Crab: A Unified Audio-Visual Scene Understanding Model with Explicit Cooperation

Henghui Du[1,3,*], **Guangyao Li**[2], **Chang Zhou**[3], **Chunjie Zhang**[3], **Alan Zhao**[3], **Di Hu**[1,✉]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
[2] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3] AI Technology Center, Online Video Business Unit, Tencent PCG
[1]{cserdu,dihu}@ruc.edu.cn, [2]guangyaoli@tsinghua.edu.cn
[3]{cserdu,chanzhou,Chunjie Zhang,Alan Zhao}@tencent.com

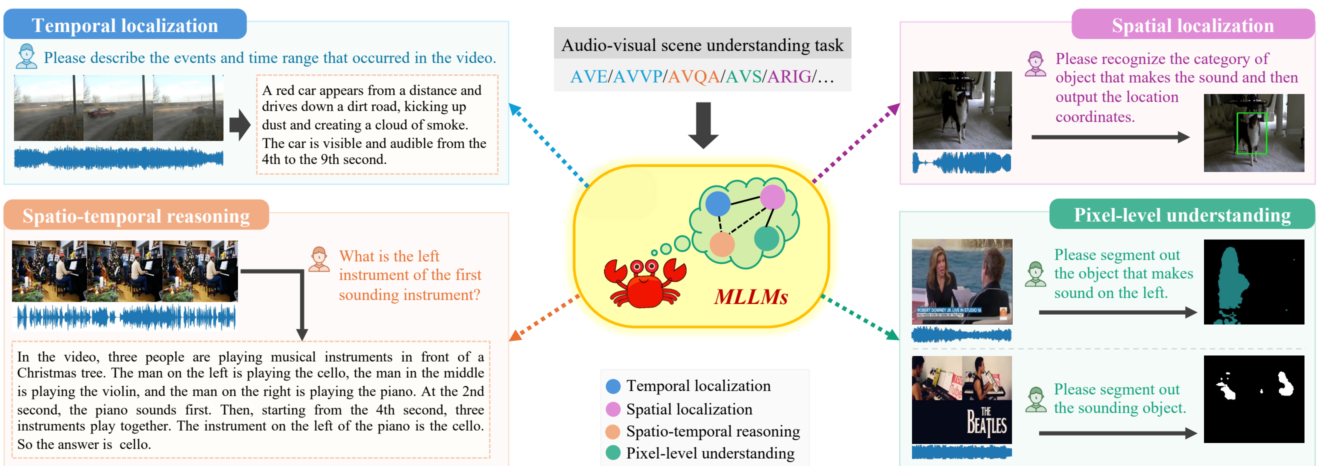


Figure 1. We present Crab, a unified audio-visual scene understanding model with explicit cooperation, which can complete various audio-visual tasks. It is trained on an instruction-tuning dataset with explicit reasoning process, which clarifies the cooperative relationship among tasks. Furthermore, to alleviate the interference caused by the learning process of complex audiovisual data and facilitate concrete cooperation, an interaction-aware LoRA structure is designed to enable the model focus on different aspects of data interaction.

Abstract

*In recent years, numerous tasks have been proposed to encourage model to develop specified capability in understanding audio-visual scene, primarily categorized into temporal localization, spatial localization, spatio-temporal reasoning, and pixel-level understanding. Instead, human possesses a unified understanding ability for diversified tasks. Therefore, designing an audio-visual model with general capability to unify these tasks is of great value. However, simply joint training for all tasks can lead to interference due to the heterogeneity of audiovisual data and complex relationship among tasks. We argue that this problem can be solved through explicit cooperation among tasks. To achieve this goal, we propose a unified learning method which achieves **explicit inter-task cooperation** from both the perspectives of data and model thoroughly. Specifically, considering the labels of existing datasets are simple words, we carefully refine these datasets and construct an **A**udio-**V**isual **U**nified **I**nstruction-tuning dataset with **E**xplicit reasoning process (**AV-UIE**), which clarifies the cooperative relationship among tasks. Subsequently, to facilitate concrete cooperation in learning stage, an interaction-aware LoRA structure with multiple LoRA heads is designed to learn different aspects of audiovisual data interaction. By unifying the explicit cooperation across the data and model aspect, our method not only surpasses existing unified audio-visual model on multiple tasks, but also outperforms most specialized models for certain tasks. Furthermore, we also visualize the process of explicit cooperation and surprisingly find that each LoRA head has certain audio-visual understanding ability. Code and dataset: https://github.com/GeWu-Lab/Crab*

[*]Intern at Tencent PCG, [✉]Corresponding author.

1. Introduction

"Alone we can do so little; together we can do so much."

— Helen Keller

Over the past few years, many works have explored the utilization of audio and visual modality to enhance general scene understanding, thereby deriving multiple tasks, primarily categorized into temporal localization, spatial localization, spatio-temporal reasoning, and pixel-level understanding. Moreover, several reviews [53] have also provided detailed analysis for these tasks from novel perspective. These tasks include audio-visual event localization (AVE) [47], audio-visual video parsing (AVVP) [48], audio referred image grounding (ARIG) [43], audio-visual question answering (AVQA) [26], audio-visual segmentation (AVS) [61, 62] and reference audio-visual segmentation (Ref-AVS) [52], *etc*. Specifically, AVE and AVVP tasks focus on temporal localization, where the model needs to accurately predict the events occurring in the video and their temporal boundaries. The purpose of ARIG task is to endow models with grounding capabilities, enabling them to locate the position of sounding objects. The AVQA task requires the model to integrate temporal and spatial information to answer questions. To complete AVS and Ref-AVS tasks, the model needs to output segmentation mask of target objects at pixel level. While significant advancements have been made in exploring these tasks, these works [13, 16, 17, 28, 29, 34, 45, 52, 57] have mainly focused on completing specific task. By contrast, humans can generally possess a general perception and understanding for complex audio-visual scene, hence we also aspire to equip the model with similar capability.

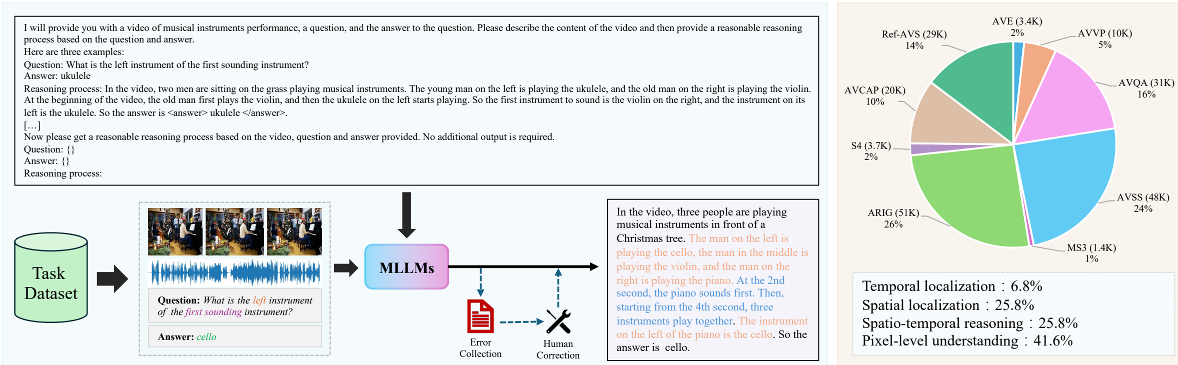
Unified general-purpose models can effectively transfer knowledge across tasks, and even perform unobserved tasks during training process. Early works primarily focused on unifying multiple language tasks, such as GPT2 [39] and PaLM [11]. Subsequently, large language models (LLMs) [10, 49, 58] have demonstrated human-level performance across various language tasks. Then researchers further extend LLMs into multimodal domain. A line of works [1, 4, 30] perform well on visual-language tasks. Recently, utilizing LLMs to complete multiple audio-visual tasks has become a goal pursued by the community [9, 12, 59]. They mainly accomplished this by naively curating instruction-tuning dataset and training jointly on multiple tasks, which is essentially an implicit approach to achieve inter-task cooperation. Considering the heterogeneity of audiovisual data and complex relationship among audio-visual tasks, it is difficult to more effectively achieve mutual cooperation among these tasks through such simple paradigm. For example, for two different types of tasks, temporal localization and spatial localization, instruction fine-tuning is effective when applied to one type of task individually, but directly training jointly on these two tasks could lead to interference.

We argue that this problem can be addressed through explicit cooperation among tasks from both data and model perspectives, thoroughly. On the one hand, considering labels of existing datasets are relatively simple words, they are insufficient for different tasks to establish relationship. Therefore, it is essential to augment existing datasets to explicitly clarify cooperative relationship among tasks. As shown in Fig. 1, for spatio-temporal reasoning task, the label is no longer single words, but includes explicit reasoning process, such as the time when the first instrument makes a sound, the position of instruments, *etc*. These information can help establish cooperation relationship among tasks. On the other hand, implicit methods may cause interference in the learning stage of heterogenous audiovisual data, hence the models needs to focus on different aspects of audio-visual data interaction, facilitating concrete cooperation in learning stage. As shown in Fig. 1, the model can focus on multiple interaction aspects such as temporal, spatial, and pixel-level, thereby facilitating concrete cooperation.

To achieve this goal, we propose a unified audio-visual scene understanding model with explicit cooperation. Specifically, we first refine existing multiple audio-visual task datasets and construct an **A**udio-**V**isual **U**nified **I**nstruction-tuning dataset with **E**xplicit reasoning process (**AV-UIE**), which clarifies concrete cooperative relationship among tasks. The learning objective is not only to predict simple labels, but also includes explicit reasoning process, which contains information from multiple spatio-temporal dimensions. The model can rely on these information to establish relationship among tasks. Then, to alleviate interference and facilitate concrete cooperation in learning stage, an interaction-aware LoRA structure with multiple LoRA heads is designed. Each head is responsible for learning certain audiovisual data interaction aspect. By unifying explicit cooperation across data and model aspect, our method not only surpasses existing unified audiovisual models on multiple tasks, but also outperforms the specialized models for certain tasks. Furthermore, we also visualize the process of explicit cooperation and surprisingly find that each LoRA head has certain audio-visual understanding ability.

Our contributions can be summarized as follows:

- Based on unification of audio-visual tasks, we propose an approach of explicit cooperation, which solves the problem of unified learning of complex audiovisual data.
- We construct AV-UIE, an audio-visual unified instruction-tuning dataset with explicit reasoning process, which clarifies the cooperation relationship among tasks.
- We design an interaction-aware LoRA structure to focus on different audiovisual data interaction aspects, thereby facilitating concrete cooperation in learning stage.



(a) The procedure of constructing dataset.

b) The data statistics for each task.

Figure 2. Our proposed AV-UIE dataset. (a) explains the specific process of dataset construction, and (b) is the data analysis for all tasks.

2. Related Works

Unified Learning for Multi-task. Pre-training language models on a large amount of unlabeled text corpora has become a standard practice, such as BERT [21]. To address certain task, the model needs to be fine-tuned on corresponding downstream task. Therefore, researchers have been interested in developing unified models to solve a wide range of language tasks [2, 11]. Inspired by this success, many efforts [7, 8, 35, 36, 55] have been devoted to building unified models that can apply to vision tasks. In this work, we also use a unified interface to accomplish multiple tasks, but we explore more complex audio-visual scene understanding tasks.

Audio-Visual Unified MLLMs. Leveraging powerful reasoning capability of LLMs to accomplish multiple audio-visual tasks has become a recent trend [9, 12, 44, 59]. VideoLLaMA 2 [9] integrated audio and video branch into model through joint training, thereby enriching the model's multimodal understanding capability. MEERKAT [12] builds a fine-grained large-scale audio-visual instruction-tuning dataset, endowing the model with grounding ability in time and space. While these works have made significant advancements, they simply construct instruction-tuning dataset and conduct joint training on audio-visual tasks. In contrast to these methods, we propose a unified learning method that achieves explicit cooperation.

Multi-task LoRA. LoRA [18] is an approach that utilizes limited computational resources to efficiently finetune large language models. Researchers further explore the combination of multiple LoRAs to handle complex tasks. LoRAhub [19] trains multiple LoRA adapters separately, and then selects appropriate combination of adapters during inference process to complete specific task. These works aim to use multiple LoRAs to address different downstream tasks. Unlike them, in a multi-task unified framework, to achieve the goal of explicit cooperation among audio-visual tasks, we design an interaction-aware LoRA structure to decouple the model's various capabilities.

3. AV-UIE Dataset

In order to clarify the explicit cooperative relationship and enable model to establish connections among tasks, we construct AV-UIE, an audio-visual unified instruction-tuning dataset with explicit reasoning process. In Section 3.1, we will introduce the construction process in detail, and in Section 3.2 we provide an analysis for our dataset.

3.1. Dataset Construction

Our dataset is primarily an augmentation of existing audio-visual task datasets. Considering the raw data labels are single words or phrases, we transform these plain labels into an explicit reasoning process. Figure 2(a) illustrates a detailed construction procedure. Specifically, we first sample an instance from the dataset of each task, which includes audio, video, and simple labels. Then, using the in-context learning method, we prompt existing powerful MLLMs to output correct reasoning process based on provided information. In this process, we ensure that the final label after transformation remains consistent with original label. Finally, to ensure the quality of data, we filter out poor quality instances and use manual annotation for correction. For instance, in the example described in Figure 2(a), this video shows a scene of three people playing musical instruments. The question is "What is the left instrument of the first sounding instrument?", and the corresponding answer is "cello". Through our dataset construction process, a detailed reasoning process can be obtained, which contains rich information in terms of temporal and spatial, conducive to temporal and spatial localization tasks.

Subsequently, following above construction procedure, we transform the MUSIC-AVQA [26], AVE [47] fully supervised, LLP [48], AVS-Bench [61] and Ref-AVS [52] datasets. Specifically, for MUSIC-AVQA and AVE datasets, we directly transform the labels as described in construction process. Since the training data of LLP dataset only contains video-level event labels and lacks the tem-

We use Gemini 1.5 Pro to complete this process.

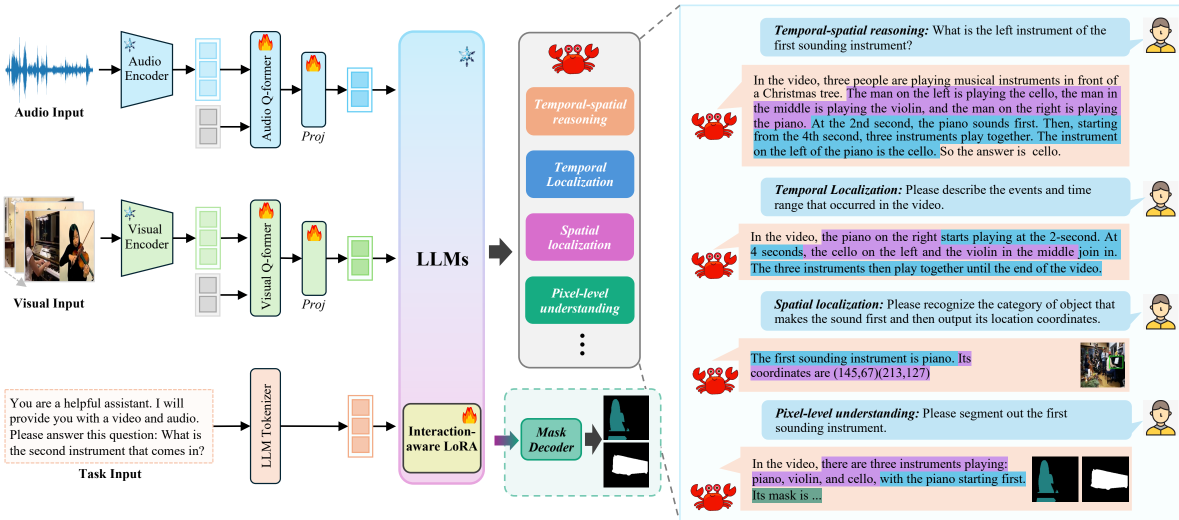


Figure 3. The architecture of our unified audio-visual scene understanding model. It mainly consists of two parts: unified audio-visual interface, which consists of three multimodal branches, and a large language model with interaction-aware LoRA structure. The audio branch and visual branch process audio and video inputs respectively, while the segmentation branch is responsible for outputting the segmentation mask. The model is trained on our AV-UIE dataset, which clarifies the cooperation relationship among tasks, as marked by different colors on the right side of the figure. Content of same color in different tasks can help model establish cooperative relationship among tasks. Furthermore, to alleviate the interference caused by the learning process of complex audiovisual data, we design an interaction-aware LoRA structure to facilitate concrete cooperation.

poral boundaries of event occurrence, We use the same MLLMs to annotate additional temporal information. The transformed LLP dataset can also benefit from data related to other temporal localization and spatio-temporal reasoning tasks, which contain precise temporal labels. For AVS-Bench and Ref-AVS datasets, we explicitly add information about the object making the sound. For spatial localization task, we collect data from the AVS-Bench dataset and obtain the bounding box of sounding object. Each box is represented by its top-left and bottom-right corners, i.e., $[x_{Left}, y_{Top}, x_{Right}, y_{Bottom}]$. To enhance the model's spatio-temporal ability, our dataset additionally includes the perception dataset proposed by VALOR [6]. To this end, we can obtain an instruction-tuning dataset with explicit reasoning process, namely AV-UIE dataset. We provide prompt and examples in the *supplementary materials*.

3.2. Dataset Analysis

Our dataset contains approximately $200K$ training samples with explicit reasoning process, as shown in Fig. 2(b). The data of pixel-level understanding tasks account for the largest proportion, reaching 41.6% (S4 1.7%, MS3 0.7%, AVSS 24.3% and Ref-AVS 14.7%), while spatial localization (ARIG 25.8%) and spatio-temporal reasoning tasks (MUSIC-AVQA 15.7% and AVCap 10.1%) account for 25.8% each. The number of temporal localization tasks is the smallest, accounting for 6.8% (AVE 1.7% and LLP 5.1%). Compared to previous instruction-tuning datasets [12, 31, 41], our dataset covers a wider range of task types, including temporal, spatial localization, spatio-temporal reasoning, and pixel-level understanding tasks. To the best of our knowledge, it is the first dataset to cover all these audio-visual scene understanding tasks. Moreover, while the total amount of our data is relatively smaller for each task, thanks to the explicit reasoning process, we can still achieve superior results with a smaller amount of data.

4. Method

In this section, we present a detailed introduction to our model. As depicted in Fig. 3, our framework comprises two parts: *i*) unified audio-visual interface, which consists of three multimodal branches, each responsible for processing audio, visual, and segmentation mask output, *ii*), a large language model with interaction-aware LoRA structure.

4.1. Unified Audio-Visual Interface

Visual branch. The visual branch comprises a visual encoder and a visual-language alignment module. We employ a pre-trained CLIP-ViT-L/14 [40] encoder $\mathcal{E}^I(\cdot)$ as our visual encoder. Given T frames of video input $\mathcal{V} = \{I_i \in \mathbb{R}^{H\times W\times C}\}_{i=1}^{T}$ where H, W, C represent the height, width and channels respectively, we extract the patch-level feature of each frame $f_v \in \mathbb{R}^{L_v \times D_v}$ where L_v and D_v denote the number of image tokens and feature dimension respectively. The visual embeddings of video can be denoted as $F_v = \{f_v^i\}_{i=1}^{T}$. Then we use a Q-Former [30] with K_v learnable query tokens to compress the visual features into a fixed number of tokens. These visual tokens will be aligned to the textual semantic space through a two-layer MLP.

Audio branch. It consists of a pre-trained audio encoder and an audio-language alignment module. The audio encoder transforms the raw audio into an audio embedding. We use BEATs [5] as audio encoder $\mathcal{E}^A(\cdot)$, which is trained to extract audio semantics information using iterative self-supervised learning. We select T audio segments corresponding to the visual frames. For each audio segment a_i, we can obtain corresponding audio feature $f_a^i \in \mathbb{R}^{L_a \times D_a}$ where L_a and D_a are token numbers and feature dimension of audio feature. The audio embeddings can be denoted as $F_a = \{f_a^i\}_{i=1}^T$. Similarly, our audio-language alignment module also comprises a Q-Former followed by a two-layer MLP. Finally, we can get K_a audio tokens aligned to the textual space where K_a is the number of learnable query tokens. These tokens will be fed into the LLM.

Segmentation branch. Inspired by the success of SAM [23] in image segmentation, recent works [24, 42] have combined the mask decoder in SAM with LLMs to complete segmentation task. Our segmentation branch mainly consists of a mask decoder $\mathcal{D}^M(\cdot)$ with a similar structure. To adapt the mask decoder to our model, we extend the original LLM vocabulary with multiple `<MASK>` tokens, i.e. $C_{seg} = \{\texttt{<MASK>}_i\}_{i=0}^N$ where N is the number of tokens, which signify the segmentation output. When the model intends to generate a segmentation mask, the output would include these `<MASK>` tokens. We then extract the LLM last-layer embeddings corresponding to these `<MASK>` tokens. These embeddings serve as a prompt and will be sent to the mask decoder together with the visual embeddings F_v. Finally the mask decoder predicts the segmentation mask $\mathcal{M} \in \mathbb{R}^{C \times H \times W}$ where C represents the number of categories, H and W represent height and width of mask image respectively.

4.2. Interaction-aware LoRA

LoRA [18] is a highly efficient method for finetuning LLMs, enabling them to acquire specific capabilities. In a standard LoRA structure, there are two trainable pairs of rank decomposition matrices $A \in \mathbb{R}^{h \times r}$ and $B \in \mathbb{R}^{r \times h}$ where h and r is the hidden dimension of LLMs and LoRA rank. Rather than employing a single LoRA, multiple LoRA structure employs multiple independent and smaller LoRA, with each LoRA responsible for addressing a specific downstream task. Both structures have drawbacks in unifying multiple tasks. The former relies on a fully parameter-shared LoRA to implicitly learn the capabilities needed for all tasks, making it difficult to achieve effective task cooperation. The latter, with independent LoRAs for each task, hinders inter-task cooperation.

Inspired by HydraLoRA [46], in order to enable the model focus on different audiovisual data interaction aspects, we design an interaction-aware LoRA structure, a specialized fine-tuning architecture for multi-task. Fig. 4 provides a detailed explanation of this structure. It mainly consists of a shared parameter matrix A and multiple independent matrices B, also known as LoRA heads. Different from HydraLoRA, the purpose of our interaction-aware LoRA structure is to alleviate interference of complex audiovisual data and facilitate concrete cooperation in learning stage. Under this structure, each head can focus on certain aspect of data interaction, with our expectations focusing on temporal, spatial, and pixel-level, *etc*. Therefore, the number of LoRA heads is not related to the task numbers, but is determined by different types of data interaction.

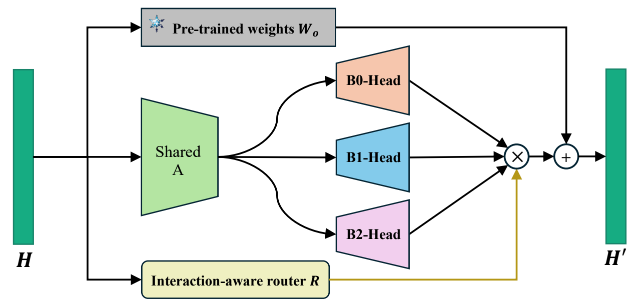


Figure 4. The architecture of interaction-aware LoRA structure. It comprises a shared matrix A and multiple matrices B, also known as LoRA head. Matrix A learns common multimodal representations, while each LoRA head is dedicated to learn certain audiovisual data interaction aspect.

Specifically, T frames of video input and corresponding audio segments first pass through the audio and visual branch to obtain corresponding representations, then they are concatenated and fed into the LLMs together with the task instruction as the initial hidden states H_0:

$$H_v = \text{MLP}_v(\text{Q-Former}_v(\mathcal{E}^I(F_v))) \in \mathbb{R}^{TK_v \times h}, \quad (1)$$

$$H_a = \text{MLP}_a(\text{Q-Former}_a(\mathcal{E}^A(F_a))) \in \mathbb{R}^{TK_a \times h}, \quad (2)$$

$$H_0 = g([H_v \circ H_a \circ H_t]) \in \mathbb{R}^{T(K_v+K_a) \times h}, \quad (3)$$

where $\circ$ denotes concatenation operation, H_t is the text embedding. Then, H_0 will forward through each transformer block in the LLMs. In each block, similar to the common use of LoRA, our interaction-aware LoRA structure is added as a bypass branch in all linear layers. For each linear layer, there is a pre-trained and frozen weight W_o. The interaction-aware LoRA structure learns the change of W_o, denoted as ΔW. Given the input hidden embedding $H \in \mathbb{R}^{T(K_v+K_v) \times h}$, we use n heads $\{B_i\}_{i=1}^n$ to learn ΔW, each head is responsible for learning a certain audiovisual data interaction aspect. In order to enable the model to determine how to effectively focus on these aspects, we additionally add an interaction-aware router structure R, which consists of a dense layer with trainable transformation matrix $W_r \in \mathbb{R}^{T(K_v+K_v) \times n}$ and a $softmax(\cdot)$ function to calculate the route scores $S = \{s_i\}_{i=1}^n$ of different interaction aspects. The output of this bypass branch is obtained

Table 1. The comparison results with other general models on all type of tasks. MS3 and AVSS are two subtasks of AVS-Bench. Seen is a subtask of the Ref-AVS test set. The X-InstructBLIP's performance on AVQA is zero-shot. ✓ indicates the model has ability to complete this type of task, but no evaluation is provided in their paper. ✗ indicates the model does not have the corresponding ability.

Method	AVE	AVVP		ARIG		AVQA	MS3(AVS)		AVSS(AVS)		Seen(Ref-AVS)	
	Acc	Segment-level	Event-level	cIoU	AUC	Acc	mIOU	F-score	mIOU	F-score	mIOU	F-score
TimeChat [41]	✓	51.28	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
MEERKAT [12]	✓	<u>54.96</u>	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
GroundingGPT [31]	✓	✓	✓	**44.02**	**0.45**	✓	✗	✗	✗	✗	✗	✗
X-InstructBLIP [38]	✗	✗	✗	✗	✗	44.50	✗	✗	✗	✗	✗	✗
VALOR [6]	✗	✗	✗	✗	✗	<u>78.90</u>	✗	✗	✗	✗	✗	✗
AnyRef [15]	✗	✗	✗	✗	✗	✗	<u>55.6</u>	**66.30**	✓	✓	✓	✓
Crab(Ours)	**80.15**	**59.00**	**54.44**	<u>41.78</u>	<u>0.42</u>	**78.94**	**58.21**	<u>66.24</u>	**26.52**	**32.10**	**40.54**	**0.58**

by the weighted sum of all heads. This is formulated as:

$$S = softmax(W_r H) \in \mathbb{R}^{T(K_v+K_a)\times n}, \tag{4}$$

$$\Delta W = \sum_{i=1}^{n} s_i B_i A H \in \mathbb{R}^{T(K_v+K_a)\times h}, \tag{5}$$

$$H' = W_o H + \Delta W, \tag{6}$$

where H' is the output of this linear layer, A is a shared matrix to learn the common multimodal representation.

4.3. Training Objective

Our overall training objective consists of two parts: the auto-regressive cross-entropy loss $\mathcal{L}_{txt}$ for text generation of all tasks, and the auxiliary loss required for AVS and Ref-AVS tasks $\mathcal{L}_{seg}$. Specifically, we use binary cross-entropy loss $\mathcal{L}_{bce}$ and $\mathcal{L}_{dice}$ for binary mask prediction and cross-entropy loss $\mathcal{L}_{ce}$ for the AVSS task. The final training loss can be denoted as follows:

$$\mathcal{L}_{seg} = \lambda_{bce} \cdot \mathcal{L}_{bce} + \lambda_{dice} \cdot \mathcal{L}_{dice} + \lambda_{ce} \cdot \mathcal{L}_{ce}, \tag{7}$$

$$\mathcal{L} = \lambda_{txt} \cdot \mathcal{L}_{txt} + \lambda_{seg} \cdot \mathcal{L}_{seg}. \tag{8}$$

Here, λ_{txt}, λ_{seg}, λ_{bce}, λ_{dice}, and λ_{ce} are hyper-parameters.

5. Experiments

5.1. Implement Details

Each video is sampled uniformly with 10 frames, and the size of each frame is 224×224. We extract the last layer patch-level representations for each frame. Following BEATs [5], we convert the sample rate of each raw waveform to $16KHz$ and extract 128-dimensional Mel-filter bank features with 25ms Povey window that shifts every 10ms as the acoustic feature. For the mask decoder, we use two `<MASK>` token groups corresponding to two scales of visual features from visual encoder, and we utilize the visual features from the 14th and second-to-last layers. Each group has three tokens. The number of query tokens in both audio Q-Former and visual Q-Former is 32. We use the LLaMA-2-7b-Chat model as our base model. The interaction-aware LoRA structure is employed in all linear layers with a rank of 8. The hyper-parameters λ_{txt}, λ_{seg}, λ_{bce}, λ_{dice}, and λ_{ce} are set to $1.0, 0.5, 1.0, 0.5, 1.0$ respectively. More details in the *supplementary materials*.

Table 2. The comparison results with specialized models on temporal localization task.

Method	AVE task	AVVP task	
	Acc	Segment-level	Event-level
AVT [33]	75.80	-	-
PSP [60]	<u>77.80</u>	-	-
MM-Pyramid [57]	<u>77.80</u>	59.20	53.04
CMBS [54]	-	55.00	48.48
MPN [56]	77.60	-	-
DHHN [20]	-	**60.32**	**55.06**
Crab(Ours)	**80.15**	59.00	<u>54.44</u>

5.2. Training Procedure

We consider a two-stage training paradigm, including pre-training for feature alignment and instruction-tuning for audio-visual tasks.

Stage 1: Pre-training for feature alignment. The purpose of pre-training is to train three multimodal branches in the unified audio-visual interface. We use pre-training data from Video-LLaVA [32] to train visual Q-Former and MLP layer, the audio-caption pairs data from AudioCaps [22] to train audio Q-Former and MLP layer, and image segmentation data from LVIS [14] to train mask decoder. During training process, the weights of visual encoder and audio encoder are frozen. We first pretrain the visual branch and audio branch, and then train the segmentation branch with frozen visual branch weight. We train for three epochs with a global batch size of 256, using the AdamW optimizer with a cosine learning rate schedule. The initial learning rate is set to $1e-4$ with a warmup ratio of 0.03.

Stage 2: Instruction-tuning for audio-visual Tasks. In the second stage, we train all downstream tasks on our AV-UIE dataset. The trainable parameters include three multi-modal branches (excluding the audio encoder and visual encoder) and interaction-aware LoRA. The data in each batch is a random combination from different tasks. We train for five epochs with a global batch size of 512. The remaining settings are the same as the first stage. We provide additional hyper-parameters in the *supplementary materials*.

5.3. Quantitative Results and Analysis

To demonstrate effectiveness of our method, we first compare with other general models across all tasks, and then compare with task-specialized models on each task. For

Table 3. The comparison results with specialized models on MUSIC-AVQA test set.

Method	Audio	Visual	Audio-Visual	Avg
ST-AVQA [26]	73.87	74.40	69.53	71.59
COCA [25]	75.42	75.23	69.96	72.33
PSTP-Net [27]	70.91	77.26	72.57	73.52
LAVISH [34]	75.97	80.22	71.26	74.46
TSPM [29]	**76.91**	83.61	73.51	76.79
Crab(Ours)	76.58	**90.73**	**74.13**	**78.94**

Table 4. The comparison results with specialized models on spatial localization task.

Method	LVS [3]	EZ-VSL [37]	FNAC [45]	Crab(Ours)
cIoU	23.69	26.43	27.15	**41.78**
AUC	0.25	0.29	0.31	**0.42**

Table 5. The comparison results with specialized models on AVS-Bench and Ref-AVS. S4, MS3 and AVSS are the subtasks of AVS-Bench. Seen, Unseen and Null are the subtasks of Ref-AVS.

Method	Backbone	MS3	AVSS	Seen	Unseen	Null(↓)
AVSBench [61]	ResNet-50	54.00	-	0.51	0.55	0.21
TPAVI [61]	PVT-v2	54.00	**29.80**	-	-	-
AVSegFormer [13]	PVT-v2	**58.40**	24.90	33.47	36.05	0.17
GAVS [51]	PVT-v2	-	-	28.93	29.82	0.19
EEMC [52]	PVT-v2	-	-	34.20	**49.54**	**0.01**
Crab(Ours)	**ViT/L-14**	58.21	26.59	**40.54**	45.55	**0.01**

Table 6. The ablation results of explicit reasoning process (denoted as ERP) and interaction-aware LoRA (denoted as IA-LoRA).

Method	AVQA	AVE	AVVP		ARIG	
	Avg	Acc	Segment	Event	cIoU	AUC
w/o. ERP	76.05	78.62	52.01	51.36	40.92	0.41
w/o. IA-LoRA	76.92	79.93	53.54	53.15	40.22	0.40
Crab(Ours)	**78.94**	**80.15**	**59.00**	**54.44**	**41.78**	**0.42**

temporal localization task, we evaluate on AVE [47] fully supervised and LLP [48] test set. For spatio-temporal reasoning task, we evaluate on MUSIC-AVQA test set. For spatial localization task, we adapt AVS-Bench [61] S4 test set to construct an audio referred image grounding test set. For pixel-level understanding task, we compare on AVS [61] and Ref-AVS [52] test set respectively. We present main metrics for AVQA, AVS, and Ref-AVS tasks. Top-2 results are highlighted. In addition to zero-shot results mentioned in Tab. 1, all baseline models include training data for corresponding tasks to ensure fair assessment. More metric results in the *supplementary materials*.

Comparison with other general models. For general models, we keep consistent with foundational LLMs used by official, including LLaMA2-7B, Vicuna-v1.5. Tab. 1 shows the comparison results of our model with other general models on all types of tasks. As shown, our model outperforms other general-purpose models by handling a broader range of audio-visual scene understanding tasks and delivering superior performance across multiple tasks. Specifically, we achieve a significant performance improvements on AVVP tasks (59.00% *vs.* 54.96%). Compared to VALOR [6], which was trained on VALOR-1M dataset, our model achieves comparable results (78.94% *vs.* 78.90%) while utilizing a smaller dataset. Since GroundingGPT [31] focuses on the general ability of fine-grained grounding task, its performance on ARIG task is slightly better than our model. On pixel-level understanding tasks, compared to AnyRef [15], we can accomplish semantic segmentation task (AVSS) and reference segmentation tasks (Ref-AVS). And we also achieve superior result on MS3 task for mIOU metric and comparable result for F-score metric.

Temporal localization task. Tab. 2 shows the event localization accuracy, segment-level and event-level F1-score on temporal localization task. It is evident that our model achieves superior performance compared to specialized methods. Compared to MM-Pyramid [57], we achieves a performance improvement of 2.35% (80.15% *vs.* 77.80%). For AVVP task, our model achieves the comparable performance (59.00% and 54.44%). As discussed in Section 3, while temporal localization task accounts for the smallest proportion in our AV-UIE dataset, explicit inter-task cooperation enables this task benefit from other related audio-visual tasks, leading to improved performance.

Spatial localization and pixel-level understanding task. Tab. 4 shows the experiment results on ARIG task. From the table, it can be seen that our model achieves superior results compared with specialized models. To validate the model's pixel-level understanding capability, we present the experimental results on two representative tasks, AVS and Ref-AVS. In order to maintain a unified audio-visual interface for integration with other tasks, we chose not to use the widely adopted PVT-V2 [50] model tailoredly designed for segmentation tasks as the visual backbone. Instead, we utilized ViT/L-14 to obtain visual features as input for mask decoder. As illustrated in Tab. 5, our model achieves comparable performance on the MS3, AVSS subtasks and Unseen test set. It performs best on the Seen and Null test set.

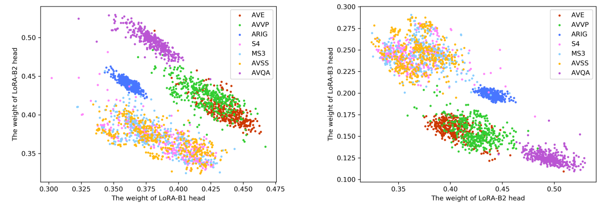
Spatio-temporal reasoning task. Tab. 3 shows the experimental results on MUSIC-AVQA [26] test set. It can be observed that our method outperforms all specialized models. Specifically, compared to recent TSPM [29], our model achieves significant overall performance improvements of 2.15% (78.94% *vs.* 76.79%). On visual subtask, our method achieves remarkable improvements of 7.12% (90.73% *vs.* 83.61%). For complex audio-visual question type, our model obtains the best overall performance. And the performance in audio subtask is also comparable.

5.4. Ablation Results

To verify the effectiveness of explicit reasoning process and interaction-aware LoRA structure, we conduct two ablation experiments, including using only simple labels from original dataset and using a standard LoRA. Tab. 6 presents the corresponding experiment results. When model does not output explicit reasoning process, the performance on AVQA, AVE, and AVVP tasks decreases significantly, while for the ARIG, AVS, and Ref-AVS tasks, the impact is

Table 7. The LoRA head drop results on interaction-aware LoRA. *drop head-1* means the weight of *head-1* is set to zero during inference process. And so on for the others.

Method	AVE	ARIG		AVQA
	Acc	cIoU	AUC	Acc
drop head-1	75.07	36.15	0.36	76.33
drop head-2	79.25	29.16	0.29	76.22
drop head-3	79.62	38.65	0.39	77.01
no drop	**80.15**	**41.78**	**0.42**	**78.94**



(a) The router weight of *LoRA-B1* and *LoRA-B2* head.

(b) The router weight of *LoRA-B2* and *LoRA-B3* head.

Figure 5. We visualize the router weights of three LoRA heads on different tasks. Figure (a) compares *head-B1* and *head-B2*, while figure (b) compares *head-B2* and *head-B3*. Different colors distinguish between tasks. The larger the router weight value, the greater the task's dependence on that LoRA head, indicating this LoRA head has a stronger ability to solve this type of task.

minimal. This is primarily because the former task demands simultaneous comprehension of complex video and audio information, which heavily depends on the reasoning process. In contrast, the latter primarily relies on the model's pixel-level understanding of visual content, where reasoning is not the main factor limiting performance. Furthermore, our interaction-aware LoRA structure consistently outperforms standard LoRA across all tasks. More results for other tasks in the *supplementary materials*.

5.5. Interaction-aware LoRA Analysis

Visualized results. To understand how tasks cooperate, we visualize the dependency of different tasks on LoRA heads during the inference process. As shown in Fig. 5, two key observations emerge: (1) Tasks of the same type exhibit similar dependencies on various LoRA heads. For instance, the temporal localization tasks like AVE and AVVP, the pixel-level understanding tasks like S4, MS3, and AVSS, tend to cluster together, while other task types, including AVQA and ARIG, form distinct clusters. It should be noted that this phenomenon is not caused by the homogeneity of data from tasks of the same type. The most intuitive example is our ARIG dataset, which is derived from AVSBench data, but these tasks do not form the same cluster. Moreover, the data from S4 and MS3 are not homogenous. However, they still cluster together because they belong to the same type of task and require similar skills to complete. 2) The varying dependence of different types of tasks on the three LoRA heads surprisingly suggests that these heads possess distinct capabilities. It can be found that the AVE and AVVP tasks rely significantly more on *head-B1* and *head-B2* than on *head-B3*. The ARIG task is most dependent on *head-B2*, followed by *head-B1*, and finally *head-B3*. Therefore, we can infer that *head-B1* has stronger temporal localization capability, *head-B2* has stronger spatial localization capability, and *head-B3* is responsible for learning pixel-level understanding capability. Similarly, since the AVQA task requires stronger spatio-temporal reasoning capability, it also mainly relies on *head-B1* and *head-B2*, and less on *head-B3*. Notably, the AVS task primarily relies on *head-B1* and *head-B2*, with a lesser dependence on *head-B3*. This distribution may also contribute to the model's performance on the AVS task. Despite this, compared to other tasks, the AVS task still has the greater dependence on *head-B3*, reflecting that *head-B3* does have pixel-level understanding capability, although it is not very strong.

Drop LoRA head results. Through above analysis, each LoRA head indeed possesses different capabilities. Furthermore, during the inference process, we set the weights of three LoRA heads to zero respectively, also known as "*drop head*". We find that setting weight directly to zero would affect the model's behavior to some extent. Therefore, we only report the comparison results under the model's normal behavior. The Tab. 7 presents corresponding results. From the table, we can observe that for temporal localization, spatial localization, and spatio-temporal reasoning tasks, dropping any head will result in a performance decline. Specifically, for AVE task, when we drop the head with temporal localization capability, namely *head-1*, the performance is the worst. For the ARIG task, the performance is the worst when we drop *head-2*. As for the AVQA task, the impact of these two heads on performance is relatively similar.

6. Conclusion

In this paper, we propose a unified audio-visual scene understanding model with explicit cooperation. To achieve this goal, we construct AV-UIE dataset, which clarifies the explicit cooperative relationship among tasks. Then we design an interaction-aware LoRA structure with multiple heads, each head is responsible for learning certain audio-visual data interaction aspect. Furthermore, we are surprised to discover that each head possesses certain capability through the visualized experimental results.

In future work, we plan to further enhance the model's pixel-level understanding capability. Additionally, while we have established connections among tasks through explicit reasoning process, its accuracy is not guaranteed, which can impact model's performance. Therefore, another possible major research direction is to enhance the reliability of reasoning process, thereby further improving the overall performance of the model.

Acknowledgement The project was supported by National Natural Science Foundation of China (NO.62106272) and sponsored by CCF-Zhipu.AI Large Model Innovation Fund.

References

[1] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. 2

[2] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[3] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16867–16876, 2021. 7

[4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[5] 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. 5, 6

[6] 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. 4, 6, 7

[7] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022. 3

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 3

[9] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 3

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 2

[11] 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. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 2, 3

[12] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. *arXiv preprint arXiv:2407.01851*, 2024. 2, 3, 4, 6

[13] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12155–12163, 2024. 2, 7

[14] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 6

[15] Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. Multi-modal instruction tuned llms with fine-grained visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13980–13990, 2024. 6, 7

[16] Wenxuan Hou, Guangyao Li, Yapeng Tian, and Di Hu. Toward long form audio-visual video understanding. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–26, 2024. 2

[17] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020. 2

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 5

[19] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023. 3

[20] Xun Jiang, Xing Xu, Zhiguo Chen, Jingran Zhang, Jingkuan Song, Fumin Shen, Huimin Lu, and Heng Tao Shen. Dhhn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 719–727, 2022. 6

[21] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, page 2. Minneapolis, Minnesota, 2019. 3

[22] 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. 6

[23] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 9579–9589, 2024. 5

[25] Mingrui Lao, Nan Pu, Yu Liu, Kai He, Erwin M Bakker, and Michael S Lew. Coca: Collaborative causal regularization for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12995–13003, 2023. 7

[26] 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. 2, 3, 7

[27] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7808–7816, 2023. 7

[28] Guangyao Li, Yixin Xu, and Di Hu. Multi-scale attention for audio question answering. *arXiv preprint arXiv:2305.17993*, 2023. 2

[29] Guangyao Li, Henghui Du, and Di Hu. Boosting audio visual question answering via key semantic-aware cues. *arXiv preprint arXiv:2407.20693*, 2024. 2, 7

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 4

[31] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6657–6678, 2024. 4, 6, 7

[32] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6

[33] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 6

[34] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023. 2, 7

[35] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[36] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024. 3

[37] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer, 2022. 7

[38] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2023. 6

[39] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[40] 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. 4

[41] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 4, 6

[42] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 5

[43] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 2

[44] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024. 3

[45] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023. 2, 7

[46] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *arXiv preprint arXiv:2404.19245*, 2024. 5

[47] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 2, 3, 7

[48] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 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*, pages 436–454. Springer, 2020. 2, 3, 7

[49] 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. 2

[50] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 7

[51] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5669–5677, 2024. 7

[52] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. *arXiv preprint arXiv:2407.10957*, 2024. 2, 3, 7

[53] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022. 2

[54] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19989–19998, 2022. 6

[55] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. 3

[56] Jiashuo Yu, Ying Cheng, and Rui Feng. Mpn: Multimodal parallel network for audio-visual event localization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 6

[57] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6241–6249, 2022. 2, 6, 7

[58] 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. 2

[59] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2, 3

[60] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. 6

[61] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 2, 3, 7

[62] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024. 2