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

Supplementary Material

1. Dataset Construction

In this section, we introduce the prompt used in the construction of our AV-UIE dataset and some examples in dataset. We have provided the code and annotation information for the dataset in the attachment. The audio and video data can be downloaded from these links: AVE, AVVP, ARIG, AVS, Ref-AVS, MUSIC-AVQA, VALOR.

1.1. Prompt Template

To construct the AV-UIE dataset, we use the in-context learning approach to prompt Gemini 1.5 Pro to transform simple labels into explicit reasoning process. Fig. 1, Fig. 2 and Fig. 3 demonstrate the prompt for AVE, AVVP, and AVQA tasks respectively. Specifically, for each instance in the dataset, our prompt first includes several input-output pairs, allowing the Gemini 1.5 Pro to generate content in a fixed format based on these examples. Subsequently, we provide audio, video, and simple labels to Gemini 1.5 Pro, which then outputs the reasoning process based on the these provided information, thereby clarifying cooperation relationship among audio-visual tasks. Throughout this process, we ensure that the transformed labels remain consistent with the original ones.

1.2. Dataset Examples

To help readers better understand the format of our dataset, we introduce a few examples below.

AVE task. Fig. 4 are two examples on AVE task in our dataset. The video in Fig. 4(a) shows a scene of a race car driving in an open space in front of a building, which lasts from the beginning to the end of the video, so the original label is *"Race car, auto racing, [0,10]"*. As can be seen, in our dataset, the transformed labels include an explicit reasoning process, which is conducive to clarify cooperative relationship among tasks.

AVVP task. Fig. 5 shows the label transformation results of the LLP dataset. Since there are only video-level event labels in the training set of LLP, such as *"Speech, Baby_laughter"* shown in the Fig. 5(b), but no specific time boundaries for the occurrence of events. We further annotate the temporal information of these events using Gemini 1.5 Pro and output the explicit reasoning process.

ARIG task. As shown in Fig. 6, we convert the segmentation mask in the AVS-Bench training set into bounding boxes, then obtain the two vertices at the top left and bottom right, and represent them in the form of $[x_{Left}, y_{Top}, x_{Right}, y_{Bottom}]$.

AVS task. As shown in Fig. 7, the transformed labels additionally contain temporal and spatial information. The model first needs to accurately understand the audiovisual scene, then point out the target object, and finally predict the segmentation mask based on these information.

AVQA task. Fig. 8 shows the results after label transformation on MUSIC-AVQA dataset. It can be found that the transformed explicit reasoning process contains spatio-temporal reasoning information, such as the localization of instruments, the accurate time of sound, and the order of instruments occurrence, *etc*. These information can help the model establish relationships and achieve mutual cooperation among tasks.

1.3. Explicit Cooperation among Tasks

To further illustrate how our AV-UIE dataset achieves explicit cooperation among tasks, we provide two data examples in Fig 11. Existing audiovisual datasets only have single words "cello, accordion". In our AV-UIE dataset, explicit reasoning process includes *temporal and spatial* information (marked with different colors in Fig 11), which can effectively achieve explicit cooperation among tasks. For example, in AVVP task, there are specific timestamps, which can enhance model's temporal localization ability, and thus help AVQA task. This is also a major difference between our dataset and existing instruction-tuning datasets, which do not emphasize concrete temporal and spatial information to achieve task cooperation.

2. Mask decoder

Fig. 9 illustrates the detailed process of the mask decoder outputting the segmentation mask. As mentioned in the main paper, we obtain the LLM last-layer hidden states corresponding to the `<mask>` tokens. Inspired by the multi-scale features of PVT-v2 [11], these hidden states are divided into two scales. Then, we use learnable weighting factors to combine all the hidden states from each scale, resulting in two embeddings. Each embedding serves as a prompt and is sent to the mask decoder along with the corresponding scale of visual features. The mask decoder primarily consists of two cross-attention blocks, each block responding for the interaction at one scale. For each mask generation, mask decoder sequentially produces a mask score map, which then guides the model to focus atten-

I will provide you with a video, the audible and visible events that occurred in the video and the time range of the events. Please describe event in the video based on this information.

Here are three examples:
Event: Church bell
Time range: 7,10
Event description: There is a church from the beginning to the end of the video. From the beginning of the video to the 7th second, the church did not make any sound. From the 7th second to the end of the video, that is, the 10th second, the church has been making sounds. So the audible and visible event in the video is <event> church bell </event>, and the time range is <range> 7,10 </range>.

Event: Male speech, man speaking
Time range: 2,6
Event description: In this video, a man is standing on the stage and speaking, and there are many audiences listening below. The man's speech lasts from the 2nd second to the 6th second, followed by the cheers of the audience below. So the audible and visible event in the video is <event> Male speech, man speaking </event>, and the time range is <range> 2,6 </range>.

Occurrence event: Bark
Time range: 0,4
Event description: At the beginning of the video, there is a dog on a leash, which is barking until the 4th second. At 7 seconds the dog reappears, but does not bark. So the audible and visible event in the video is <event> Bark </event>, and the time range is <range> 0,4 </range>.

Now please give a description based on the video, occurrence event and time range provided. No other additional output is required.
Occurrence event: {}
Time range: {}
Event description:

Figure 1. The prompt used to convert labels on AVE task.

I will provide you with a video and the events that occurred in the video. Please determine the time range of the event from the visual and audio information of the video. Here are two examples:

Events: Motorcycle, Speech, Singing
Event description: The video shows a person riding a motorcycle on the top of a mountain, which lasts from the beginning to the end of the video, that is, from the 0th second to the 10th second. Therefore, the visual event includes: <visual>Motorcycle,(0 10)</visual>. The sound of the video includes the sound of the person speaking, the sound of the motorcycle engine starting, and the sound of singing. Among them, the sound of the person speaking starts from the 1st second to the 7th second, and then starts again at the 9th second until the end of the video, that is, the 10th second. The sound of the motorcycle engine starts from the 4th second and ends at the 6th second. The singing sound lasts from the 5th second to the 10th second. Therefore, the sound events include: <audio>Speech,(1 7),(9 10)</audio> <audio>Motorcycle,(4 6)</audio> <audio>Singing,(5 10)</audio>.

Events: Violin_fiddle, Cello, Singing, Clapping, Speech
Event description: From the video screen, we can see four people playing musical instruments, including violin_fiddle and cello, which are always present in the screen. Therefore, the visual events include: <visual>Violin_fiddle,(0 10)</visual> <visual>Cello,(0 10)</visual>. From the sound of the video, we can hear the playing of Violin_fiddle and Cello, from the beginning of the video to the end of the 9th second. The sound of the character singing is accompanied by the sound of the playing, from the beginning of the video to the end of the 8th second, the sound of the character speaking appears at the 9th second, and finally the sound of clapping appears at the 9th second and ends at the 10th second. So the sound events include: <audio>Violin_fiddle,(0 9)</audio> <audio>Cello,(0 9)</audio> <audio>Singing,(0 8)</audio> <audio>Clapping,(9 10)</audio> <audio>Speech,(9 9)</audio>.

Now please give the event description according to the provided video and events in the format of examples.
Events: {}
Event description:

Figure 2. The prompt used to convert labels on AVVP task.

tion on areas of higher relevance at the next scale, thereby promoting more accurate mask generation. Finally, we use learnable weighting factors to combine the mask maps from all scales to obtain the final segmentation mask.

3. Experiment Results

3.1. More Comprehensive Ablation Results

Multitask interference was mentioned in previous works [1]. It has often been overlooked in MLLMs. We provide experiments in Tab **??**. Comparing single task and LoRA baseline (training jointly on multitask), increasing task numbers indeed improves model's performance (AVE and AVVP), but it could also introduce the issue of interference (AVQA and ARIG).

Our interaction-aware LoRA structure is a special MoE structure, where each decoder head can be seen as an expert with specific ability. We also compared results of LoRA baseline and LoRA MoE in Tab **??**, which can also prove effectiveness of this structure.

3.2. Pixel-level understanding

Tab. 1 shows the experimental results compared with specialized models on the AVS-Bench [15] and Ref-AVS[13] test sets. It can be seen that our model achieves comparable results on the MS3, AVSS, and Unseen subtasks, and performed best on the Seen and Null subtasks, but performed poorly on the S4 subtask. The AVS task uses audio as guiding information to find out the target object to be segmented, while text reference expressions are used in the Ref-AVS task. Since LLMs naturally have stronger understanding and reasoning capabilities for text, the performance on the three subtasks of Ref-AVS is generally better. In addition, while the model can accurately determine the position of target object, the mask decoder is responsible for outputting the segmentation mask, which will also affect the final re-

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: Is the flute on the left louder than the piano on the right?
Answer: yes
Reasoning process: In the video, two girls are sitting together playing musical instruments. The girl on the left is playing the flute, and the girl on the right is playing the piano. They are playing from the beginning to the end of the video. From the sound, we can judge that the sound of the flute is louder than that of the piano. So the answer is <answer> yes </answer>.

Question: Where is the lowest instrument?
Answer: left
Reasoning process: There are two people standing on the stage in the video. The woman on the left is holding a flute and the man on the right is holding a tuba. There is a piano behind them, but no one is playing it. At the beginning of the video, the woman plays the flute first, and then the man starts playing the tuba. The sound of the flute on the left is the lowest. So the answer is <answer> left </answer>.

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:

Figure 3. The prompt used to convert labels on AVQA task.

(a) 

Answer: Race car, auto racing, [0,10]
Converted answer: A race car is driving and drifting around a paved area in front of a large building. The car is visible and audible from the beginning of the video, at 0 seconds, until the end of the video, at 10 seconds. So the event in the video is <event> Race car, auto racing </event>, and the time range is <range> 0,10 </range>.

(b) 

Answer: Helicopter, [3,6]
Converted answer: A helicopter is visible and audible from the third second to the sixth second of the video. After the sixth second, the helicopter can still be seen, but it is no longer clearly audible. Therefore, the audible and visible event in the video is <event> Helicopter </event> and the time range is <range> 3,6 </range>.

Figure 4. Two examples on AVE task. (a) A race car is driving in front of a building. (b) A scene of a helicopter flying.

sult on these two tasks.

3.3. Spatio-temporal reasoning

Tab. 2 shows detailed experiment results on MUSIC-AVQA [4] test set. It can be seen that our method outperforms all specialized models. Specifically, compared to recent TSPM [6], our model achieves significant overall performance improvements of 2.15% (78.94% *vs.* 76.79%). In all visual subtasks, including *Count* and *Localization*, our method acheives remarkable improvements of 5.22% (87.51% *vs.* 82.29%) and 9.02% (93.92% *vs.* 84.90%) respectively. In complex audio-visual question types, our model obtains the best overall performance (74.13%) and the performance in audio subtasks is also comparable. It is worth noting that our model generally achieves superior results in spatial localization, counting and temporal question types. This is mainly due to the cooperation of temporal and spatial localization tasks, which also proves that our method can achieve mutual cooperation among tasks. Moreover, our model performs pooly in the *Comp* question type, which mainly involves subjective questions such as comparing the melody, pitch and intensity of sounds. Differences among annotators can effect the accuracy of original labels. This also makes it difficult for our model to learn a unified standard for evaluation, leading to a disordered reasoning process and resulting in decreased performance.

3.4. Ablation results on LoRA head numbers

In order to explore the impact of different numbers of LoRA heads on model performance, we compare the experimental results of using three, four, and five LoRA heads respectively. Tab. 3 presents the corresponding experimental results. From the table, it can be seen that different tasks achieve the best results on different numbers of LoRA heads. The increase in the number of LoRA heads does not necessarily improve the model's performance. A possible reason is that when the number of LoRA heads is too large, exceeding the types of audiovisual data interactions, the each additional LoRA head may focus on the same or multiple aspects of data interaction. The former improves the performance of the corresponding tasks, while the latter causes the same ability to be distributed among multiple LoRA heads, making it difficult for model to acquire all the abilities to solve the corresponding tasks.

3.5. The visualized results on LoRA head numbers

Furthermore, we also perform a visualized analysis of the router weights for different numbers of LoRA heads.

(a) **Label:** Speech, Clapping
Converted label: The office workers are clapping for a man in a red polo shirt and cap. From the beginning to about 0:03 seconds someone is talking but it's inaudible. Clapping from the group begins around 0:00 and ends around 0:03. The man being clapped for begins speaking around 0:03 and continues to the end of the video. The clapping starts again briefly at the very end, around 0:09 - 0:10 seconds. Therefore: <audio>Speech,(0 3),(3 10)</audio> <audio>Clapping,(0 3),(9 10)</audio>

(b) **Label:** Speech, Baby_laughter
Converted Label: The video shows a baby sitting in a small inflatable pool placed on a table. The baby is visible from the beginning (0 seconds) to the end (10 seconds) of the video. We can hear the baby laughing and making baby noises almost throughout the entire video, with occasional pauses. An adult can be heard speaking to the baby from 0 to 1, 2 to 3 and 7 to 8 seconds. Therefore, the events are: <audio>Speech,(0 1),(2 3),(7 8)</audio> <audio>Baby_laughter,(0 1),(1 2),(3 6),(6 7),(8 10)</audio>.

Figure 5. Two examples of label transformation on AVVP task. (a) A scene of office workers clapping and someone talking. (b) A scene of a child playing and laughing in the water.

Table 1. The comparison results with specific models on AVS-Bench and Ref-AVS test set. S4, MS3 and AVSS are the subtasks of AVS-Bench. Seen, Unseen and Null are the subtasks of Ref-AVS test set.

Method	Visual Backbone	S4		MS3		AVSS		Seen		Unseen		Null
		miou	F-score	miou	F-score	miou	F-score	miou	F-score	miou	F-score	S(↓)
TPAVI [15]	ResNet-50	72.80	84.80	47.90	57.80	20.20	25.20	-	-	-	-	-
CATR [7]	ResNet-50	74.80	86.60	52.80	65.30	-	-	-	-	-	-	-
BAVS [9]	ResNet-50	78.00	85.30	50.20	62.40	24.70	29.60	-	-	-	-	-
iGAN [10]	Swin-T	61.60	77.80	42.90	54.40	-	-	-	-	-	-	-
LGVT [14]	Swin-T	74.90	87.30	40.70	59.30	-	-	-	-	-	-	-
TPAVI [15]	PVT-v2	78.70	87.90	54.00	64.50	**29.80**	**35.20**	-	-	-	-	-
AVSBench [15]	PVT-v2	78.70	87.90	54.00	23.20	-	-	0.51	32.36	0.55	0.21	0.21
AVSegFormer [2]	PVT-v2	**82.10**	**89.90**	**58.40**	**69.30**	24.90	29.30	33.47	0.47	36.05	0.50	0.17
GAVS [12]	PVT-v2	-	-	-	-	-	-	28.93	0.50	29.82	0.50	0.19
EEMC [13]	PVT-v2	-	-	-	-	-	-	34.20	0.51	**49.54**	**0.65**	**0.01**
Crab(Ours)	**ViT/L-14**	73.25	86.81	58.21	66.24	26.59	32.10	**40.54**	**0.58**	45.55	0.63	**0.01**

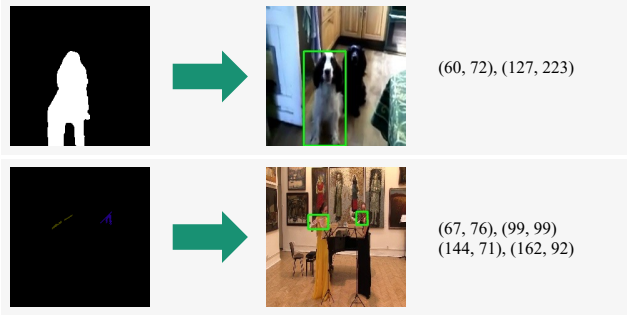


Figure 6. Two examples of converting segmentation mask to obtain bounding boxes on AVS-Bench dataset. (a) The example on S4 subset. (b) The example on AVSS subset.

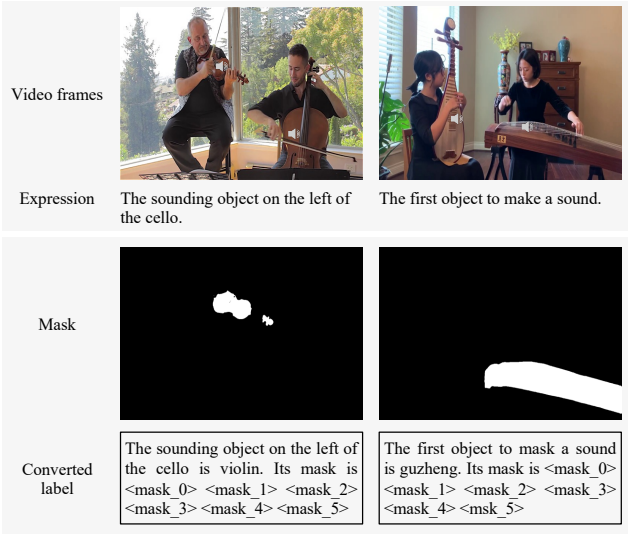


Figure 7. Two examples of label transformation on Ref-AVS task.

Fig. 10 and Fig. 12 demonstrate the router weights when using four heads and five heads. Similar to the analysis in the main paper, we can see that tasks of the same type form a cluster, indicating that their dependence on the same head is similar. Different types of tasks have different dependencies on these heads, indicating that different heads have different types of capabilities. Moreover, as discussed in Section 3.4, when the number of LoRA heads is too large, each head may focus on a specific aspect of data interaction, thus possessing a specific type of capability, such as the *head-B2* and *head-B3* in four heads, *head-B3* in five heads. It may also

(a)

Question: Which xylophone makes the sound first?
Answer: right
Converted answer: The video shows two xylophones and a white drum placed between the two xylophones. Two men are standing there, the man in gray is playing the left xylophone and the man in black is playing the right xylophone. The right xylophone sounds first at the beginning of the video. So the answer is <answer> right </answer>.

(b)

Question: What is the second instrument that comes in?
Answer: violin
Converted answer: In the video, a girl is playing the piano at the beginning. Then, a girl playing the violin joins in at [00:00:23]. So the second instrument is the violin. So the answer is <answer> violin </answer>.

Figure 8. Two examples of label transformation on AVQA task.

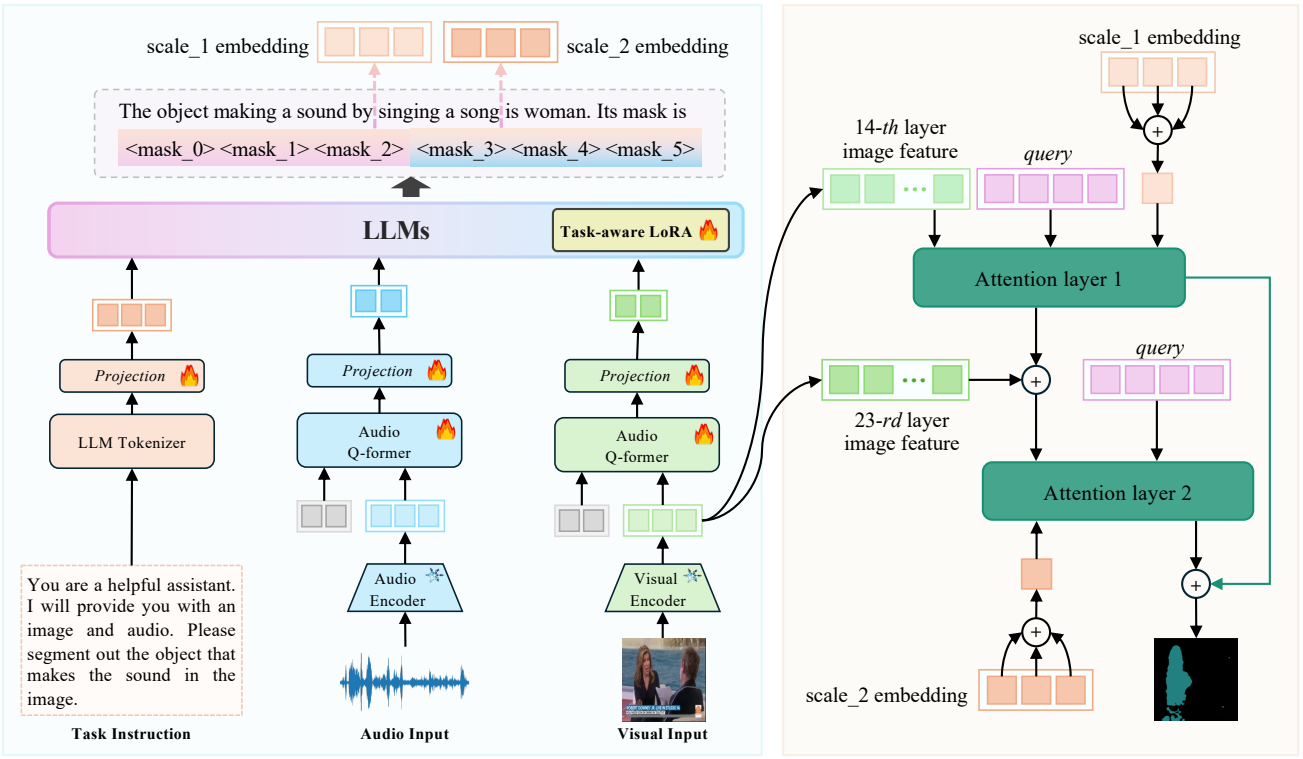


Figure 9. The overview of mask decoder.

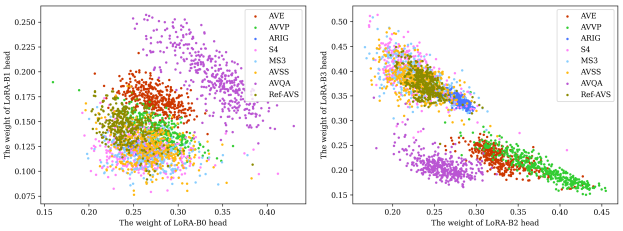


Figure 10. The router weights of four heads.

have multiple capabilities at the same time, such as *head-B0* in four heads, *head-B1*, *head-B2* and *head-B3* in five heads. Therefore, how to more precisely control each head to have specific capability may be a meaningful direction for future research.

3.6. Visualized results on all tasks

Fig. 13 and Fig. 14 present some visualized results on all tasks.

References

[1] Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. Octavius: Mitigating task interference in mllms via

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

Table 2. The comparison results with specific models on MUSIC-AVQA test set.

Method	Audio			Visual			Audio-Visual						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
ST-AVQA [4]	77.78	67.17	73.87	73.52	75.27	74.40	82.49	69.88	64.24	**64.67**	65.82	69.53	71.59
COCA [3]	79.35	**67.68**	75.42	75.10	75.43	75.23	**83.50**	66.63	69.72	64.12	65.57	69.96	72.33
PSTP-Net [5]	73.97	65.59	70.91	77.15	77.36	77.26	76.18	72.23	71.80	71.79	69.00	72.57	73.52
LAVISH [8]	82.09	65.56	75.97	78.98	81.43	80.22	81.71	75.51	66.13	63.77	67.96	71.26	74.46
TSPM [6]	84.07	64.65	**76.91**	82.29	84.90	83.61	82.19	76.21	71.85	**65.76**	**71.17**	73.51	76.79
Crab(Ours)	**85.55**	61.21	76.58	**87.51**	**93.92**	**90.73**	82.88	**81.26**	**71.95**	62.13	71.11	**74.13**	**78.94**

Table 3. The ablation results on task-aware LoRA. "three heads" means the head numbers of task-aware LoRA is three.

Method	AVE	ARIG		AVQA	S4		MS3		AVSS		Seen		Unseen		Null
	Acc	cIoU	AUC	Acc	mIoU	F-score	mIoU	F-score	mIoU	F-score	mIoU	F-score	mIoU	F-score	S↓
three heads	80.15	41.78	0.42	**78.94**	73.52	86.81	58.21	66.24	**26.59**	**32.10**	40.54	**0.58**	**45.55**	**0.63**	**0.01**
four heads	**80.25**	42.33	0.42	77.76	72.59	86.21	56.29	65.63	24.76	30.25	**42.17**	**0.58**	42.01	0.58	**0.01**
five heads	80.17	**43.59**	**0.44**	78.13	**73.78**	**86.88**	**59.97**	**68.43**	24.58	30.16	41.98	**0.58**	44.13	0.61	**0.01**

lora-moe. *arXiv preprint arXiv:2311.02684*, 2023.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1485–1494, 2023.

[8] 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.

[9] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *IEEE Transactions on Multimedia*, 2024.

[10] Yuxin Mao, Jing Zhang, Zhexiong Wan, Yuchao Dai, Aixuan Li, Yunqiu Lv, Xinyu Tian, Deng-Ping Fan, and Nick Barnes. Transformer transforms salient object detection and camouflaged object detection. *arXiv preprint arXiv:2104.10127*, 1 (2):5, 2021.

[11] 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.

[12] 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.

[13] 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.

[14] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. *Advances in Neural Information Processing Systems*, 34:15448–15463, 2021.

[15] 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.

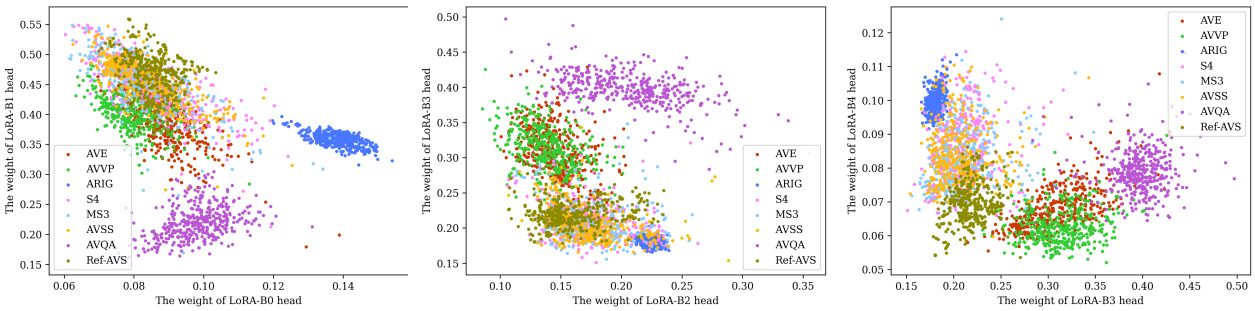


Figure 12. The router weights of five heads.

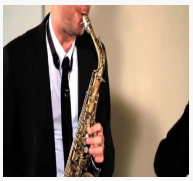
AVE

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>.

Figure 13. The visualized results on AVE, AVVP and AVQA tasks.

ARIG MS3 S4 AVSS Ref-AVS

Figure 14. The visualized results on ARIG, AVS and Ref-AVS 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**