Improving Sound Source Localization with Joint Slot Attention on Image and Audio

Supplementary Material

1. Model Size

Unlike previous studies, we use an additional module for joint slot attention, including auxiliary decoders. The joint slot attention module consists of 6.84M parameters and the image decoder and audio decoder each have 1.05M parameters. Compared to EZ-VSL [6], which uses only two ResNet-18 [4], our approach requires 39% more parameters during training, and 30% more parameters during inference because the decoders are not used for inference.

Method	# of parameters (M)
EZ-VSL [6]	22.87
Ours	31.82
Ours†	29.71

Table 1. **Model size.** † denotes the number of used parameters during the inference.

2. Experiments on Multi-Source Dataset

To more precisely evaluate the performance of the multi-source dataset, we trained our model on VGGSound-144k and evaluated in a zero-shot manner on AVSBench MS3. As shown in Table 2, our method showed strong performance on the multi-source dataset and effectively captured distinct objects when two target slots were used. However, the impact of increasing target slots was limited since the training dataset, VGGSound-144k, is a single-source dataset.

3. Additional Ablation Studies

In this section, we conduct ablation studies on the slot attention iterations N, predicted false negatives k, masking ratio, the number of each slot. All experiments are trained on VGGSound-144k [2] and tested on VGG-SS [2].

Slot Attention Iteration. We analyze the effect of N by varying slot attention iterations. Table 3 shows that 5 iterations achieve the best performance for both sound source localization and cross-modal retrieval. Similar to the findings in the original slot attention [5], performance declines when the number of iterations is too low or too high.

Predicted False Negatives. We measure the impact of k-reciprocal nearest neighbors on false negative mitigation by varying k. Table 4 shows that an appropriate k improves performance by excluding false negatives. However, a large k may also exclude true negatives, reducing the number of samples available for contrastive learning and degrading performance. We set $k = 20$ for the best results in sound source localization.

Masking Ratio. During training, we randomly replace 10% of the input features with learnable mask tokens to prevent overfitting and improve performance. We investigate the impact of this approach by varying the masking ratio. Table 5 shows that using a small ratio of learnable masks yields better performance compared to not using masks at all. However, excessively replacing input features leads to a loss of information, resulting in performance degradation.

The Number of Each Slot. We analyze the effect of the slots by changing the number of target slot and off-target slot. Table 6 shows that the performance drops as the number of slots increases. This is mainly because increasing the number of slots too much intensifies their competition in input decomposition. For example, with four slots as shown

Testset	Method	mIoU	F-Score
MS3	LVS [3]	18.54	17.4
	EZ-VSL [6]	21.36	21.6
	FNAC [9]	21.98	22.5
	SLAVC [7]	24.37	25.56
	Ours$_{(1,1)}$	<u>24.45</u>	**36.87**
	Ours$_{(2,1)}$	**24.73**	<u>36.56</u>

Table 2. **Results on MS3.** Subscripts indicate the numbers of target and off-target slots, respectively.

N	cIoU	AUC	Audio → Image			Image → Audio		
			R@1	R@5	R@10	R@1	R@5	R@10
1	16.36	27.81	1.12	5.78	9.89	3.47	11.65	17.82
3	39.59	40.78	16.17	36.27	47.91	19.45	41.55	53.88
5	**40.71**	**41.62**	**30.61**	**57.87**	**69.37**	**31.70**	**58.43**	**69.81**
7	39.74	41.21	23.03	46.88	58.14	24.97	49.17	61.25
10	34.24	38.60	7.12	19.58	28.44	7.43	24.37	35.69

Table 3. **Ablation studies of the number of iterations N.** The gray row indicates the settings used in the main paper.

k	cIoU	AUC	Audio → Image			Image → Audio		
			R@1	R@5	R@10	R@1	R@5	R@10
10	40.29	41.16	**33.08**	**60.78**	71.42	33.00	**60.06**	**71.73**
20	**40.71**	**41.62**	30.61	57.87	69.37	31.70	58.43	69.81
30	40.15	41.14	31.87	60.06	**72.04**	**33.29**	59.78	71.23
50	40.44	41.56	28.13	55.58	67.29	29.47	56.65	68.77

Table 4. **Ablation studies of k-reciprocal nearest neighbors.** The gray row indicates the settings used in the main paper.

Ratio	cIoU	AUC	Audio → Image			Image → Audio		
			R@1	R@5	R@10	R@1	R@5	R@10
0.0	37.46	40.26	8.90	24.72	34.45	10.08	29.35	42.52
0.05	40.07	40.96	17.39	37.53	48.64	20.24	44.65	56.34
0.1	**40.71**	**41.62**	**30.61**	**57.87**	**69.37**	**31.70**	**58.43**	**69.81**
0.2	33.50	38.34	9.25	23.23	32.92	9.52	27.82	39.47
0.3	31.72	37.48	9.13	22.55	31.72	9.13	27.05	39.20
0.5	16.56	29.22	12.54	30.46	40.29	15.08	36.51	48.31

Table 5. **Ablation studies of the masking ratio.** The gray row indicates the settings used in the main paper.

in Fig. 2, a target area is often fragmented, with no target slot clearly capturing the target, and target or off-target slots tend to intrude into background or target area, respectively.

4. Failure Cases

Fig. 1 presents failure cases from VGG-SS. Due to the limited resolution of attention, the model sometimes has trouble capturing small objects. Also, it struggles with strong co-occurrence bias. For example, as instruments and humans frequently co-occur in training data, it sometimes identifies both simultaneously when an instrument sound is given.

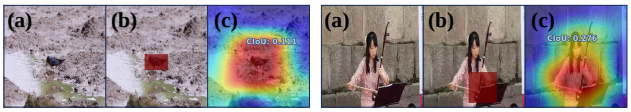

Figure 1. **Failure cases.** (a) Input (b) Ground-Truth (c) Prediction

5. More Qualitative Results

In this section, we present additional qualitative results. Fig. 3 shows the intra-modal attention result, showing the target and off-target slots effectively decompose the input by attending to certain regions. Additionally, in the case of audio, we qualitatively confirmed that the target slot focuses on time intervals related to bird sounds in the spectrogram. Also, we present additional qualitative results on SoundNet-Flickr-Test [1] and VGG-SS [2] in Fig. 4, and additional cross-modal retrieval on VGG-SS in Fig. 5. Furthermore, we visualize the cross-modal attention between the audio target slot and image features with and without $\mathcal{L}_{\text{match}}$ in Fig. 6 to show the effect of $\mathcal{L}_{\text{match}}$. Fig. 6 demonstrates that $\mathcal{L}_{\text{match}}$ encourages the attention map to focus on the sound source. Finally, Fig. 7 presents some examples of predicted false negatives within a batch. Fig. 7 demonstrates that k-reciprocal nearest neighbors are likely to belong to false negatives.

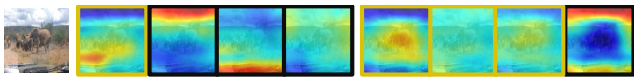

Figure 2. The yellow boundaries indicate target slot attention, while the black boundaries indicate off-target slot attention.

Target	Off-Target	VGG-SS	
		cIoU	AUC
1	1	**40.71**	**41.62**
1	2	38.91	40.69
1	3	34.52	38.66
2	1	40.36	41.17
3	1	37.50	40.11

Table 6. **Ablation studies of the number of each slots.** The gray row indicates the settings used in the main paper.

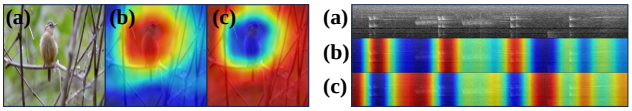

Figure 3. **Intra-modal attention.** (a) Input image-audio pair (b) Target slot attention (c) Off-target slot attention

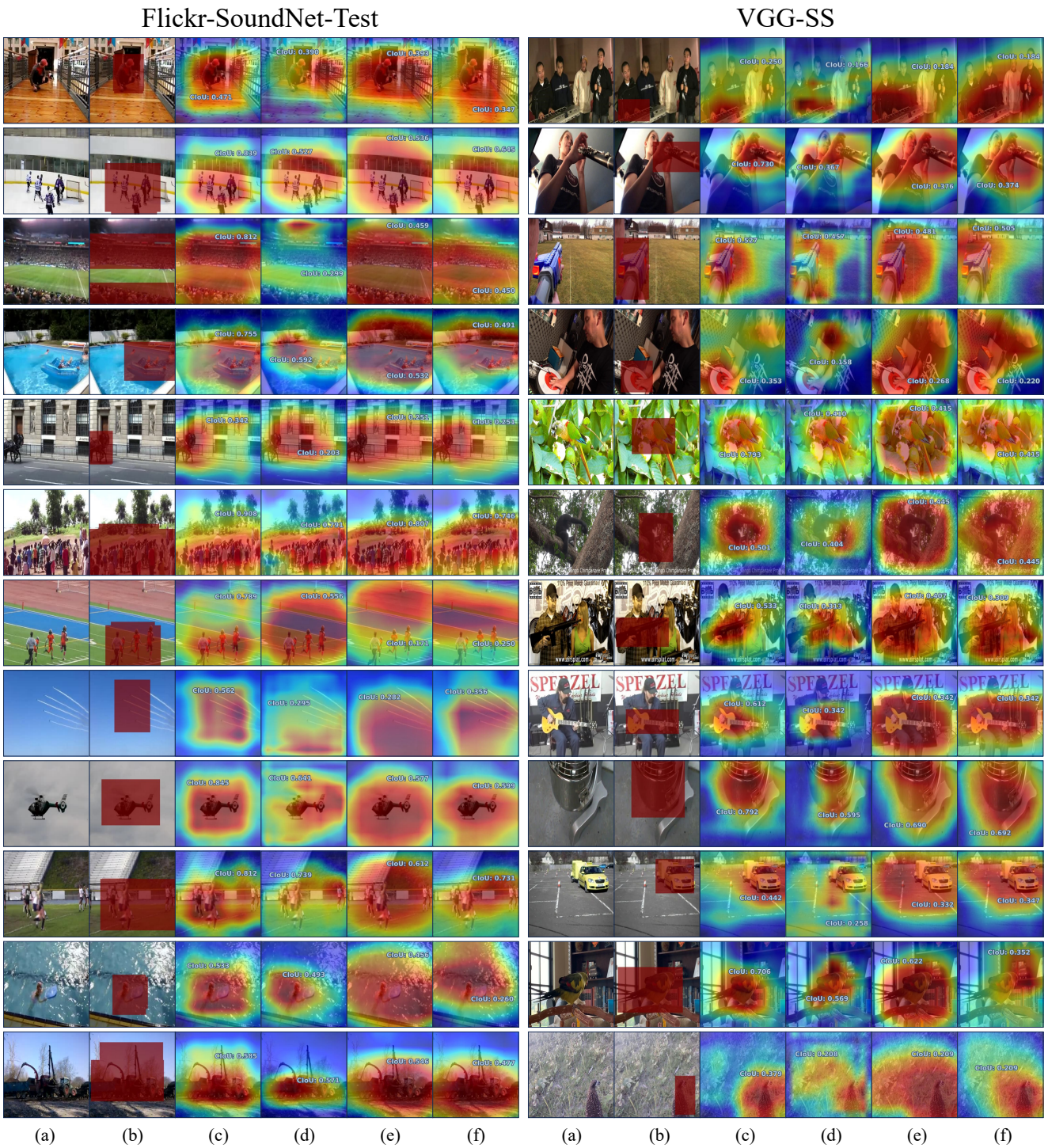

Figure 4. **Additional sound localization results on Flickr-SoundNet-Test [1] and VGG-SS [2].** (a) Input image. (b) Ground-Truth. (c) Ours. (d) Alignment [8]. (e) FNAC [9]. (f) EZ-VSL [6]. The qualitative results are obtained by the model trained on Flickr-144k and the model trained on VGGSound-144k, respectively. Note that all visualizations are obtained without refinement.

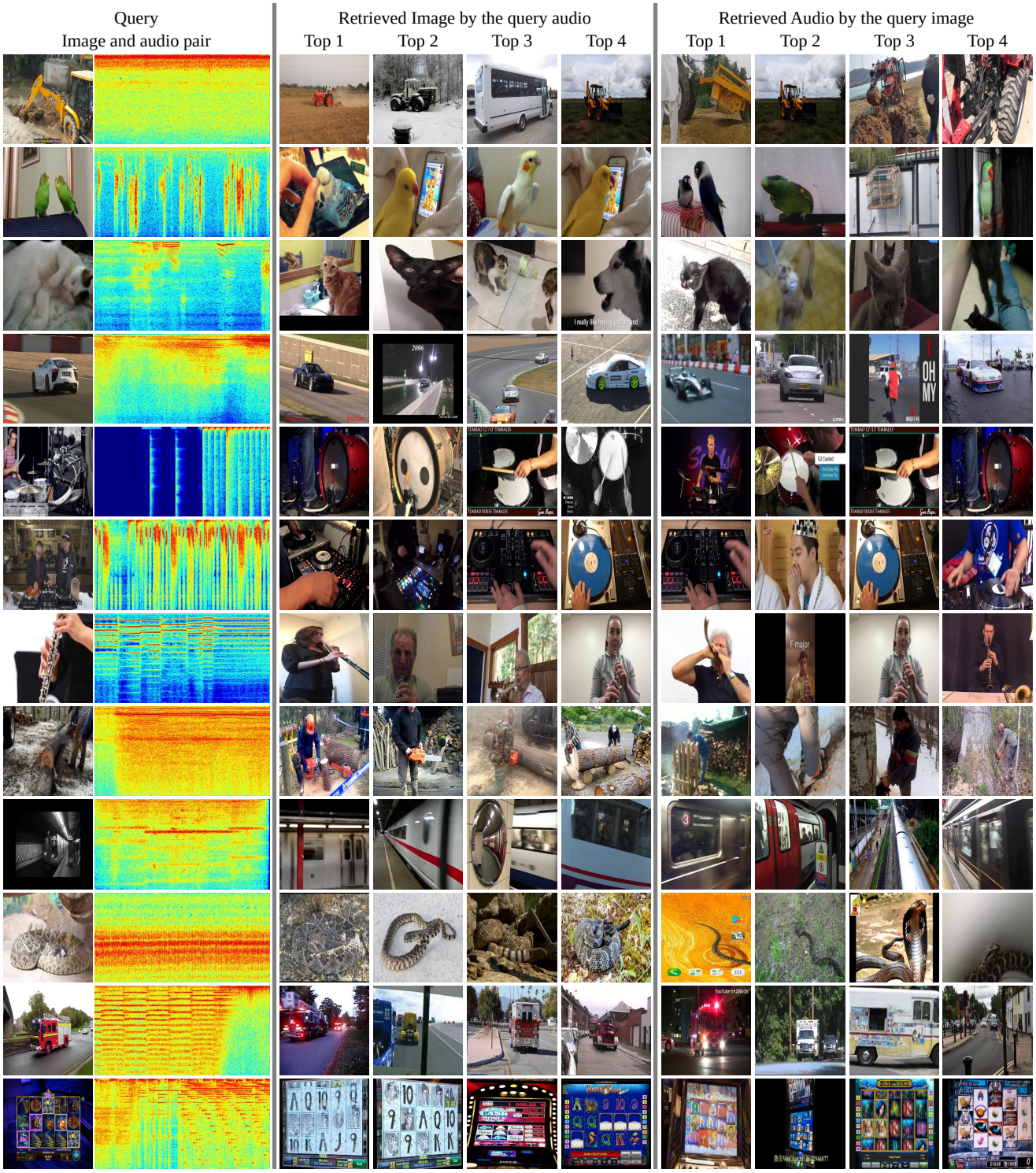

Figure 5. **Additional Qualitative results of cross-modal retrieval on VGG-SS [2].** Since the audio spectrograms are difficult to understand, we visualize its corresponding image instead.

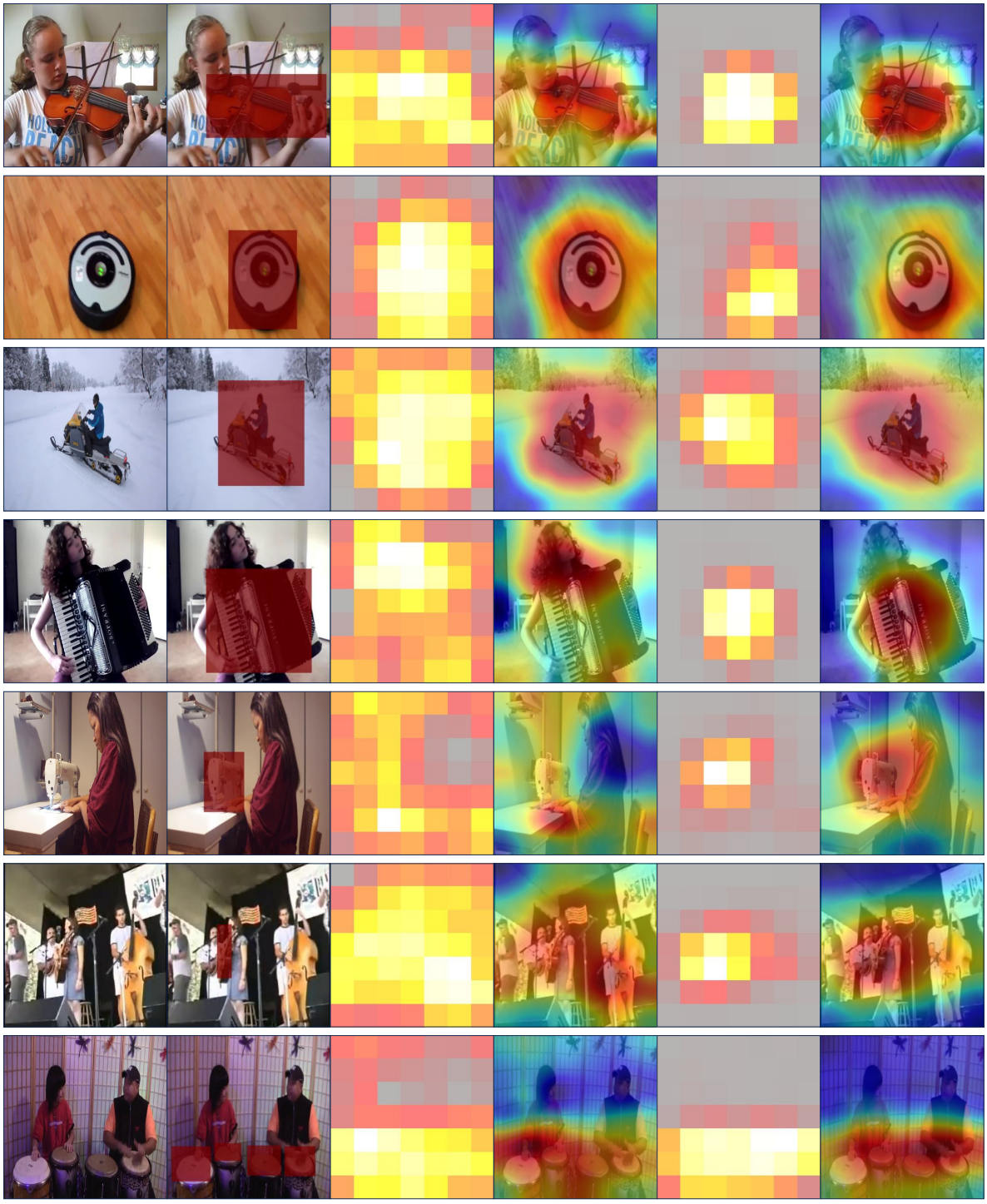

Figure 6. **Additional qualitative results to show the impact of $\mathcal{L}_{\text{match}}$ on VGG-SS [2].** (a) Input image. (b) Ground-Truth. (c) Attention map of 7×7 size without cross-modal attention matching. (d) Attention map of 224×224 size without cross-modal attention matching. (e) Attention map of 7×7 size with cross-modal attention matching. (f) Attention map of 224×224 size with cross-modal attention matching.

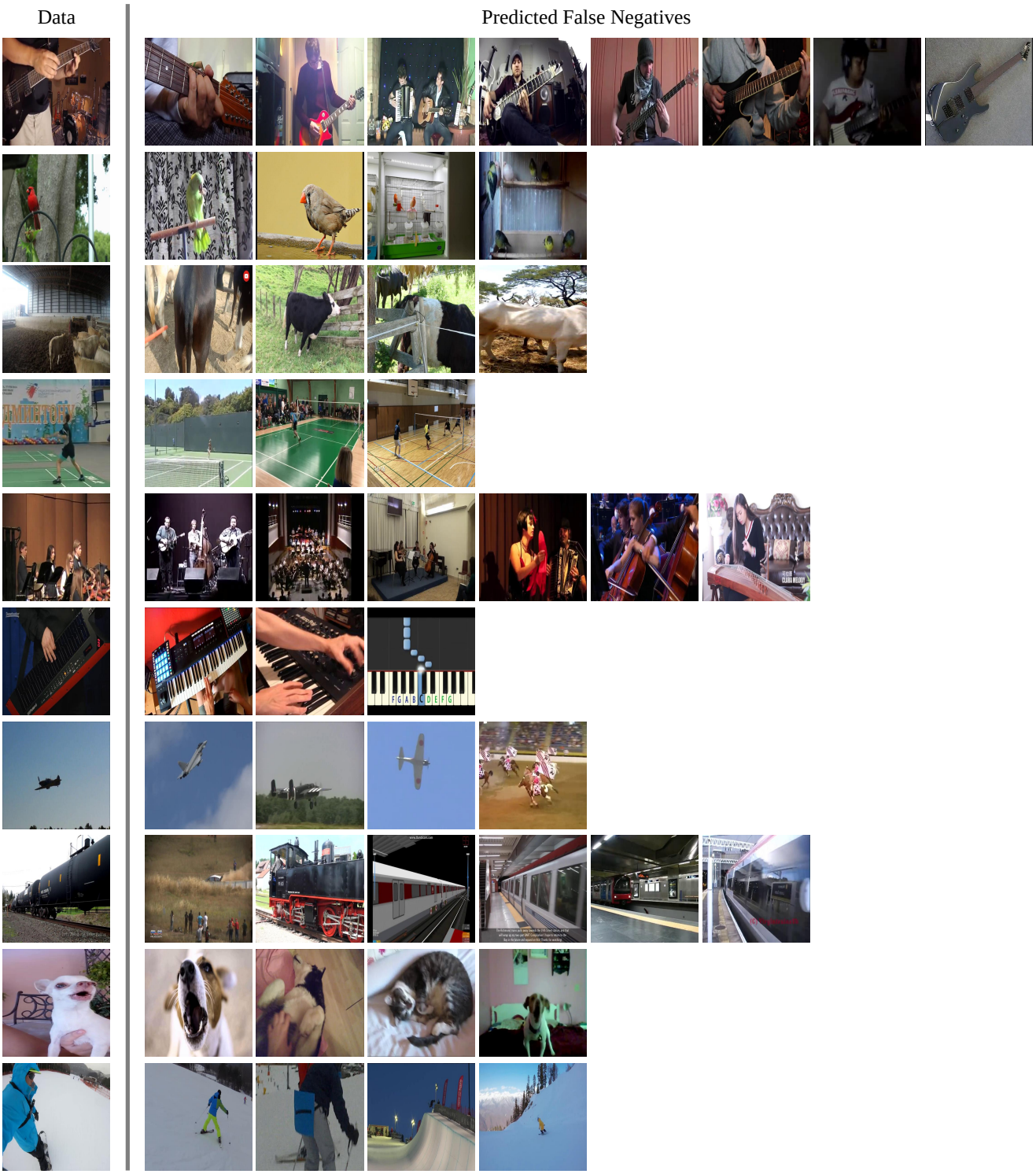

Figure 7. **Qualitative results of predicted false negative samples of VGGSound [2].** Samples in the same row are predicted as false negatives by using k-reciprocal nearest neighbors within a batch. They have both similar image and audio target slot representations, so they are not used as negative pairs for contrastive learning.

References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Proc. Neural Information Processing Systems (NeurIPS)*, 29, 2016. 2, 3

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 2, 3, 4, 5, 6

[3] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[5] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Proc. Neural Information Processing Systems (NeurIPS)*, 33:11525–11538, 2020. 1

[6] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 1, 3

[7] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2022. 1

[8] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[9] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3