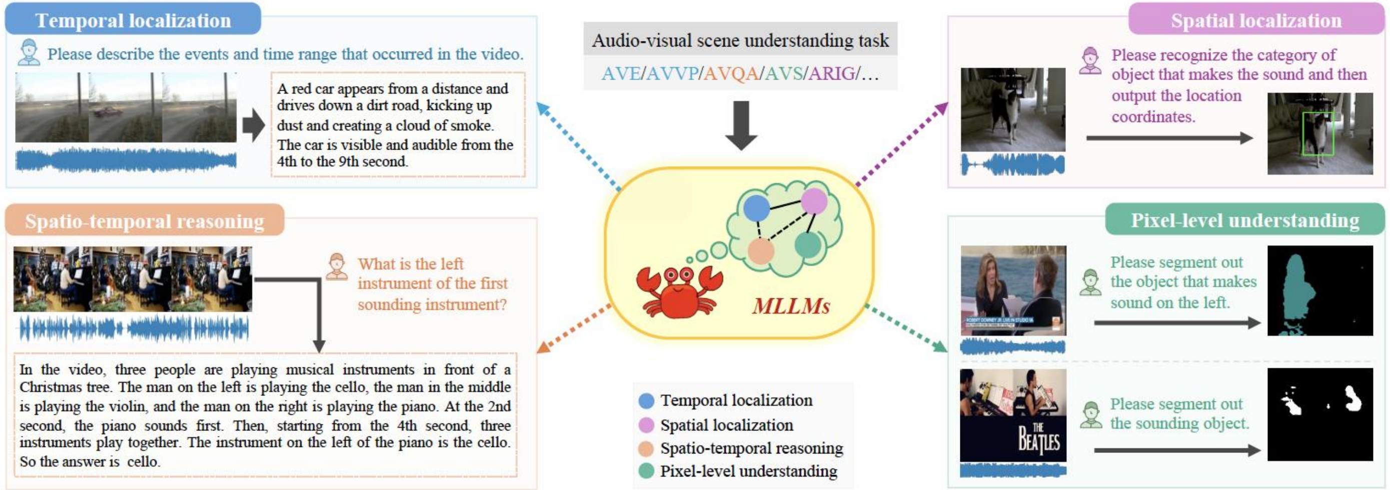
Crab: A Unified Audio-Visual Scene Understanding Model with Explicit Cooperation

Henghui Du[1], 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
[3] AI Technology Center, Online Video Business Unit, Tencent PCG, Beijing



Temporal localization

Please describe the events and time range that occurred in the video.

A red car appears from a distance and drives down a dirt road, kicking up dust and creating a cloud of smoke. The car is visible and audible from the 4th to the 9th second.

Spatio-temporal reasoning

What is the left instrument of the first sounding instrument?

In the video, three people are playing musical instruments in front of a Christmas tree. The man on the left is playing the cello, the man in the middle is playing the violin, and the man on the right is playing the piano. At the 2nd second, the piano sounds first. Then, starting from the 4th second, three instruments play together. The instrument on the left of the piano is the cello. So the answer is cello.

Audio-visual scene understanding task

AVE/AVVP/AVQA/AVS/ARIG/…

- ● Temporal localization
- ● Spatial localization
- ● Spatio-temporal reasoning
- ● Pixel-level understanding

Spatial localization

Please recognize the category of object that makes the sound and then output the location coordinates.

Pixel-level understanding

Please segment out the object that makes sound on the left.

Please segment out the sounding object.

Content

- Introduction
- Related work
- Motivation
- Method
- Experiments

Introduction

■ Audio-visual scene understanding tasks

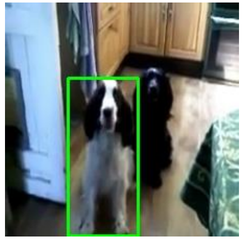
➢ Temporal localization

Audio-Visual Event Localization (AVE)
Audio-Visual Video Parsing (AVVP)

Helicopter, [3,6]

➢ Spatial localization

Audio Referred Image Grounding (ARIG)



Top left: (60, 72)
Bottom right: (127, 223)

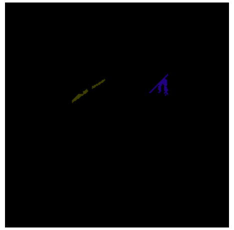
➢ Spatial-temporal reasoning

Audio-Visual Question Answering (AVQA)

Question: Which xylophone makes the sound first?
Answer: Right

➢ Pixel-level understanding

Audio-Visual Segmentation (AVS)
Reference Audio-Visual Segmentation (Ref-AVS)

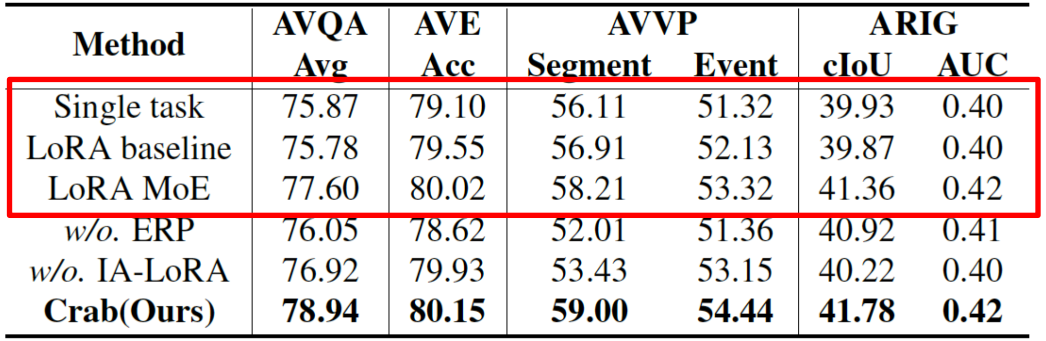


Motivation

- Is the mainstream learning paradigm of building a large-scale instruction-tuning dataset and directly doing multi-task instruction-tuning the best?

- Especially for multimodal scene understanding tasks with large task differences, how to effectively alleviate the mutual interference among tasks and promote explicit cooperation among tasks?

Table 4. More comprehensive ablation results. ERP represents reasoning process. IA-LoRA represents interaction-aware LoRA.

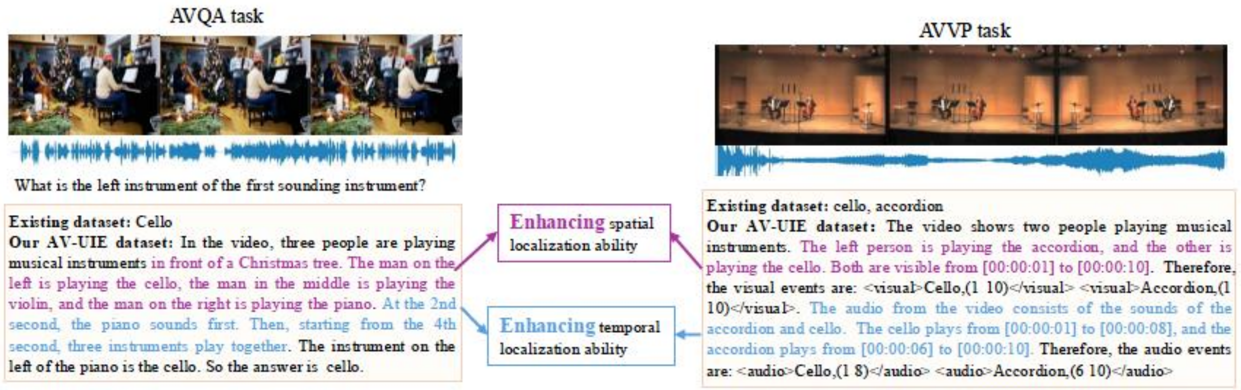


Method	AVQA	AVE	AVVP		ARIG	
	Avg	Acc	Segment	Event	cIoU	AUC
Single task	75.87	79.10	56.11	51.32	39.93	0.40
LoRA baseline	75.78	79.55	56.91	52.13	39.87	0.40
LoRA MoE	77.60	80.02	58.21	53.32	41.36	0.42
w/o. ERP	76.05	78.62	52.01	51.36	40.92	0.41
w/o. IA-LoRA	76.92	79.93	53.43	53.15	40.22	0.40
Crab(Ours)	**78.94**	**80.15**	**59.00**	**54.44**	**41.78**	**0.42**

Method

■ Data perspective

Refine the labels of existing datasets to include specific spatiotemporal information and clarify the explicit cooperation relationship among tasks.


AVQA task
AVVP task
Enhancing spatial localization ability
Enhancing temporal localization ability

What is the left instrument of the first sounding instrument?

Existing dataset: Cello
Our AV-UIE dataset: In the video, three people are playing musical instruments in front of a Christmas tree. The man on the left is playing the cello, the man in the middle is playing the violin, and the man on the right is playing the piano. At the 2nd second, the piano sounds first. Then, starting from the 4th second, three instruments play together. The instrument on the left of the piano is the cello. So the answer is cello.

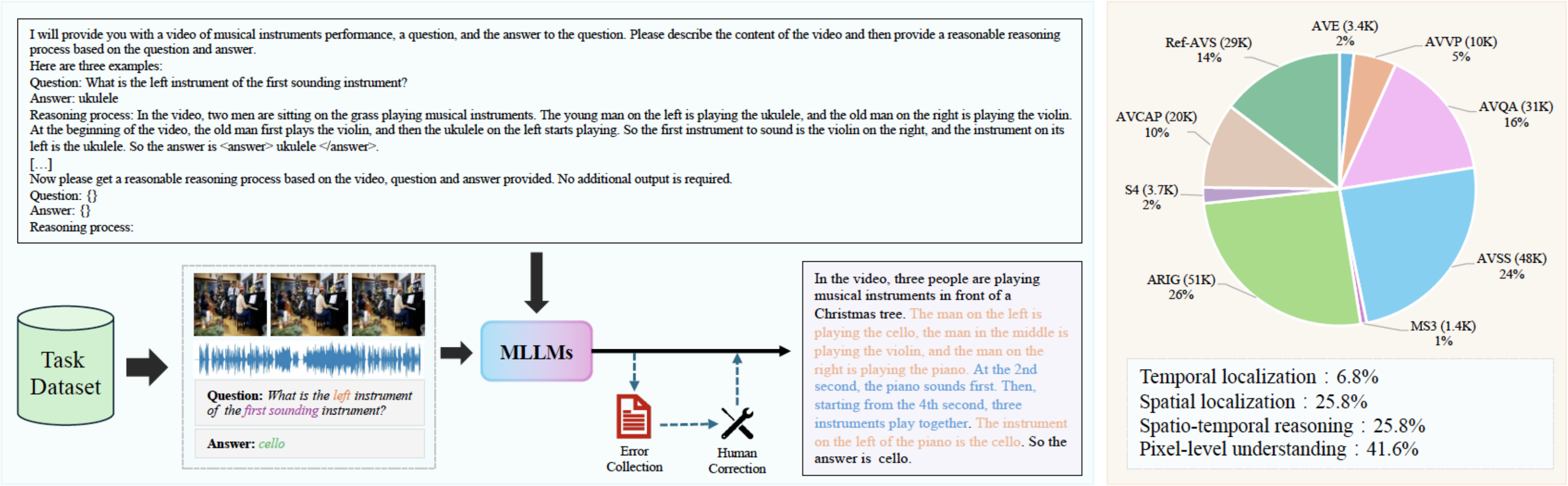
Existing dataset: cello, accordion
Our AV-UIE dataset: The video shows two people playing musical instruments. The left person is playing the accordion, and the other is playing the cello. Both are visible from [00:00:01] to [00:00:10]. Therefore, the visual events are: <visual>Cello,(1 10)</visual> <visual>Accordion,(1 10)</visual>. The audio from the video consists of the sounds of the accordion and cello. The cello plays from [00:00:01] to [00:00:08], and the accordion plays from [00:00:06] to [00:00:10]. Therefore, the audio events are: <audio>Cello,(1 8)</audio> <audio>Accordion,(6 10)</audio>

Figure 11. AVQA and AVVP tasks achieve explicit cooperation through explicit reasoning process.

Method

Data perspective

Audio-**V**isual **U**nified **I**nstruction-tuning dataset with **E**xplicit reasoning process (**AV-UIE dataset**)



I will provide you with a video of musical instruments performance, a question, and the answer to the question. Please describe the content of the video and then provide a reasonable reasoning process based on the question and answer.
Here are three examples:
Question: What is the left instrument of the first sounding instrument?
Answer: ukulele
Reasoning process: In the video, two men are sitting on the grass playing musical instruments. The young man on the left is playing the ukulele, and the old man on the right is playing the violin. At the beginning of the video, the old man first plays the violin, and then the ukulele on the left starts playing. So the first instrument to sound is the violin on the right, and the instrument on its left is the ukulele. So the answer is <answer> ukulele </answer>.
[…]
Now please get a reasonable reasoning process based on the video, question and answer provided. No additional output is required.
Question: {}
Answer: {}
Reasoning process:

Question: *What is the left instrument of the first sounding instrument?*

Answer: cello

In the video, three people are playing musical instruments in front of a Christmas tree. The man on the left is playing the cello, the man in the middle is playing the violin, and the man on the right is playing the piano. At the 2nd second, the piano sounds first. Then, starting from the 4th second, three instruments play together. The instrument on the left of the piano is the cello. So the answer is cello.

Temporal localization：6.8%
Spatial localization：25.8%
Spatio-temporal reasoning：25.8%
Pixel-level understanding：41.6%

(a) The procedure of constructing dataset.

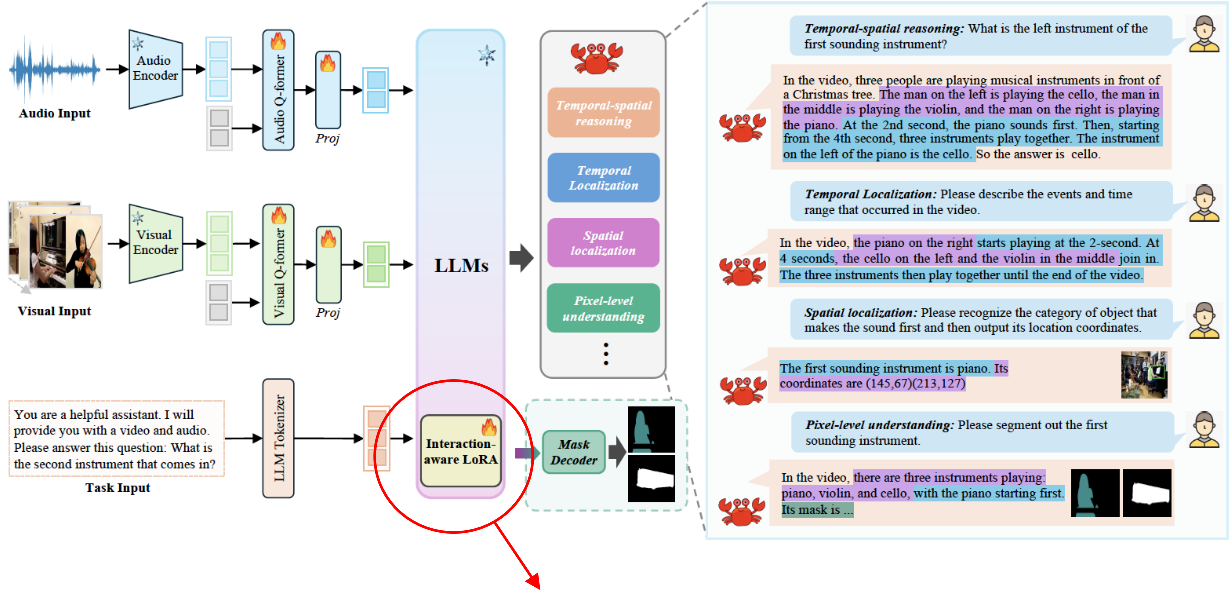
b) The data statistics for each task.

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

Method

Model perspective

Each LoRA head is responsible for learning different types of data interactions, decoupling the model’s capabilities.
During the learning process, when the capabilities of a head are enhanced, other types of tasks can benefit from the same hea d.



You are a helpful assistant. I will provide you with a video and audio. Please answer this question: What is the second instrument that comes in?

Temporal-spatial reasoning: What is the left instrument of the first sounding instrument?

In the video, three people are playing musical instruments in front of a Christmas tree. The man on the left is playing the cello, the man in the middle is playing the violin, and the man on the right is playing the piano. At the 2nd second, the piano sounds first. Then, starting from the 4th second, three instruments play together. The instrument on the left of the piano is the cello. So the answer is cello.

Temporal Localization: Please describe the events and time range that occurred in the video.

In the video, the piano on the right starts playing at the 2-second. At 4 seconds, the cello on the left and the violin in the middle join in. The three instruments then play together until the end of the video.

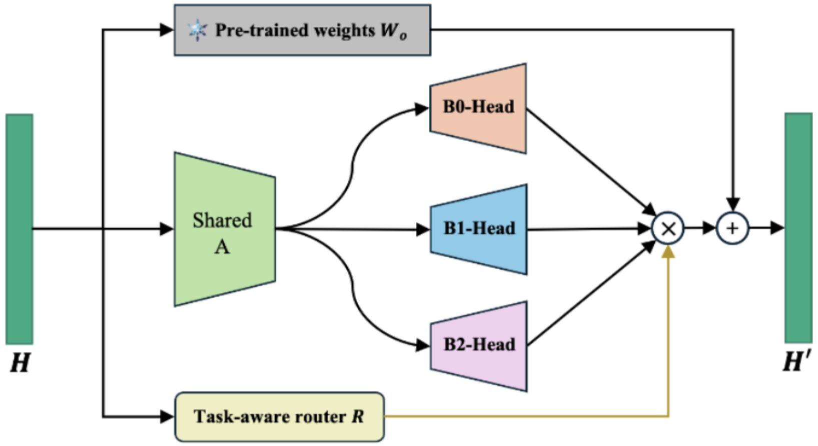
Spatial localization: Please recognize the category of object that makes the sound first and then output its location coordinates.

The first sounding instrument is piano. Its coordinates are (145,67)(213,127)

Pixel-level understanding: Please segment out the first sounding instrument.

In the video, there are three instruments playing: piano, violin, and cello, with the piano starting first. Its mask is ...

Decoupling model’s capabilities



Experiments

■ Comparison with general models

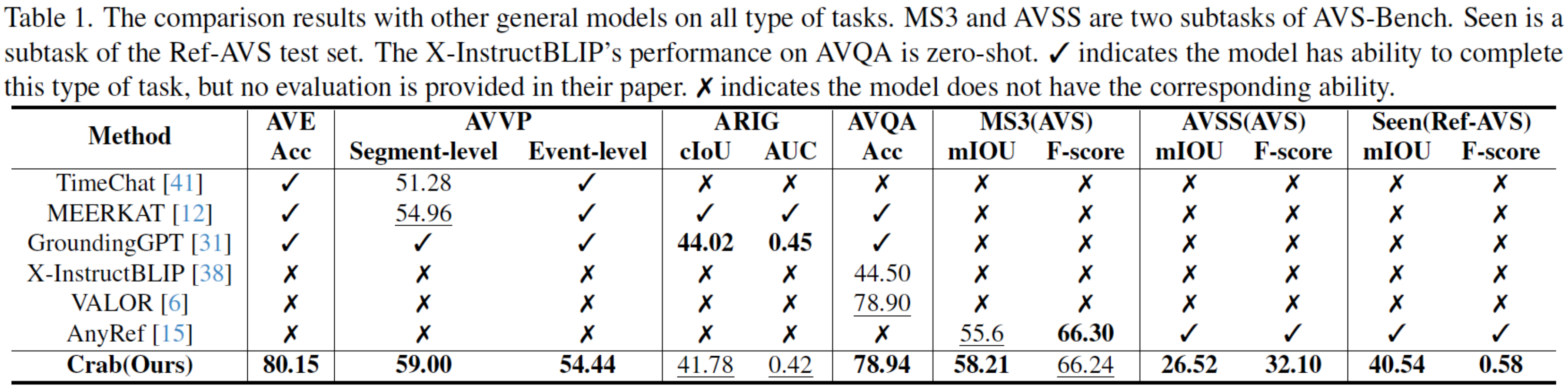


Table 1. The comparison results with other general models on all type of tasks. MS3 and AVSS are two subtasks of AVS-Bench. Seen is a subtask of the Ref-AVS test set. The X-InstructBLIP's performance on AVQA is zero-shot. ✓ indicates the model has ability to complete this type of task, but no evaluation 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]	✓	54.96	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
GroundingGPT [31]	✓	✓	✓	**44.02**	**0.45**	✓	✗	✗	✗	✗	✗	✗
X-InstructBLIP [38]	✗	✗	✗	✗	✗	44.50	✗	✗	✗	✗	✗	✗
VALOR [6]	✗	✗	✗	✗	✗	78.90	✗	✗	✗	✗	✗	✗
AnyRef [15]	✗	✗	✗	✗	✗	✗	55.6	**66.30**	✓	✓	✓	✓
Crab(Ours)	**80.15**	**59.00**	**54.44**	41.78	0.42	**78.94**	**58.21**	66.24	**26.52**	**32.10**	**40.54**	**0.58**

Experiments

■ Comparison with specialized models

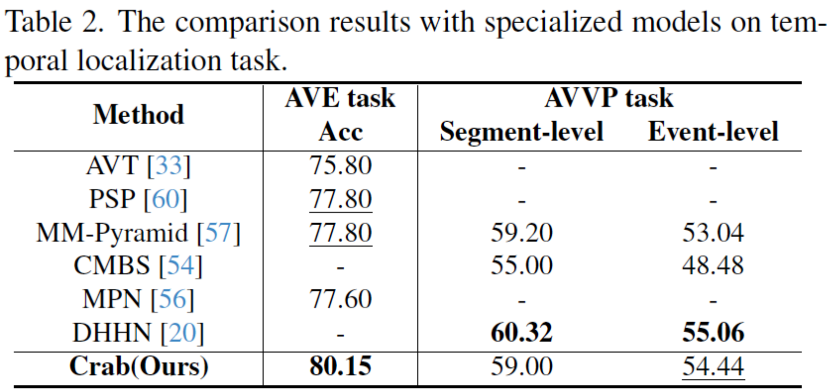


Table 2. The comparison results with specialized models on temporal localization task.

Method	AVE task	AVVP task	
	Acc	Segment-level	Event-level
AVT [33]	75.80	-	-
PSP [60]	77.80	-	-
MM-Pyramid [57]	77.80	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	54.44

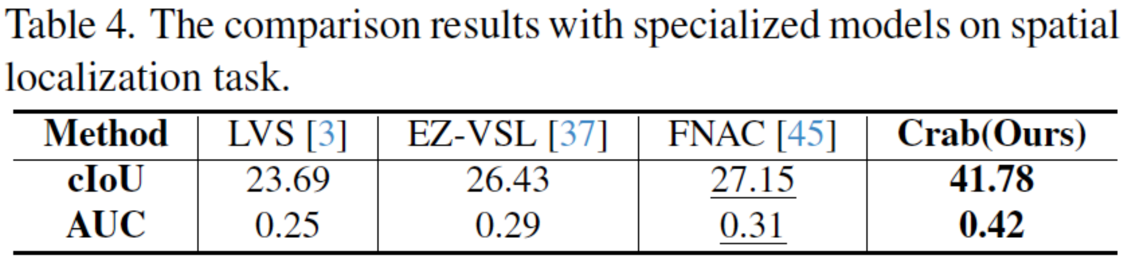


Table 4. The comparison results with specialized models on spatial localization task.

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**

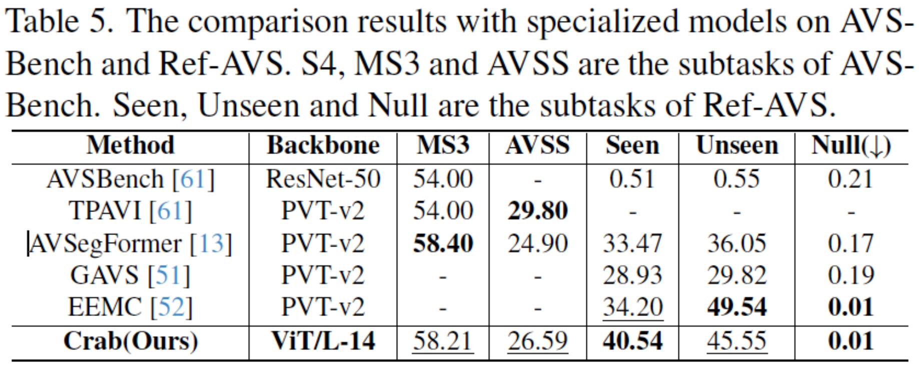


Table 5. The comparison results with specialized models on AVS-Bench and Ref-AVS. S4, MS3 and AVSS are the subtasks of AVS-Bench. Seen, Unseen and Null are the subtasks of Ref-AVS.

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**

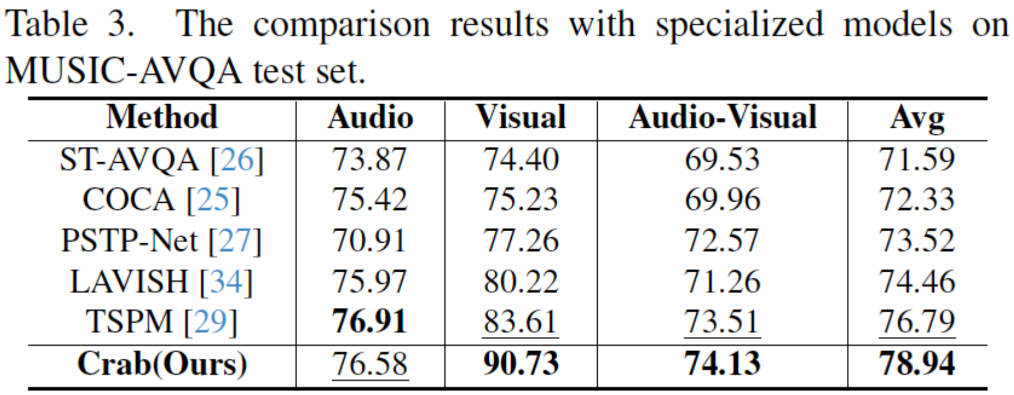
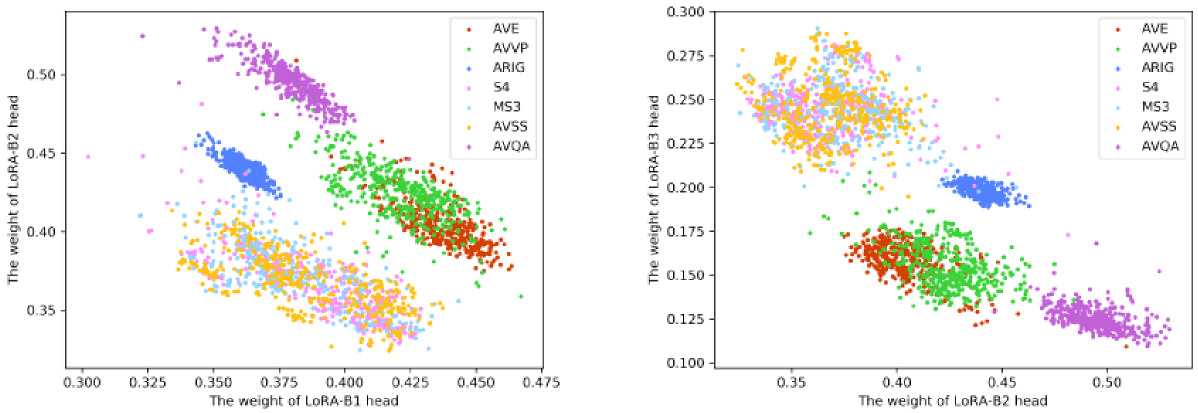


Table 3. The comparison results with specialized models on MUSIC-AVQA test set.

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**

Experiments

■ Visualization results



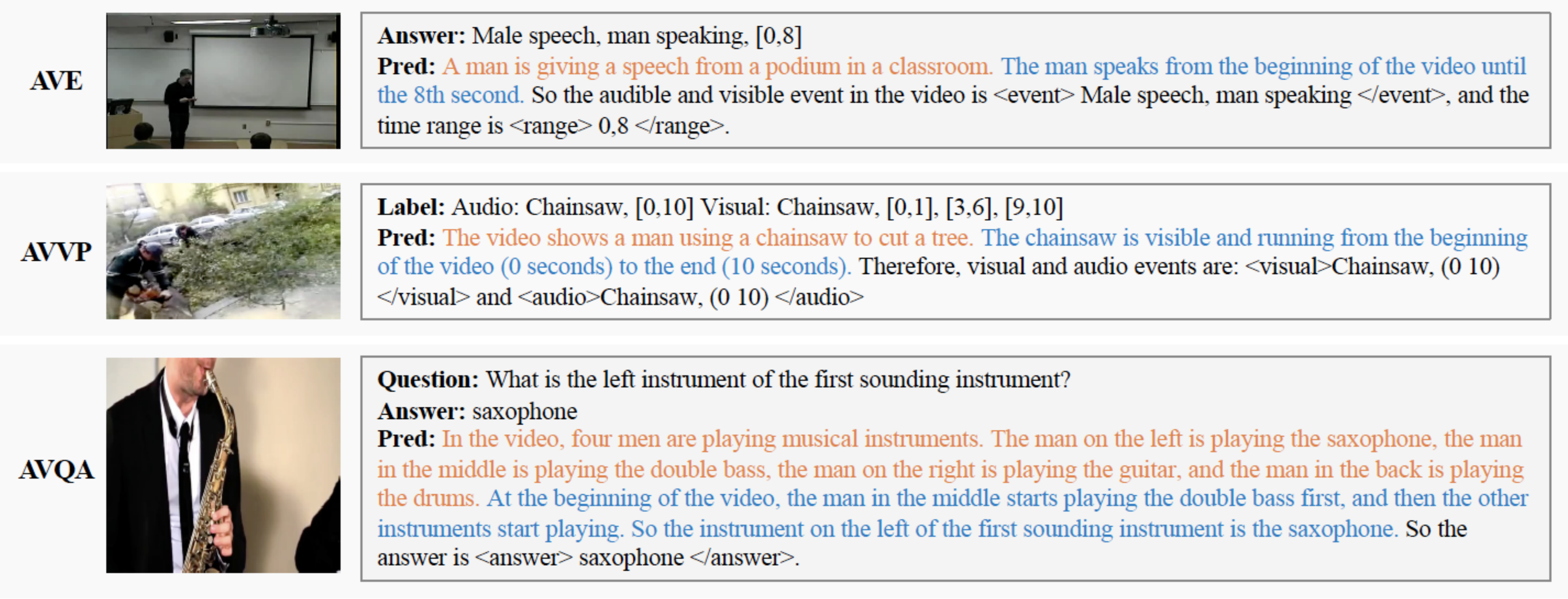
(a) The router weight of *LoRA-B1* and *LoRA-B2* head.

(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.

Experiments

Visualization results



AVE

Answer: Male speech, man speaking, [0,8]
Pred: A man is giving a speech from a podium in a classroom. The man speaks from the beginning of the video until the 8th second. So the audible and visible event in the video is <event> Male speech, man speaking </event>, and the time range is <range> 0,8 </range>.

AVVP

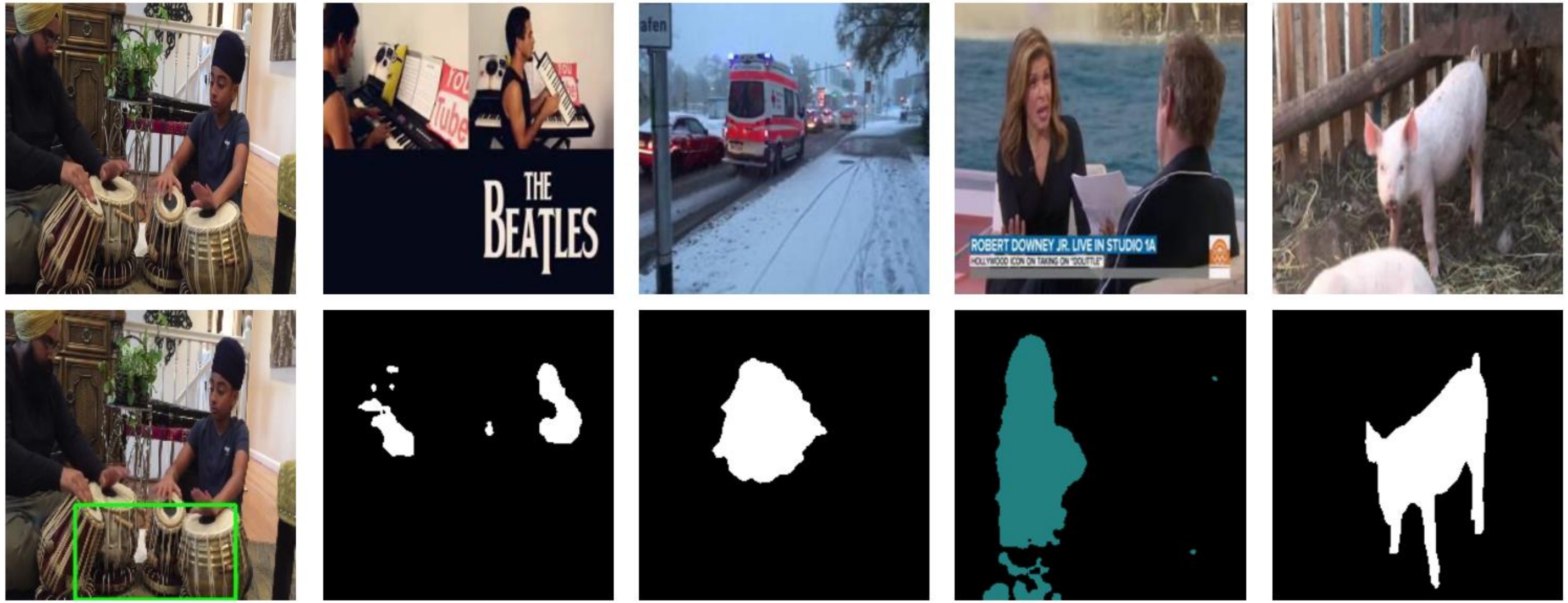
Label: Audio: Chainsaw, [0,10] Visual: Chainsaw, [0,1], [3,6], [9,10]
Pred: The video shows a man using a chainsaw to cut a tree. The chainsaw is visible and running from the beginning of the video (0 seconds) to the end (10 seconds). Therefore, visual and audio events are: <visual>Chainsaw, (0 10) </visual> and <audio>Chainsaw, (0 10) </audio>

AVQA

Question: What is the left instrument of the first sounding instrument?
Answer: saxophone
Pred: In the video, four men are playing musical instruments. The man on the left is playing the saxophone, the man in the middle is playing the double bass, the man on the right is playing the guitar, and the man in the back is playing the drums. At the beginning of the video, the man in the middle starts playing the double bass first, and then the other instruments start playing. So the instrument on the left of the first sounding instrument is the saxophone. So the answer is <answer> saxophone </answer>.

Experiments

- **Visualization results**



ARIG MS3 S4 AVSS Ref-AVS

中国人民大學
RENMIN UNIVERSITY OF CHINA

GeWu-Lab
Gaoling School of Artificial Intelligence
Renmin University of China

Thank You for Listening!

Project: https://gewu-lab.github.io/Crab/
Model: https://huggingface.co/ahsgdxhs/Crab
Dataset: https://huggingface.co/datasets/ahsgdxhs/AVUIE
Arxiv: https://arxiv.org/pdf/2503.13068