruction-guided image, audio and object categories respectively to generate prefix-based prompting.

5 Experiments and Results

5.1 Baselines

To the best of our knowledge, Meerkat is the first MLLM that unifies audio-visual spatial and temporal grounding, alongside possessing strong reasoning capabilities. We carefully choose the closest baseline for each task and suitably adapt them for fair comparisons. Owing to BuboGPT's [86] spatial localization ability, we select it as our baseline for the audio referred image grounding task. Most similar in spirit to our image guided audio-temporal localization task is TimeChat [59]. It leverages the pre-trained VideoLlama model and suitably instruction-tune it to tackle temporal grounding tasks. Due to their audio-visual comprehension abilities, we resort to X-InstructBLIP [50], Macaw-LLM [43], PandaGPT [65], and VideoLlama [82] as baselines for audio-visual fact-checking, AV question answering, and AV captioning tasks respectively. Please refer to Tab. 1 for an overview of the characteristics of the generalist baselines. For specialist baselines, refer to the corresponding task tables. We finetune all baselines on our datasets except for using Openimages-AudioSet and Openimages-VGGSound train splits from the audio-referred visual grounding task.

5.2 Main Results

Audio Referred Image Grounding (ARIG) This task involves visual grounding by predicting the coordinates of a bounding box around the object of interest guided by the input audio. We prepare 1.2M image-audio-instruction pairs using steps explained in Sec. 4.2. We add details of the input instruction format and model output in the appendix. Meerkat achieves superior performance in sounding object localization task, setting a new benchmark as shown in Tab. 3.

Models	Generalist?	VGG-SS cIoU ↑	VGG-SS AUC ↑	Flickr-SoundNet cIoU ↑	Flickr-SoundNet AUC ↑	PascalSound cIoU ↑	PascalSound AUC ↑	AVSBench cIoU ↑	AVSBench AUC ↑
SSPL [64]	✗	33.90	38.00	76.70	60.50	51.72	39.79	61.32	48.44
EZ-VSL [46]	✗	38.85	39.54	83.94	63.60	51.90	40.25	60.06	49.64
SSL-TIE [39]	✗	38.63	39.65	79.50	61.20	52.14	40.44	62.88	51.28
SLAVC [45]	✗	39.80	–	86.00	–	52.29	42.19	63.39	51.07
MarginNCE [52]	✗	39.78	40.01	85.14	64.55	53.61	45.52	65.85	52.92
HearTheFlow [21]	✗	39.40	40.00	84.80	64.00	55.48	47.40	67.49	54.39
FNAC [66]	✗	41.85	40.80	85.14	64.30	57.38	48.03	68.78	56.19
Alignment [62]	✗	42.64	41.48	82.40	64.60	58.34	49.86	71.57	57.52
BuboGPT [86]	✓	40.31	39.68	81.17	62.29	58.52	51.63	74.33	59.49
MEERKAT (ours)	✓	**48.51**	**45.62**	**88.35**	**67.88**	**65.23**	**56.10**	**79.82**	**65.35**
$\Delta_{\text{MEERKAT}-\text{BuboGPT}}$	✓	+20.34%	+14.97%	+8.85%	+8.97%	+11.47%	+8.66%	+7.39%	+9.85%

Table 3: Audio referred image grounding results. For AVSBench we follow the same train/test splits for all methods. We use the VGG-SS, Flickr-SoundNet, and PascalSound datasets only for evaluation.

Models	Generalist?	LLP F1-score ↑	AudioSet Strong F1-score ↑
AVE [70]	✗	35.47	37.42
AVSDN [36]	✗	37.15	41.48
AVVP [69]	✗	48.93	49.20
TimeChat [59]	✓	51.28	54.66
MEERKAT (ours)	✓	**54.96**	**56.85**
$\Delta_{\text{MEERKAT}-\text{TimeChat}}$	✓	+7.18%	+4.01%

Table 4: Image guided audio temporal localization results. We report the segment level F1-scores and attribute our performance gain over specialist models to our multi-task learning strategy.

Model	Type 1 F1-score ↑	Type 2 F1-score ↑	Type 3 F1-score ↑	Type 4 F1-score ↑
Macaw-LLM [43]	0.65	0.70	0.56	0.77
PandaGPT [65]	0.67	0.70	0.66	0.70
VideoLlama [82]	0.71	0.72	0.72	0.78
BuboGPT [86]	0.72	0.66	0.67	0.70
X-InstructBLIP [50]	0.73	0.72	0.72	0.80
TimeChat [59]	0.74	0.76	0.74	0.82
MEERKAT (ours)	**0.85**	**0.83**	**0.84**	**0.88**
$\Delta_{\text{MEERKAT}-\text{TimeChat}}$	+14.86%	+9.21%	+13.51%	+7.32%

Table 5: Audio-Visual fact-checking requires powerful reasoning capabilities across audio-visual modalities.

Image Guided Audio Temporal Localization (IGATL). When prompted to indicate a time interval within which a certain audio event occurs, MEERKAT is capable of producing accurate time bounds in the form `[tStart, tEnd]`, where `tStart` and `tEnd` are the start and end times, respectively. For all our experiments, we maintain the audio duration to be 30s. Different from prior visual grounding-based approaches [8, 56, 79], we present a new audio event localization task by setting a new baseline. We attribute the superior performance of our method on fine-grained audio temporal localization task to our specially designed AVOpT and AVACE modules, which ensure superior modality-specific guidance. Fig. 3 demonstrates our model can locate a precise time interval associated with an audio event. Tab. 4 reports the quantitative comparison of our method against other baselines.

Audio-Visual Fact-checking (AVFact). In this section we introduce a new suite of tasks that involves a strong comprehension of the audio-visual semantic information. These tasks broadly require the model to analyze and verify whether a given statement about an audio-visual scenario holds or not. Although we do not use GT spatio-temporal annotations to train the model, we classify this task under the fine-grained category as the task requires the model to attend to a specific region/time interval as passed in the query. To alleviate inconsistencies

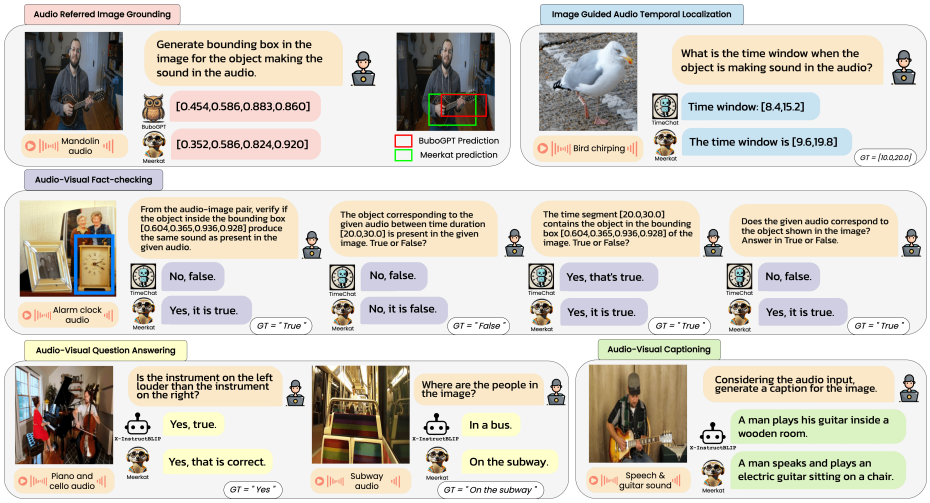

Fig. 3: Qualitative results. We compare our method against its closest baselines on all downstream tasks. MEERKAT aided by our novel design approach and instruction tuning datasets achieves superior performance on spatio-temporal grounding as well as coarse-grained tasks by outperforming prior approaches.

in evaluation, we restrict the model's response to binary *True/False* only. We divide these tasks into the following 4 categories:
Type 1: Given an audio-image pair, verify if the object within the bounding box produces sound that corresponds to the input audio.
Type 2: Given an audio-image pair, verify if the object in the image is related to the audio present within the given time segment.
Type 3: Given an audio-image pair, verify if the object present within the provided bounding box produces sound that corresponds to the audio within a given time segment.
Type 4: Given an audio snippet, verify whether its visual counterpart is present in the image or not.
In Tab. 5 we contrast the performance of other baselines against MEERKAT on all four types of AVFact tasks.
Audio-Visual Question Answering (AVQA). Audio-visual question answering aims to answer questions encompassing both audio and visual modalities. We collect question-answer pairs from the AVQA [77] and MusicAVQA [32] datasets and augment them with instruction tuning templates (details in appendix) to prepare the data samples. We contrast our method against SoTA generalist and specialist models on the AVQA task in Tab. 6. We report the evaluation results on the other metrics like Count and Comp in the appendix.
Audio-Visual Captioning (AVC). This task learns how to generate text tokens conditioned on audio-visual inputs. In contrast to image/audio-only captioning methods, this requires strong multi-modal understanding and reasoning capabilities. We note that MEERKAT outperforms existing specialist and gen-

Model	Generalist?	AVQA			MUSIC AVQA			VALOR-32K			
		Exist ↑	Localis ↑	Temp ↑	Exist ↑	Localis ↑	Temp ↑	BLEU@4 ↑	METEOR ↑	ROUGE ↑	CIDEr ↑
AVSD [60]	✗	81.61	58.79	61.41	-	-	-	-	-	-	-
PanoAVQA [80]	✗	81.21	59.33	63.23	-	-	-	-	-	-	-
ST-AVQA [32]	✗	81.81	64.51	63.23	-	-	-	-	-	-	-
CAD [47]	✗	83.42	73.97	76.16	-	-	-	-	-	-	-
AVST [32]	✗	-	-	-	72.44	65.54	59.36	-	-	-	-
LAVISH [37]	✗	-	-	-	73.83	65.00	60.81	-	-	-	-
LAST [40]	✗	-	-	-	76.21	68.91	60.60	-	-	-	-
SMPFF [12]	✗	-	-	-	-	-	-	7.59	12.64	28.69	37.18
VALOR [11]	✗	-	-	-	-	-	-	8.97	14.88	30.86	55.73
Macaw-LLM [43]	✓	82.19	74.86	78.98	72.99	71.28	59.36	9.36	15.28	33.31	58.98
PandaGPT [65]	✓	83.38	76.81	79.11	78.48	73.12	65.85	10.35	16.92	34.88	61.22
VideoLlama [82]	✓	84.48	77.06	81.36	81.21	76.10	67.52	11.45	17.39	35.14	63.63
X-InstructBLIP [50]	✓	85.53	80.09	83.91	80.28	77.45	68.83	12.31	18.82	37.93	65.73
MEERKAT (ours)	✓	**88.24**	**86.65**	**86.55**	**83.62**	**80.51**	**73.33**	**16.88**	**23.18**	**45.67**	**76.84**
$\Delta_{\text{MEERKAT}-\text{X-InstructBLIP}}$	✓	+3.17%	+8.19%	+3.15%	+4.16%	+3.95%	+6.54%	+37.12%	+23.17%	+20.41%	+16.9%

Table 6: Quantitative results on AVQA and AV captioning tasks. The reported numbers on AVQA dataset [77] are on the val split. For the MUSIC-AVQA dataset [32], results are reported on the balanced test set. Here, Exist: Existential, Localis: Localisation, Temp: Temporal. Evaluation for AV captioning is done on VALOR-32K [11] val set. MEERKAT demonstrates strong coarse-grained understanding abilities.

eralist models by a considerable margin and sets a new baseline on a recent benchmark dataset VALOR [11], as shown in Tab. 6.

We argue that the seamless extension of MEERKAT to coarse-grained tasks is facilitated by the strong semantic understanding acquired by our model during training. This comprehension ability enables our model to effectively navigate and interpret the complexities inherent in coarse-grained tasks, showcasing the versatility and easy extensibility of our approach.

5.3 Ablation Study

Evaluation on Pre-training Tasks. To study the effect of *unified* pre-training, we evaluate our model under single task vs. multi-task learning setting. We gradually add datasets for each task and assess the model's performance. On quantitative evaluation, we note that our multi-task setting is indeed benefiting from each other in achieving superior performance as shown in Tab. 7. While the model trained on fine-grained tasks performs significantly well on the coarse-grained tasks, introducing the coarse-grained tasks in the training set doesn't have a considerable impact on ARIG, IGATL, and AVFact - underlining the importance of our collected fine-grained datasets.

Full vs. LoRA Finetuning We conduct experiments on different modes of LLM fine-tuning. As shown in Fig. 4, LoRA [27] based fine-tuning with r=32 achieves optimal performance. Lower values of r (4,16) performs poorly compared to 32 and we empirically find full-finetuning performs slightly worse than LoRA (r=32). We add more ablation results in the appendix.

5.4 Qualitative Analysis

Fig. 3 illustrates the comparison of MEERKAT with its closest baseline on all downstream tasks. We observe that our model powered by the combination of

Pre-training Task					VGG-SS	LLP	AVFact	AVQA	VALOR
ARIG	IGATL	AVFC	AVQA	AVC	cIOU ↑	F1-score ↑	Avg F1-score ↑	Avg Acc. ↑	CIDEr ↑
✓	✗	✗	✗	✗	47.53	18.73	0.71	77.22	67.82
✓	✓	✗	✗	✗	47.75	54.26	0.74	79.74	70.19
✓	✓	✓	✗	✗	48.17	54.65	0.83	81.11	72.13
✓	✓	✓	✓	✗	48.29	54.82	0.83	86.68	74.14
✓	✓	✓	✓	✓	**48.51**	**54.96**	**0.85**	**87.14**	**76.84**

Table 7: We systematically analyze the effect of multi-task learning. Here ARIG: audio referred image grounding, IGATL: image guided audio temporal localization, AVFC: audio-visual fact-checking, AVQA: audio-visual question answering, and AVC: audio-visual captioning. AVQA avg accuracy calculated over Exist, Localis, and Temp.

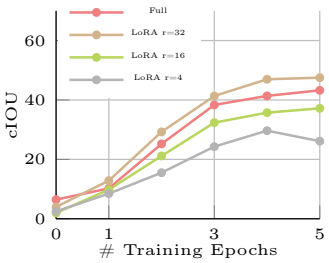

Fig. 4: cIoU upper bound on VGG-SS for Full vs. LoRA based finetuning.

AVOpT and AVACE is equipped with finer region-level understanding compared to Bubo-GPT [86]. Similarly, on image-guided audio temporal localization, our method outperforms TimeChat [59]. We attribute the excellent performance of Meerkat to the strong AV association learning backed by the instruction tuning data and multi-task learning set-up. For the AVQA task, the recently proposed X-InstructBLIP [50] achieves comparable results. We argue that fuelled by a strong fine-grained understanding acquired through the pre-training stages, Meerkat can extract additional contextual information from the visual modality. Our training paradigm emphasizes on both audio and visual modalities facilitating precise audio understanding by the model when compared against Video-LLaMA [82]. Finally, on the AVFact tasks, our approach achieves superior performance due to its better multi-modal comprehension skills.

6 Conclusions and Future Works

We presented Meerkat, a powerful multi-modal LLM adept at processing audio-visual inputs to comprehend fine-grained spatio-temporal information. Our novel audio-visual alignment strategy powered by the AVOpT and AVACE modules instill strong compositional understanding into Meerkat, thereby making it suitable for various challenging tasks. To pave the way for future research in this direction, we collect AVFIT comprising 3M instruction tuning samples and introduce MeerkatBench. Extensive experiments demonstrate the effectiveness of our approach on a wide range of downstream tasks, consistently achieving state-of-the-art performance.

In future work, we plan to equip our model to address more challenging tasks like LLM guided AV segmentation. We also plan to extend the model's capability to operate on videos and handle associated tasks such as video temporal grounding, and video summarization. Future work can also focus on collecting video-centric multi-modal training data and reasoning benchmarks for evaluation at scale. Finally, our work opens up avenues to study robustness and compositional understanding of AV LLMs with fine-grained comprehension abilities.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 10
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 2, 4
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 1, 9, 10
4. Chen, F., Han, M., Zhao, H., Zhang, Q., Shi, J., Xu, S., Xu, B.: X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv preprint arXiv:2305.04160 (2023) 2, 4
5. Chen, G., et al: Plot: Prompt learning with optimal transport for vision-language models. ICLR (2023) 7
6. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16867–16876 (2021) 9
7. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020) 9
8. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023) 2, 4, 6, 11
9. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023) 2, 4
10. Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., Liu, J.: Graph optimal transport for cross-domain alignment. In: International Conference on Machine Learning. pp. 1542–1553. PMLR (2020) 6
11. Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J., Liu, J.: Valor: Vision-audio-language omni-perception pretraining model and dataset. arXiv preprint arXiv:2304.08345 (2023) 9, 13
12. Chen, S., Zhu, X., Hao, D., Liu, W., Liu, J., Zhao, Z., Guo, L., Liu, J.: Mm21 pre-training for video understanding challenge: Video captioning with pretraining techniques. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4853–4857 (2021) 13
13. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020) 6
14. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/ 1, 4
15. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling lan-

guage modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023) 1, 4
16. Chowdhury, S., Nag, S., Manocha, D.: Apollo: Unified adapter and prompt learning for vision language models. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023) 7
17. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022) 2
18. Dou, Z.Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., et al.: Coarse-to-fine vision-language pre-training with fusion in the backbone. Advances in neural information processing systems **35**, 32942–32956 (2022) 2, 6, 7
19. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: Clap learning audio concepts from natural language supervision. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 5
20. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**, 98–136 (2015) 9
21. Fedorishin, D., Mohan, D.D., Jawade, B., Setlur, S., Govindaraju, V.: Hear the flow: Optical flow-based self-supervised visual sound source localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2278–2287 (2023) 11
22. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017) 9
23. Georgescu, M.I., Fonseca, E., Ionescu, R.T., Lucic, M., Schmid, C., Arnab, A.: Audiovisual masked autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16144–16154 (2023) 6
24. Gong, Y., Luo, H., Liu, A.H., Karlinsky, L., Glass, J.: Listen, think, and understand. arXiv preprint arXiv:2305.10790 (2023) 2
25. Gutmann, M.U., Hyvärinen, A.: Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. Journal of machine learning research **13**(2) (2012) 6
26. Honovich, O., Scialom, T., Levy, O., Schick, T.: Unnatural instructions: Tuning language models with (almost) no human labor. arXiv preprint arXiv:2212.09689 (2022) 10
27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 13
28. Huang, S., Qin, L., Wang, B., Tu, G., Xu, R.: Sdif-da: A shallow-to-deep interaction framework with data augmentation for multi-modal intent detection. arXiv preprint arXiv:2401.00424 (2023) 7
29. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., 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) 9
30. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023) 2, 4

31. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023) 2, 4
32. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19108–19118 (2022) 9, 12, 13
33. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021) 6
34. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023) 4
35. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022) 2, 6
36. Lin, Y.B., Li, Y.J., Wang, Y.C.F.: Dual-modality seq2seq network for audio-visual event localization. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2002–2006. IEEE (2019) 11
37. Lin, Y.B., Sung, Y.L., Lei, J., Bansal, M., Bertasius, G.: Vision transformers are parameter-efficient audio-visual learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2299–2309 (2023) 13
38. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 2, 4, 10
39. Liu, J., Ju, C., Xie, W., Zhang, Y.: Exploiting transformation invariance and equivariance for self-supervised sound localisation. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3742–3753 (2022) 11
40. Liu, X., Dong, Z., Zhang, P.: Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4478–4487 (2024) 13
41. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022) 10
42. Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M., Lu, P., Wang, T., Wei, Z.: Valley: Video assistant with large language model enhanced ability. arXiv preprint arXiv:2306.07207 (2023) 2
43. Lyu, C., Wu, M., Wang, L., Huang, X., Liu, B., Du, Z., Shi, S., Tu, Z.: Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. arXiv preprint arXiv:2306.09093 (2023) 2, 3, 10, 11, 13
44. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023) 2, 4
45. Mo, S., Morgado, P.: A closer look at weakly-supervised audio-visual source localization. Advances in Neural Information Processing Systems **35**, 37524–37536 (2022) 11
46. Mo, S., Morgado, P.: Localizing visual sounds the easy way. In: European Conference on Computer Vision. pp. 218–234. Springer (2022) 11
47. Nadeem, A., Hilton, A., Dawes, R., Thomas, G., Mustafa, A.: Cad-contextual multi-modal alignment for dynamic avqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7251–7263 (2024) 13

48. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 6
49. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022) 2, 4
50. Panagopoulou, A., Xue, L., Yu, N., Li, J., Li, D., Joty, S., Xu, R., Savarese, S., Xiong, C., Niebles, J.C.: X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. arXiv preprint arXiv:2311.18799 (2023) 2, 3, 10, 11, 13, 14
51. Park, J., Lee, J., Sohn, K.: Bridging vision and language spaces with assignment prediction. arXiv preprint arXiv:2404.09632 (2024) 6
52. Park, S., Senocak, A., Chung, J.S.: Marginnce: Robust sound localization with a negative margin. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 11
53. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277 (2023) 2
54. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023) 4
55. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015) 10
56. Pramanick, S., Han, G., Hou, R., Nag, S., Lim, S.N., Ballas, N., Wang, Q., Chellappa, R., Almahairi, A.: Jack of all tasks, master of many: Designing general-purpose coarse-to-fine vision-language model. arXiv preprint arXiv:2312.12423 (2023) 2, 11
57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 5
58. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020) 1
59. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. arXiv preprint arXiv:2312.02051 (2023) 3, 10, 11, 14
60. Schwartz, I., Schwing, A.G., Hazan, T.: A simple baseline for audio-visual scene-aware dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12548–12558 (2019) 13
61. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4358–4366 (2018) 9
62. Senocak, A., Ryu, H., Kim, J., Oh, T.H., Pfister, H., Chung, J.S.: Sound source localization is all about cross-modal alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7777–7787 (2023) 7, 11
63. Shu, F., Zhang, L., Jiang, H., Xie, C.: Audio-visual llm for video understanding. arXiv preprint arXiv:2312.06720 (2023) 2, 3

64. Song, Z., Wang, Y., Fan, J., Tan, T., Zhang, Z.: Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3222–3231 (2022) 11
65. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355 (2023) 2, 3, 4, 10, 11, 13
66. Sun, W., Zhang, J., Wang, J., Liu, Z., Zhong, Y., Feng, T., Guo, Y., Zhang, Y., Barnes, N.: Learning audio-visual source localization via false negative aware contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6420–6429 (2023) 11
67. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023) 4
68. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., Stojnic, R.: Galactica: A large language model for science. arXiv preprint arXiv:2211.09085 (2022) 4
69. Tian, Y., Li, D., Xu, C.: Unified multisensory perception: Weakly-supervised audio-visual video parsing. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 436–454. Springer (2020) 9, 11
70. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European conference on computer vision (ECCV). pp. 247–263 (2018) 11
71. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 1, 4, 6
72. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv. org/abs/2307.09288 (2023) 4
73. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023) 2, 4
74. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems **36** (2024) 2, 4
75. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021) 4
76. Workshop, B., Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022) 4
77. Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: Avqa: A dataset for audio-visual question answering on videos. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3480–3491 (2022) 9, 12, 13
78. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023) 4

79. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023) 2, 4, 11
80. Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-avqa: Grounded audio-visual question answering on 360deg videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2031–2041 (2021) 13
81. Zhang, C., Cai, Y., Lin, G., Shen, C.: Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12203–12213 (2020) 6
82. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023) 2, 3, 10, 11, 13, 14
83. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023) 2, 4
84. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023) 4
85. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022) 4
86. Zhao, Y., Lin, Z., Zhou, D., Huang, Z., Feng, J., Kang, B.: Bubogpt: Enabling visual grounding in multi-modal llms. arXiv preprint arXiv:2307.08581 (2023) 2, 3, 4, 10, 11, 14
87. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio–visual segmentation. In: European Conference on Computer Vision. pp. 386–403. Springer (2022) 9
88. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 2, 4