CLAPSep: Leveraging Contrastive Pre-trained Model for Multi-Modal Query-Conditioned Target Sound Extraction

Hao Ma, Zhiyuan Peng, Xu Li, Mingjie Shao, Xixin Wu, and Ju Liu

TASLP 2024

IEEE/ACM Transactions on Audio, Speech, and Language Processing

Introduction

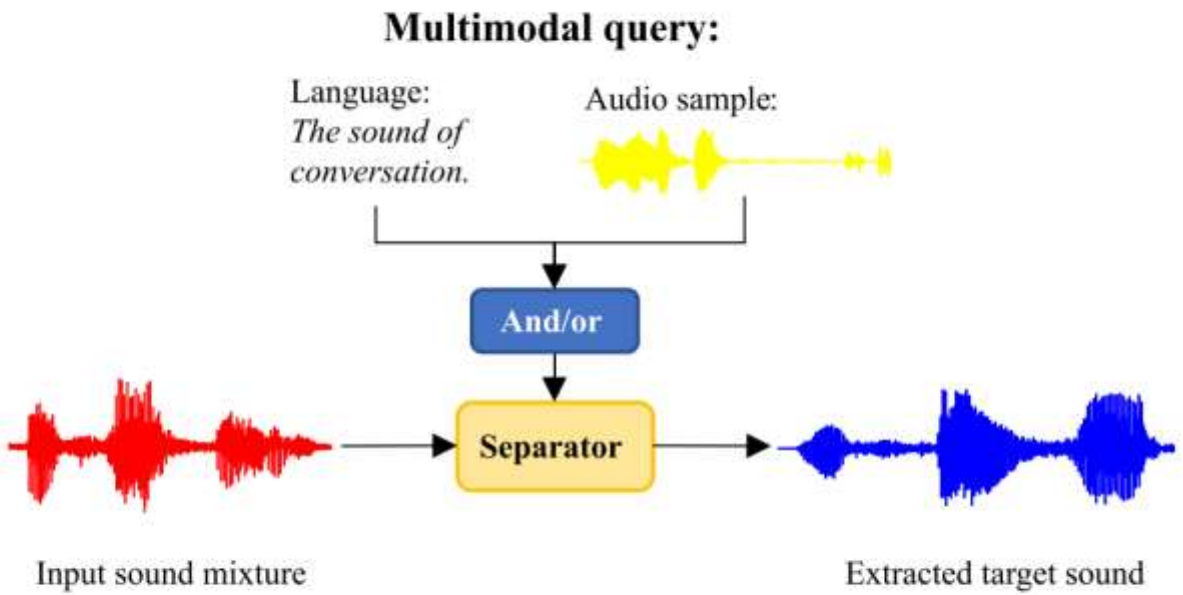


Fig. 1. Illustration of query-conditioned target sound extraction.

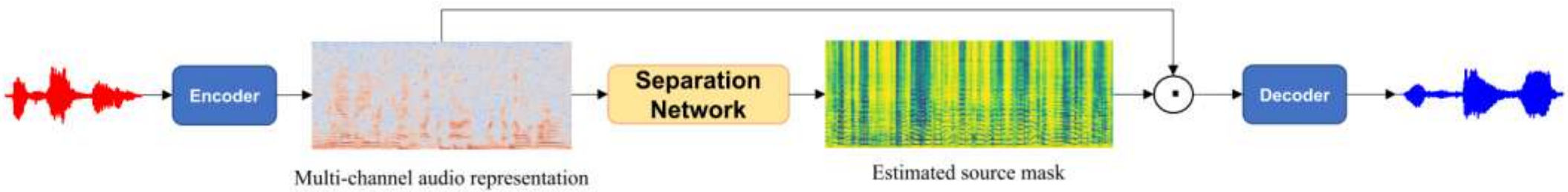


Fig. 2. Illustration of mask-based sound separation pipeline, where $\odot$ denotes element-wise production.

Method

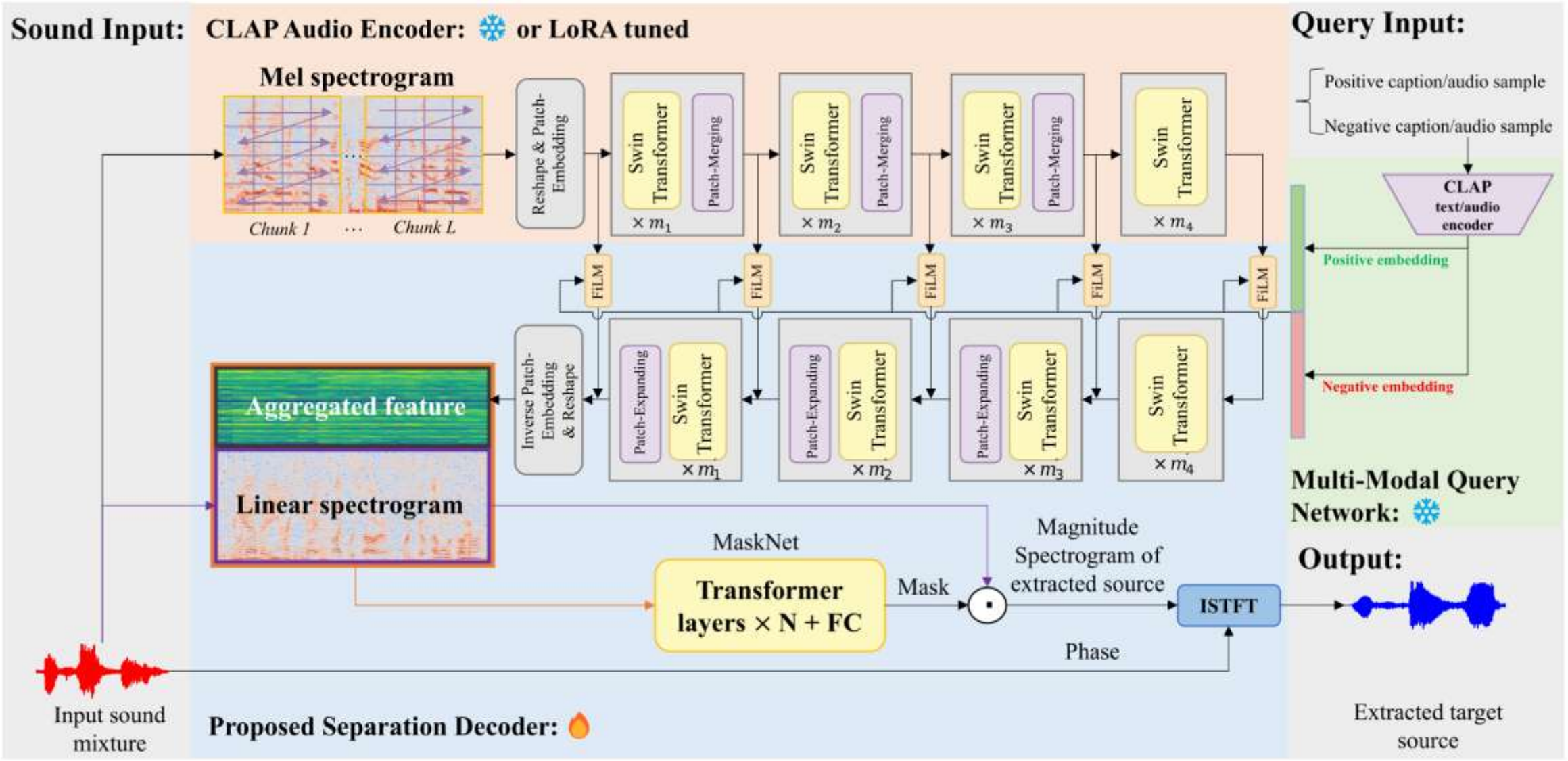


Fig. 3. Overview of proposed CLAPSep model.

Method

训练数据参考+与基线方法对比

TABLE I
TRAINING DATA COUNTING.

Methods	Training corpus	# clips	Hours
AudioSep [21]	AudioSet+AudioCaps+others	2 342 568	14 100
USS [9]	AudioSet	2 063 839	5 800
Waveformer [14]	FSDKaggle2018	9 500	18
LASS [20]	AudioCaps (subset)	6 244	17
Ours	AudioCaps	49 274	145

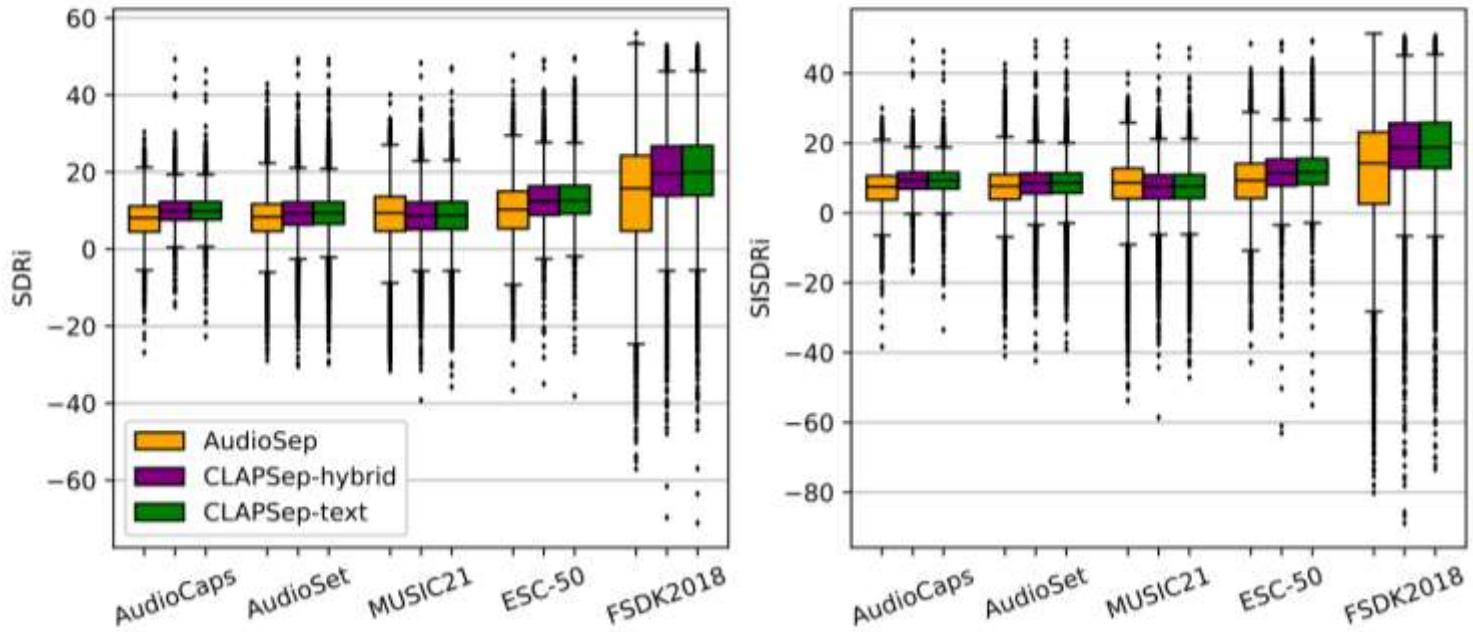


Fig. 4. Illustration of SDRi and SISIDRi distributions w.r.t. different language-queried TSE models on 5 evaluation datasets.

Experiments

仅文本查询：指标对比 和 音频-文本对齐情况

TABLE II

LANGUAGE-QUERIED TARGET SOUND EXTRACTION PERFORMANCE EVALUATION WITH SDRS.

Methods	Query Modality	Query Polarity	AudioCaps		AudioSet		MUSIC21		ESC-50		FSDKaggle2018	
			SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi
AudioSep [21]	Text	P	7.75±5.59	7.04±5.72	8.02±6.23	7.26±6.44	**8.73±7.71**	**7.84±8.09**	10.33±7.61	9.20±7.97	13.90±15.44	11.57±17.72
AudioSep†	Text	P	6.63±5.46	5.55±5.77	3.81±6.64	2.30±7.29	0.98±5.48	0.15±6.04	7.66±7.92	5.81±8.78	7.59±11.45	5.45±12.88
LASS [20]	Text	P	0.33±7.39	-1.11±8.18	-2.57±7.86	-4.41±8.60	-6.63±7.42	-9.83±7.91	-0.48±9.97	-2.60±11.31	-3.89±14.05	-9.70±16.76
LASS†	Text	P	6.75±5.59	6.05±5.86	3.12±7.61	2.02±8.29	-1.70±8.14	-3.62±9.06	7.49±9.06	6.07±10.18	6.55±16.48	3.27±19.35
Waveformer [14]	Label	P	-	-	-	-	-	-	-	-	7.77±11.33	5.68±12.11
CLAPSep-hybrid	Text	P	9.51±5.19	8.81±5.35	7.86±7.18	6.88±7.63	3.62±9.75	1.88±10.97	10.56±9.09	9.23±10.04	16.06±16.89	14.03±19.76
		N	9.55±5.07	8.85±5.19	7.90±7.11	6.99±7.53	4.00±9.25	2.32±10.20	10.14±9.83	8.70±11.13	15.86±17.42	13.69±20.59
		P+N	10.05±4.41	**9.40±4.41**	9.15±5.71	8.31±5.86	8.40±6.21	7.23±6.36	12.81±6.42	11.74±6.68	20.01±12.48	18.75±13.64
CLAPSep-text	Text	P	9.64±5.09	8.92±5.27	8.02±7.17	7.05±7.60	5.34±9.13	3.78±9.89	12.23±7.52	11.14±8.01	16.92±15.83	15.14±18.25
		N	9.65±5.03	8.94±5.17	7.98±7.21	7.05±7.64	6.24±8.12	4.99±8.70	12.19±7.41	11.12±7.97	16.42±16.88	14.27±19.94
		P+N	**10.08±4.42**	9.40±4.45	**9.29±5.61**	**8.44±5.75**	8.32±6.56	7.10±6.71	**13.09±6.22**	**12.10±6.37**	**20.17±12.43**	**18.91±13.38**

TABLE III

LANGUAGE-QUERIED TARGET SOUND EXTRACTION PERFORMANCE EVALUATION WITH CLAPSCORE.

Methods	Query Modality	Query Polarity	AudioCaps		AudioSet		MUSIC21		ESC-50		FSDKaggle2018	
			CLAPScore	ΔCLAPScore	CLAPScore	ΔCLAPScore	CLAPScore	ΔCLAPScore	CLAPScore	ΔCLAPScore	CLAPScore	ΔCLAPScore
AudioSep [21]	Text	P	0.365±0.118	0.328±0.194	**0.338±0.120**	0.274±0.188	**0.232±0.110**	**0.186±0.167**	**0.317±0.123**	0.281±0.184	0.202±0.126	0.167±0.185
CLAPSep	Text	P	**0.369±0.115**	0.353±0.175	0.329±0.120	0.254±0.189	0.206±0.138	0.081±0.230	0.312±0.134	0.265±0.219	0.205±0.130	0.192±0.187
		N	0.347±0.125	0.341±0.180	0.302±0.132	0.247±0.193	0.155±0.130	0.118±0.160	0.275±0.150	0.269±0.193	0.190±0.135	0.189±0.187
		P+N	0.367±0.116	**0.365±0.167**	0.330±0.120	**0.281±0.172**	0.218±0.126	0.180±0.165	0.313±0.131	**0.313±0.159**	**0.212±0.124**	**0.224±0.155**

CLAPScore：提取的音频与查询之间的语义相似度（越高越相关）
ΔCLAPScore：对比混音输入的提升幅度

Experiments

多模态查询：与其他方法的对比

TABLE IV

MULTI-MODAL CUES QUERIED TARGET SOUND EXTRACTION PERFORMANCE EVALUATION AND SOTA COMPARISON.

Methods	Query Modality	Shots	Query Polarity	MUSIC21		ESC-50		FSDKaggle2018	
				SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi
USS-ResUNet30 [9]	Audio	1	P	6.96±7.44	6.17±8.01	8.17±7.67	7.08±8.05	9.99±13.04	8.00±15.03
		5		8.06±6.56	7.38±6.84	9.29±6.94	8.38±7.06	12.10±11.15	11.04±11.45
		10		8.32±6.38	7.69±6.61	9.51±6.69	8.68±6.75	12.02±11.16	11.03±11.43
CLAPSep-text	Audio	10	P+N	5.39±5.45	4.36±5.47	6.93±6.57	6.16±6.81	7.79±10.49	6.79±11.09
	Audio+Text	10	P+N	8.36±5.67	7.22±5.69	12.12±6.13	11.17±6.26	19.64±11.56	18.43±12.27
CLAPSep-hybrid	Audio	1	P	6.34±7.95	5.02±8.51	12.08±7.57	10.88±7.99	15.41±17.52	13.01±20.86
			N	6.33±8.23	4.93±8.76	11.76±7.75	10.80±8.29	15.21±17.89	13.23±20.60
			P+N	8.41±6.34	7.21±6.50	12.89±6.41	11.94±6.53	19.04±13.53	17.64±14.93
		5	P	6.75±7.91	5.41±8.43	12.72±6.67	11.65±6.83	17.77±15.18	15.84±17.70
			N	6.77±8.00	5.42±8.53	12.49±6.78	11.61±6.97	17.56±15.36	16.01±17.37
			P+N	9.07±5.88	7.86±6.00	13.26±6.10	12.34±6.14	20.12±12.12	18.91±13.01
		10	P	6.98±7.72	5.67±8.19	12.79±6.63	11.73±6.78	17.97±14.89	16.15±17.12
			N	7.05±7.82	5.71±8.29	12.64±6.64	11.75±6.81	17.83±15.06	16.33±16.93
			P+N	9.35±5.59	8.16±5.65	**13.29±6.09**	**12.37±6.13**	20.25±11.88	19.07±12.71
	Audio+Text	10	P	6.91±7.92	5.54±8.56	12.46±7.16	11.30±7.47	18.87±14.02	17.28±15.80
			N	7.20±7.72	5.86±8.23	12.28±7.18	11.29±7.66	18.95±13.95	17.54±15.67
			P+N	**9.47±5.53**	**8.26±5.62**	13.21±6.16	12.24±6.23	**21.11±11.22**	**20.01±11.84**

Experiments

训练集中有无该类：

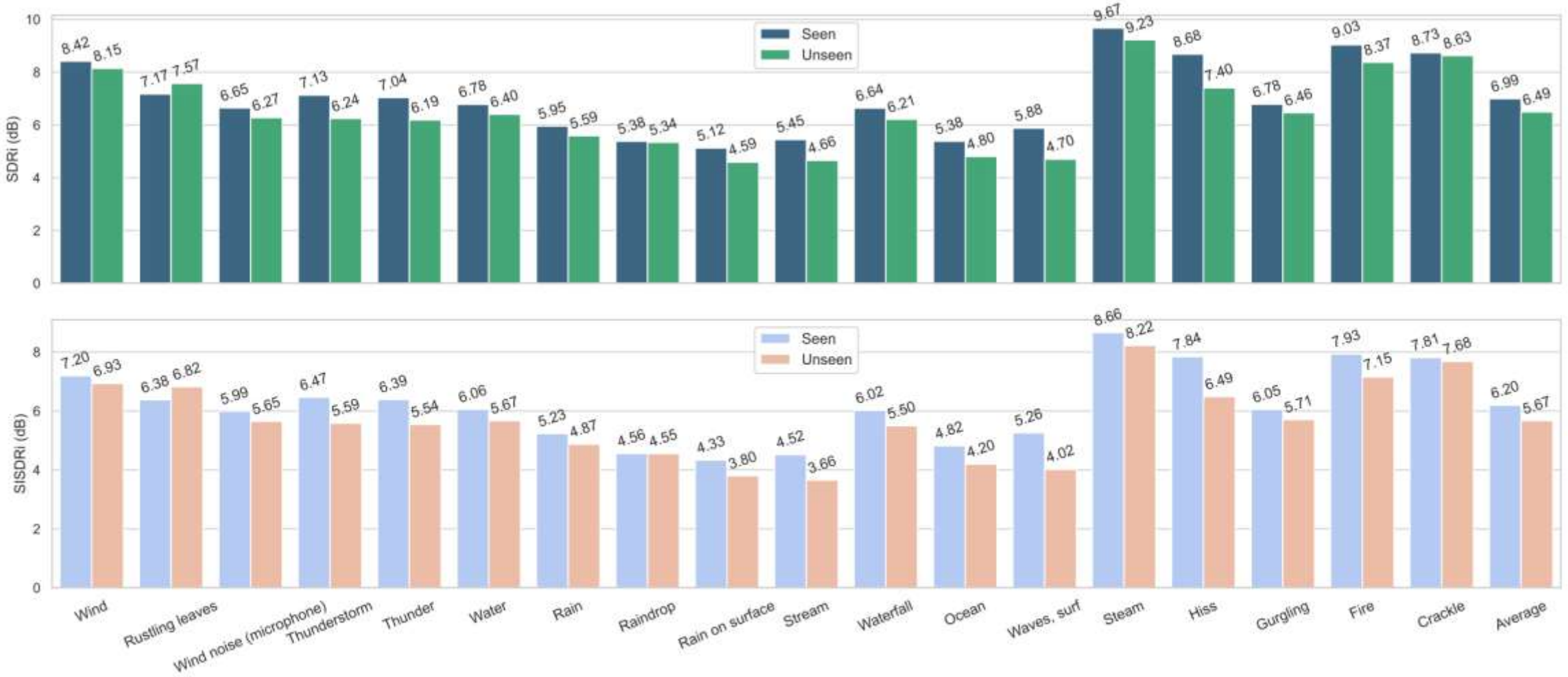


Fig. 6. Zero-shot generalizability evaluation.

Experiments

多模态查询：与其他方法的对比

TABLE V
MODEL PERFORMANCE VARYING WITH DIFFERENT NUMBERS OF SOUND SOURCES.

Methods	2-sources separation		3-sources separation				4-sources separation					
	1-target		1-target		2-target		1-target		2-taget		3-target	
	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi
AudioSep [21]	8.60±8.86	7.38±9.48	4.90±6.13	4.20±6.42	5.30±5.38	2.96±5.54	3.52±5.21	2.81±5.76	2.37±4.26	0.65±4.32	4.20±4.51	2.03±4.27
CLAPSep-P	10.03±11.27	8.70±12.64	6.48±7.82	5.73±8.50	6.35±7.36	4.88±8.10	5.94±7.58	4.85±8.29	2.91±5.54	1.37±5.98	5.34±6.70	3.79±7.09
CLAPSep-PN	**12.67±7.80**	**11.79±8.23**	**8.96±6.13**	**8.09±6.17**	**8.90±5.78**	**7.82±6.32**	**7.73±5.79**	**6.80±6.00**	**5.20±4.63**	**4.01±4.95**	**8.00±4.87**	**6.73±5.20**

TABLE VI
MODEL PERFORMANCE WITH INPUT SDR VARYING FROM -10 TO 10 DB.

Methods	Input SDR=-10dB		Input SDR=–5dB		Input SDR=0dB		Input SDR=5dB		Input SDR=10dB	
	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi
AudioSep [21]	10.55±7.24	9.86±7.18	9.54±6.29	8.82±6.30	7.75±5.59	7.04±5.72	5.60±5.21	4.86±5.52	2.87±5.34	1.94±5.93
CLAPSep	3.09±3.38	2.95±3.51	6.18±4.03	6.02±4.02	**10.05±4.41**	**9.40±4.41**	4.37±3.76	2.19±3.85	-1.62±3.40	-4.53±3.69
CLAPSep†	**13.03±6.23**	**12.11±6.20**	**11.68±5.15**	**10.82±5.16**	9.65±4.59	8.91±4.67	**7.40±4.33**	**6.73±4.52**	**5.07±4.29**	**4.31±4.76**

模型过拟合在「嘈杂环境」分离任务，在干净环境下反而无法发挥作用！！！

Experiments

TABLE VII
ABLATION STUDY OF LANGUAGE-QUERIED TSE PERFORMANCE.

Methods	AudioCaps		AudioSet		MUSIC21		ESC-50		FSDKaggle2018	
	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi	SDRi	SISDRi
CLAPSep (best)	**10.08±4.42**	**9.40±4.45**	**9.29±5.61**	**8.44±5.75**	**8.32±6.56**	**7.10±6.17**	**13.09±6.22**	**12.10±6.37**	**20.17±12.43**	**18.91±13.38**
w/o CLAP audio encoder	7.49±5.37	6.62±5.48	5.50±6.83	4.40±7.06	4.60±7.26	3.29±7.50	8.10±9.51	6.88±10.16	11.32±15.16	9.60±16.68
w/o pre-trained weights	9.33±4.71	8.58±4.77	7.78±6.26	6.79±6.46	6.20±7.58	4.73±8.00	11.40±7.36	10.21±7.96	15.10±15.22	13.48±17.40
frozen weights (w/o LoRA)	9.86±4.47	9.17±4.48	8.80±5.79	7.93±5.95	7.48±6.54	6.23±6.81	12.54±6.54	11.46±6.82	19.59±12.41	18.33±13.32
replace text encoder to CLIP	9.96±4.40	9.28±4.40	8.33±6.82	7.40±7.19	4.27±9.85	2.79±10.86	12.95±6.34	12.05±6.48	16.14±17.08	14.06±20.47

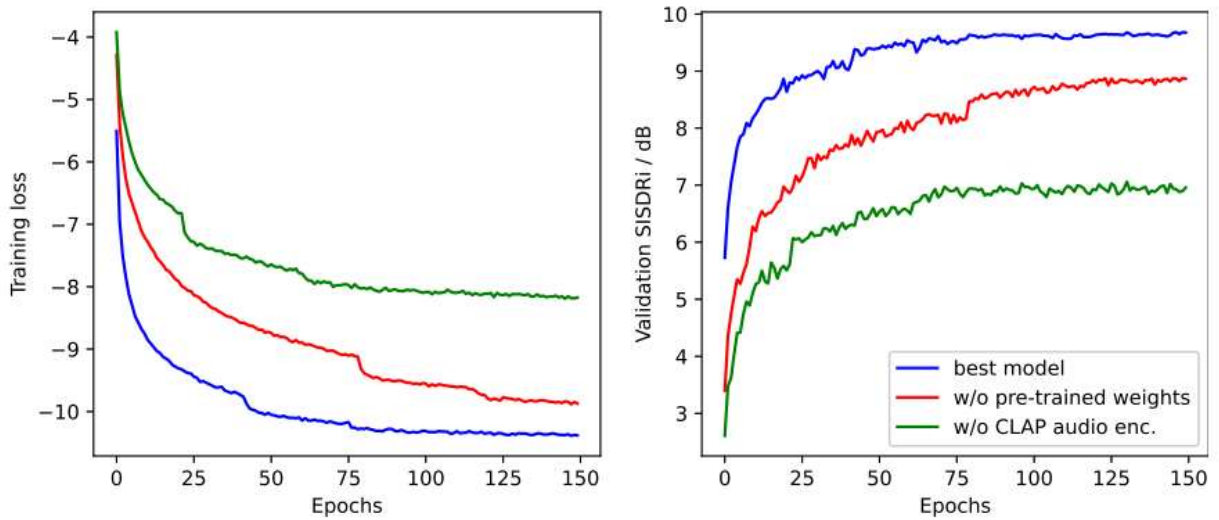


Fig. 7. Illustration of training loss and validation SISDRi.

Experiments

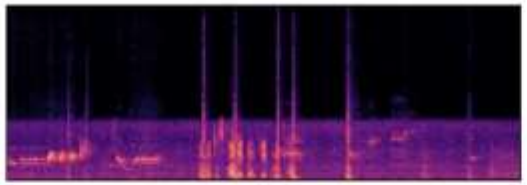
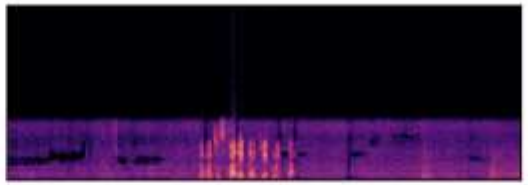
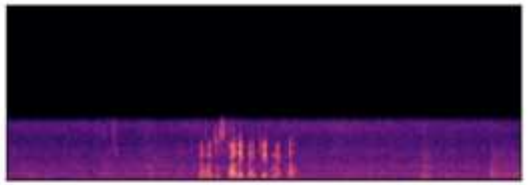
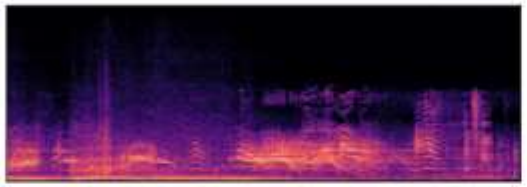
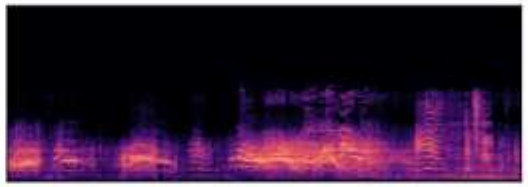
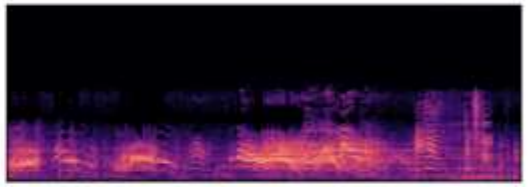
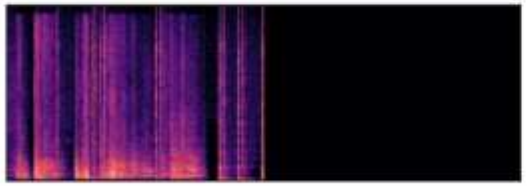
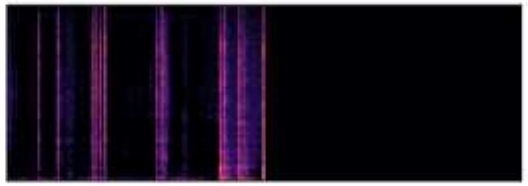
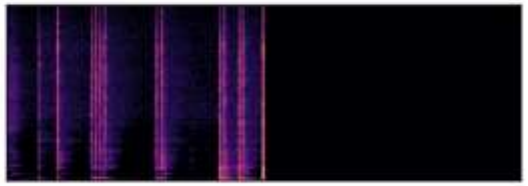
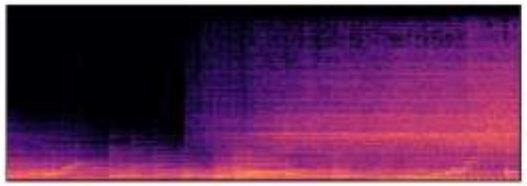
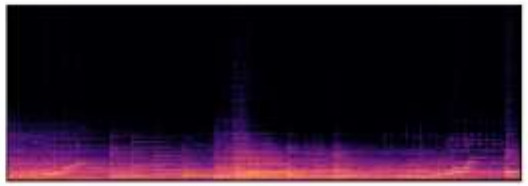
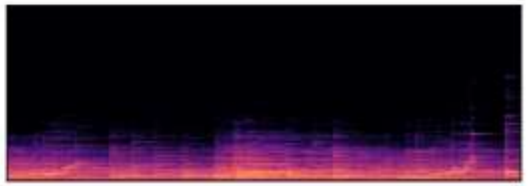
Query	Mixture	Extracted	Target
Positive text: *A man talking as insects buzz by.* **Negative text:** *Plastic crackling as a bird is singing and chirping.*			
Positive text: *Several cats meowing followed by a man singing.*			
Negative text: *A vehicle engine revving then powering down.*			
Positive text: *The sound of **Mouse click.*** **Negative text:** *The sound of **Insects.***			
Positive audio:  *(Piano)* **Negative audio:**  *(Drum)*			

Fig. 9. Visualization of spectrograms depicting sound mixtures, separated sources, and ground truth targets.

Thank you!