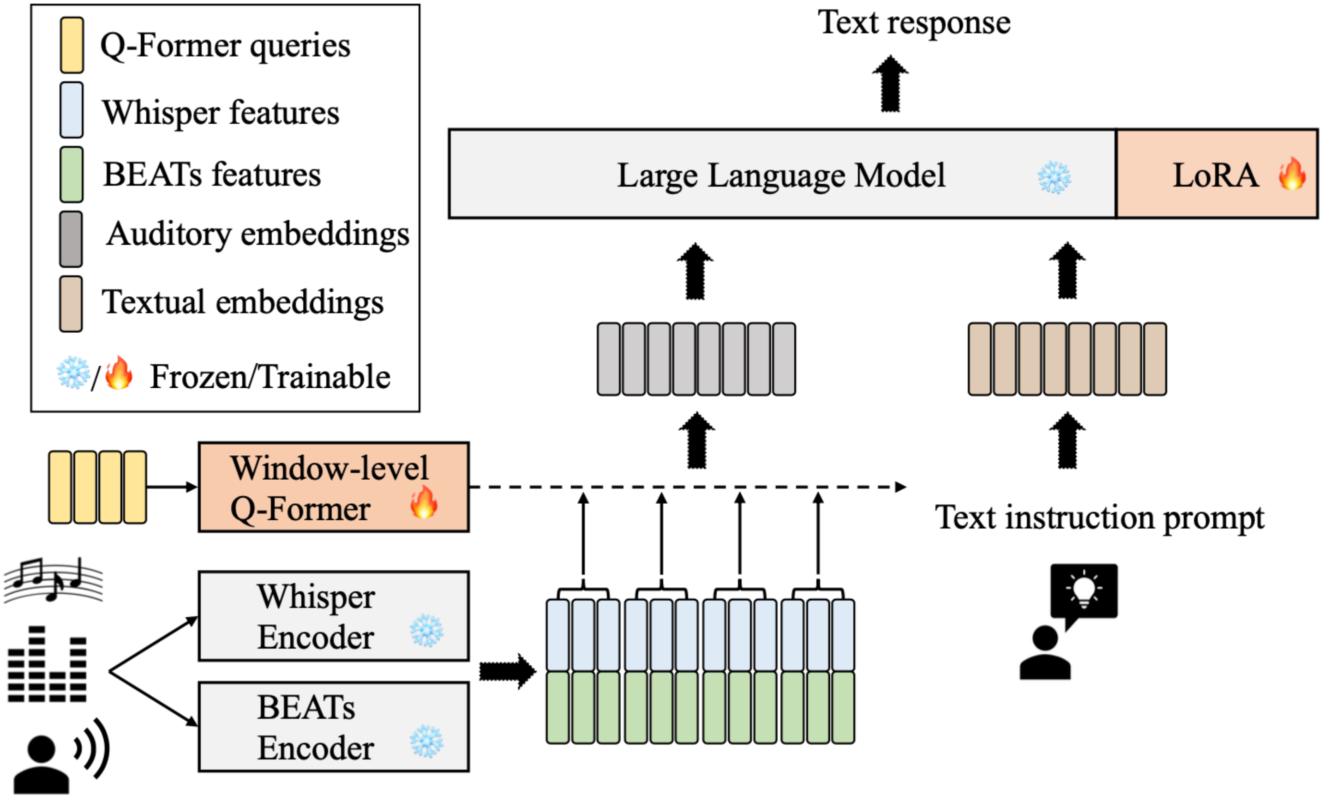
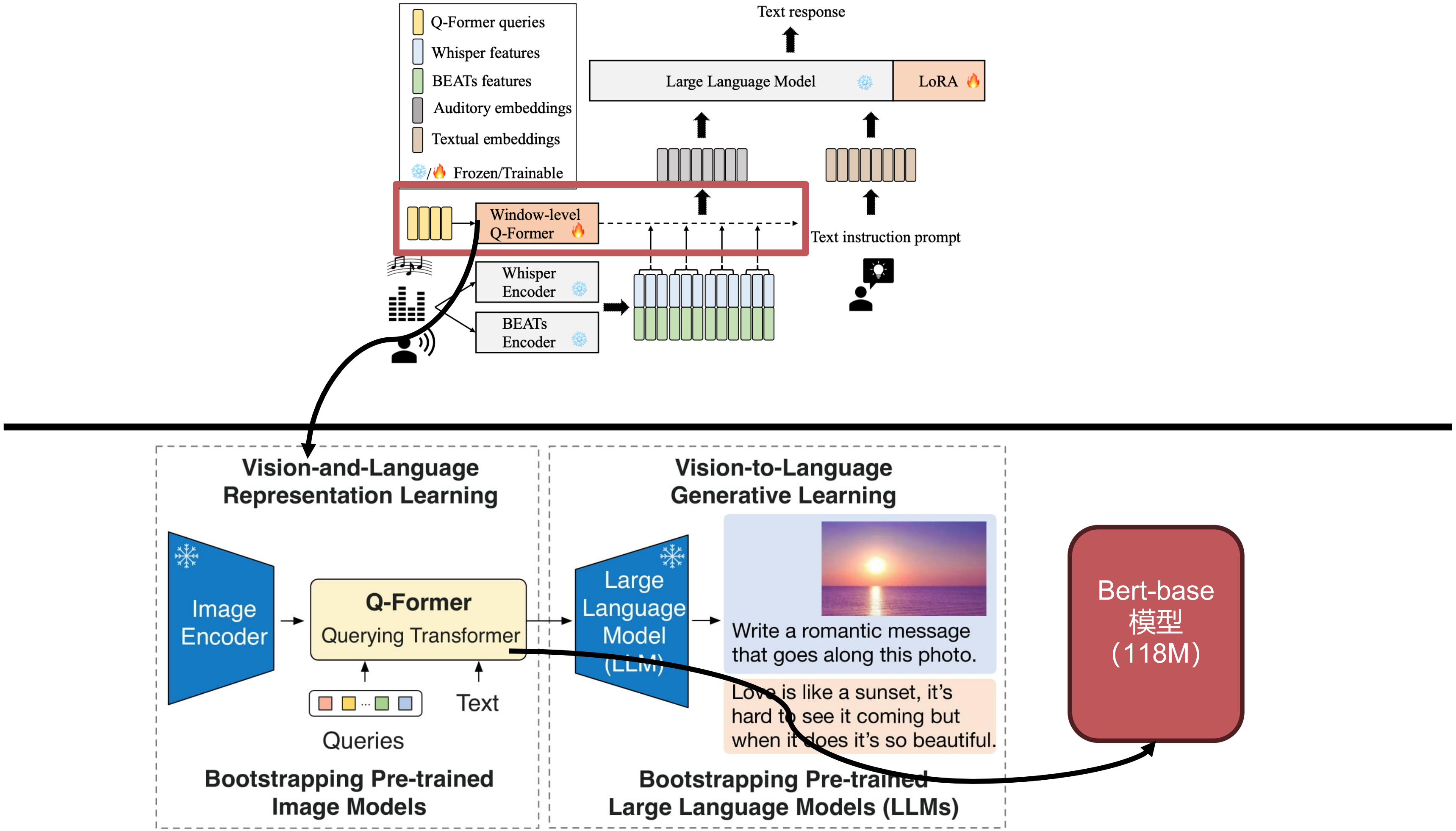
SALMONN: TOWARDS GENERIC HEARING ABILITIES FOR LARGE LANGUAGE MODELS

陈安东

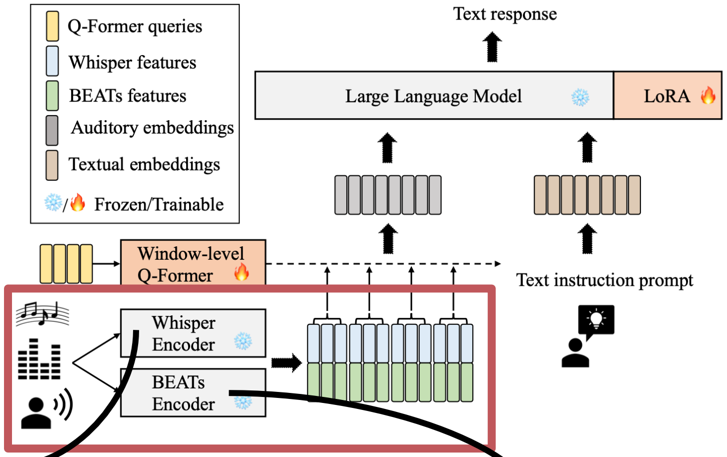
模型架构及方法

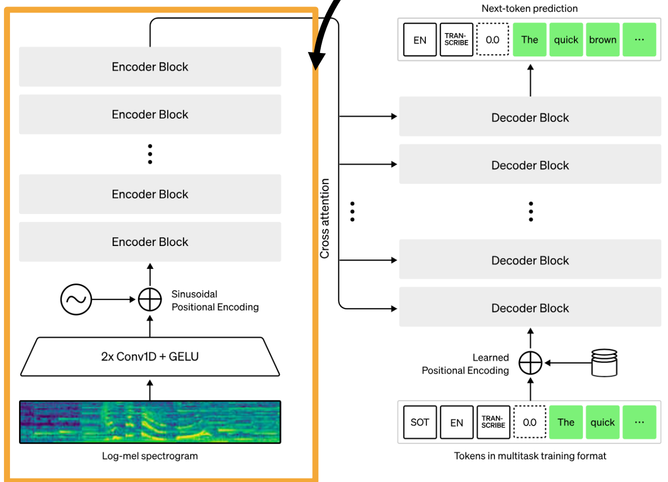

Q-Former queries
Whisper features
BEATs features
Auditory embeddings
Textual embeddings
Frozen/Trainable
Text response
Large Language Model
LoRA
Window-level Q-Former
Whisper Encoder
BEATs Encoder
Text instruction prompt


Q-Former queries
Whisper features
BEATs features
Auditory embeddings
Textual embeddings
Frozen/Trainable
Text response
Large Language Model
LoRA
Window-level Q-Former
Text instruction prompt
Whisper Encoder
BEATs Encoder
Vision-and-Language Representation Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Bootstrapping Pre-trained Image Models
Vision-to-Language Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained Large Language Models (LLMs)

Bert-base
模型
（118M）

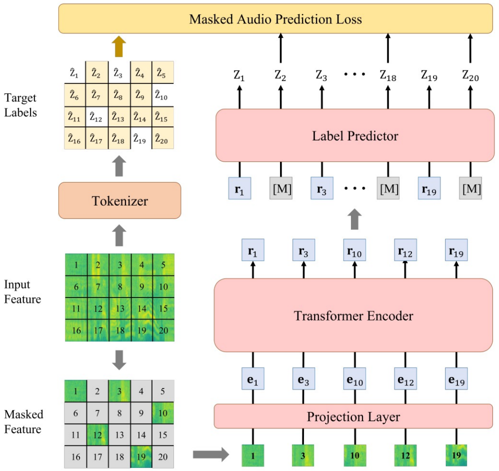
BLIP-2 architecture.



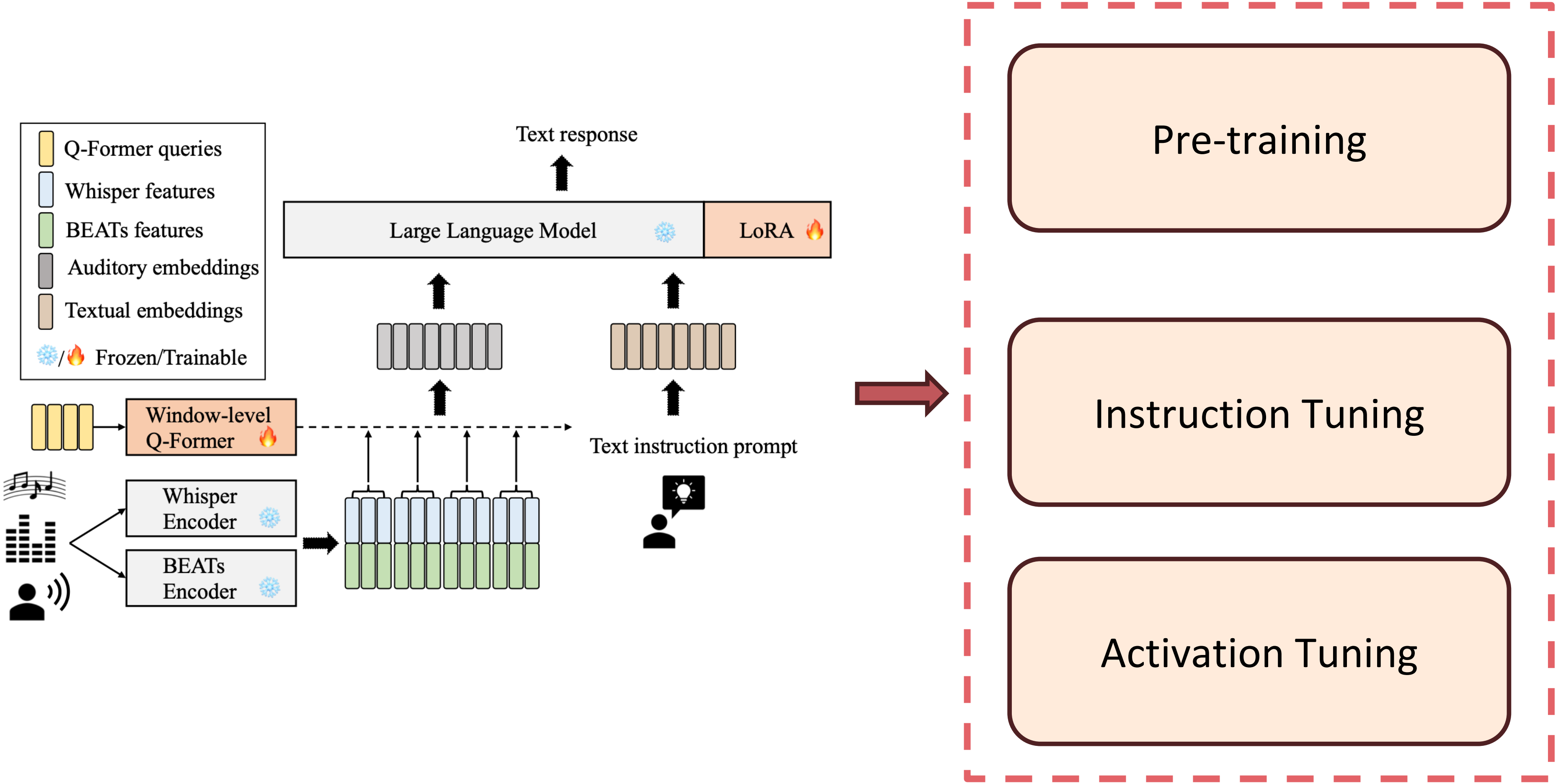


多语言
多任务
语音
预训练模型

Bidirectional
语音特征
编码器


Masked Audio Prediction Loss
Target Labels
Tokenizer
Input Feature
Masked Feature
Label Predictor
Transformer Encoder
Projection Layer
Z_1 Z_2 Z_3 ... Z_{18} Z_{19} Z_{20}
r_1 [M] r_3 ... [M] r_{19} [M]
r_1 r_3 r_{10} r_{12} r_{19}
e_1 e_3 e_{10} e_{12} e_{19}
1 3 10 12 19

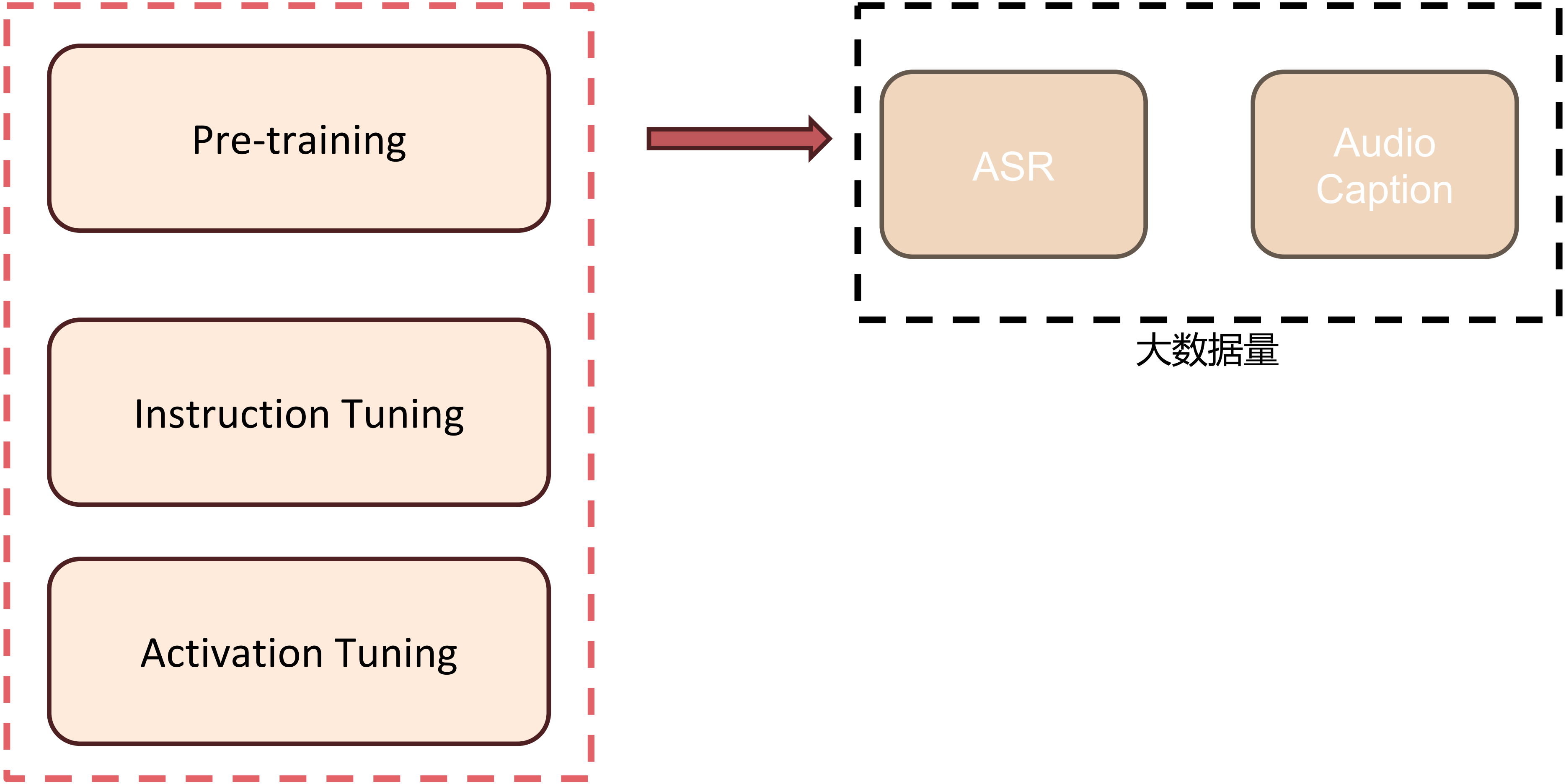
训练方案



Pre-training

Instruction Tuning

Activation Tuning



Pre-training

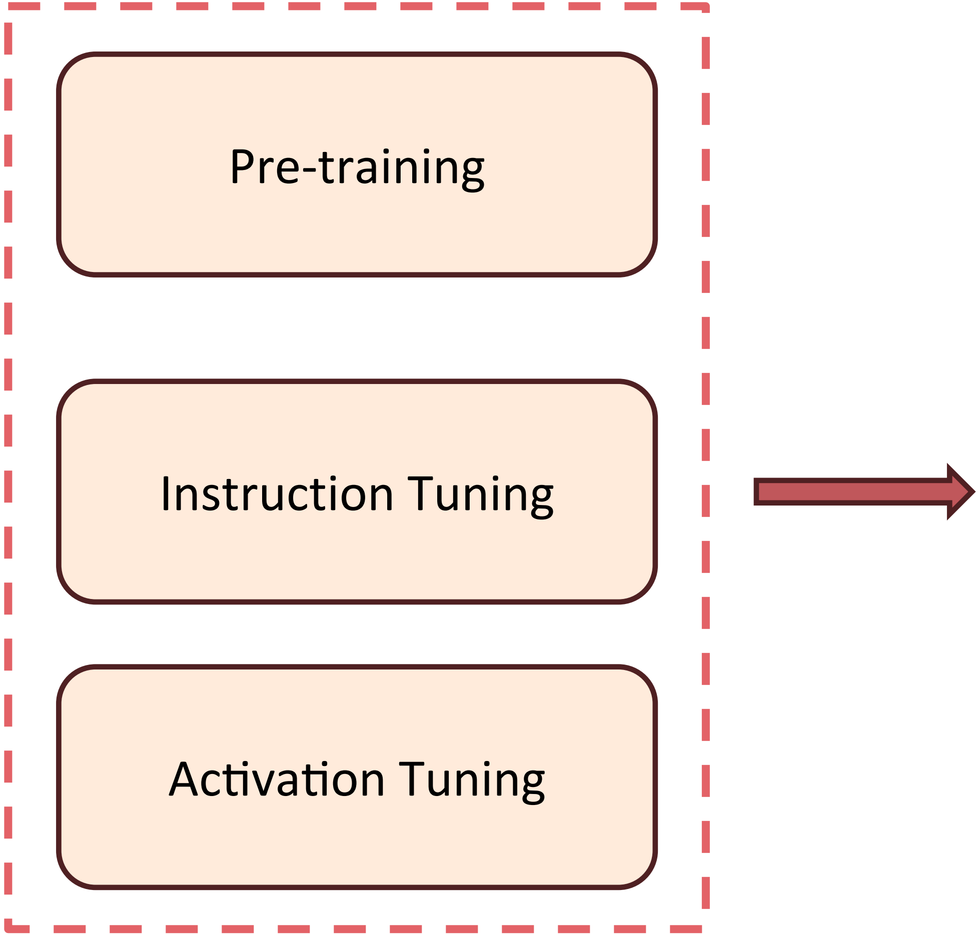
Instruction Tuning

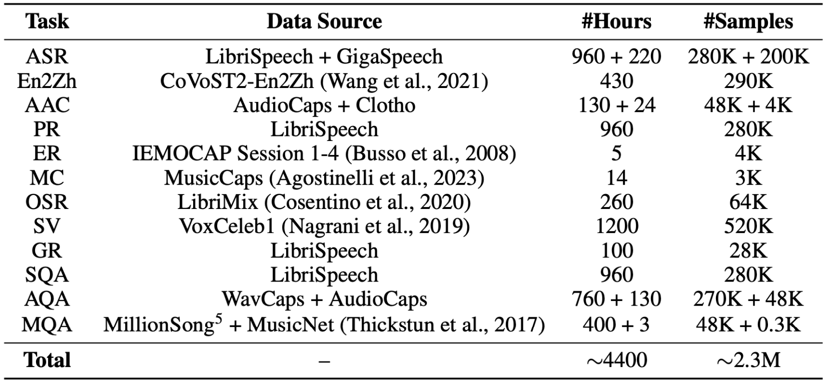
Activation Tuning

ASR

Audio Caption

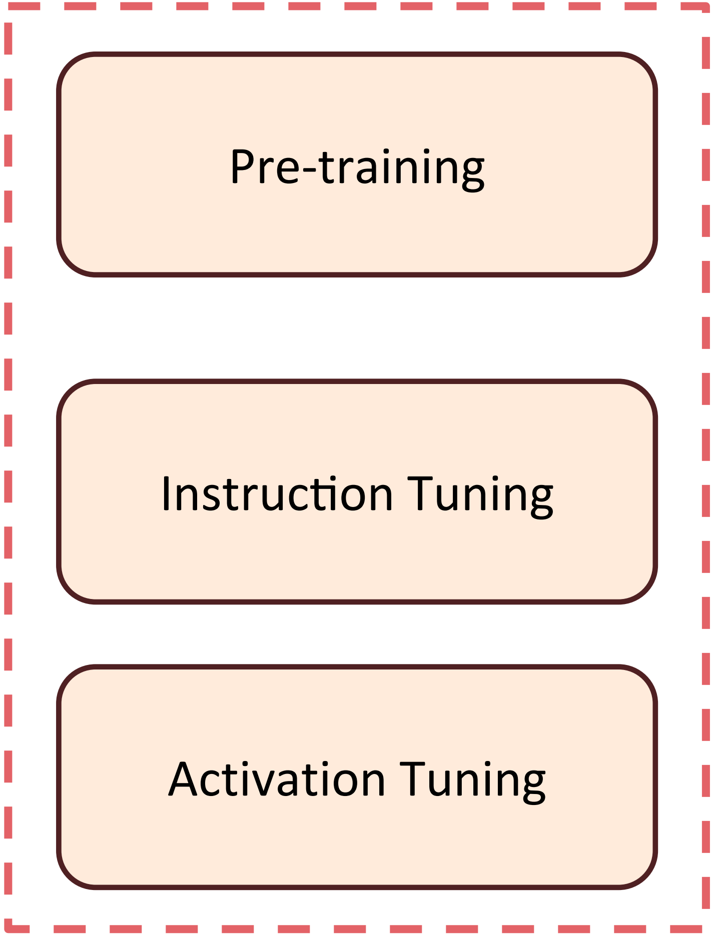
大数据量

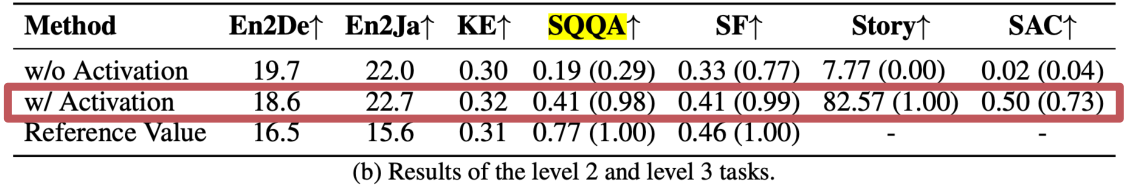




Task	Data Source	#Hours	#Samples
ASR	LibriSpeech + GigaSpeech	960 + 220	280K + 200K
En2Zh	CoVoST2-En2Zh (Wang et al., 2021)	430	290K
AAC	AudioCaps + Clotho	130 + 24	48K + 4K
PR	LibriSpeech	960	280K
ER	IEMOCAP Session 1-4 (Busso et al., 2008)	5	4K
MC	MusicCaps (Agostinelli et al., 2023)	14	3K
OSR	LibriMix (Cosentino et al., 2020)	260	64K
SV	VoxCeleb1 (Nagrani et al., 2019)	1200	520K
GR	LibriSpeech	100	28K
SQA	LibriSpeech	960	280K
AQA	WavCaps + AudioCaps	760 + 130	270K + 48K
MQA	MillionSong[5] + MusicNet (Thickstun et al., 2017)	400 + 3	48K + 0.3K
Total	–	∼4400	∼2.3M

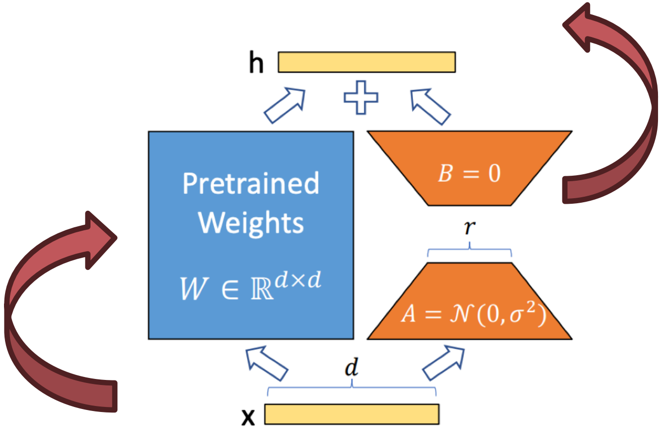
多任务

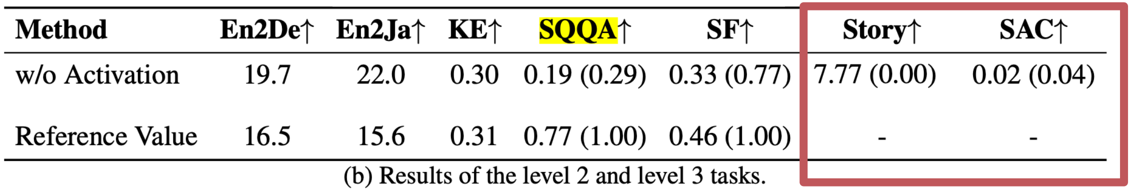




Method	En2De↑	En2Ja↑	KE↑	SQQA↑	SF↑	Story↑	SAC↑
w/o Activation	19.7	22.0	0.30	0.19 (0.29)	0.33 (0.77)	7.77 (0.00)	0.02 (0.04)
w/ Activation	18.6	22.7	0.32	0.41 (0.98)	0.41 (0.99)	82.57 (1.00)	0.50 (0.73)
Reference Value	16.5	15.6	0.31	0.77 (1.00)	0.46 (1.00)	-	-

(b) Results of the level 2 and level 3 tasks.

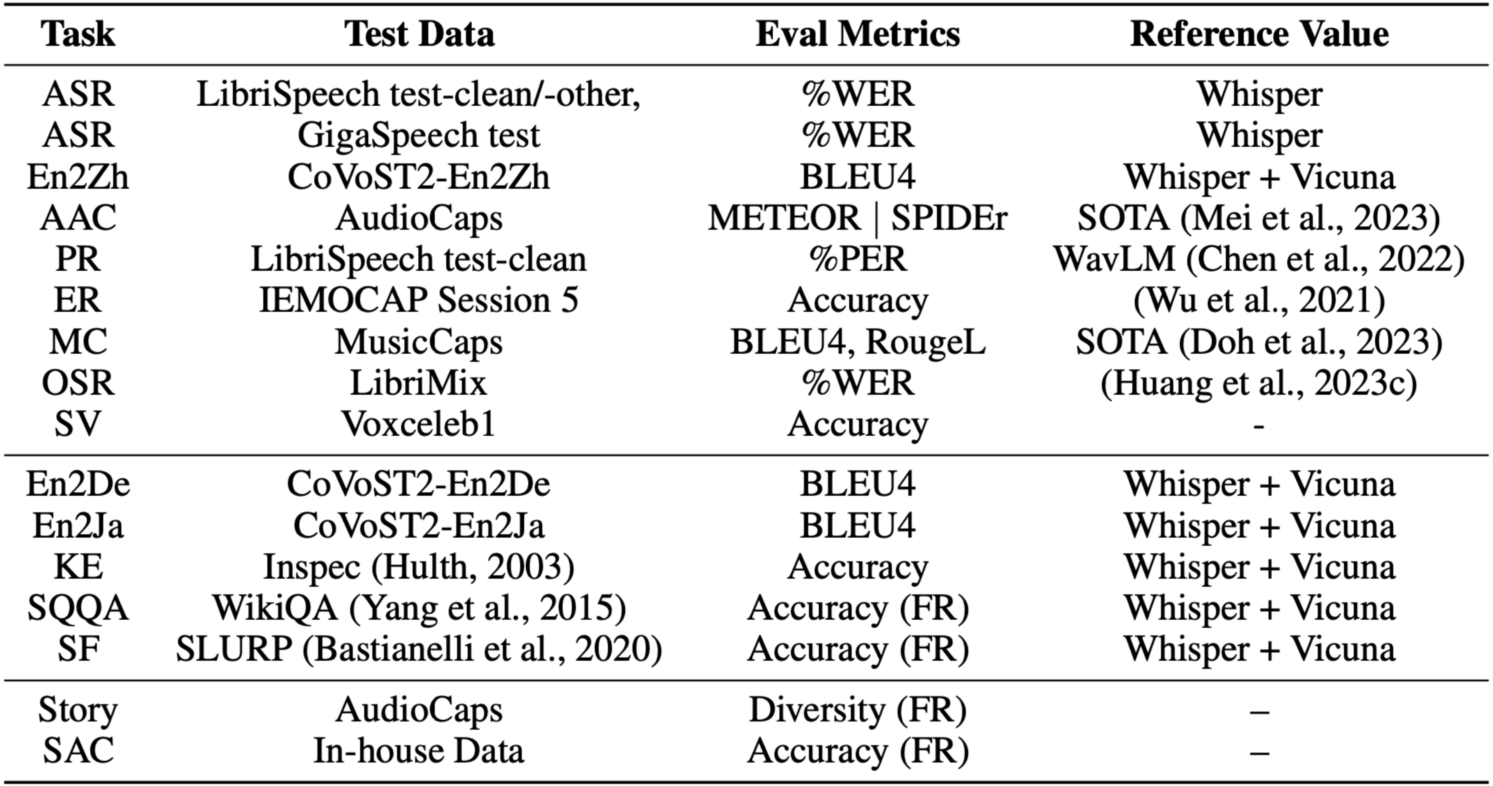




Method	En2De↑	En2Ja↑	KE↑	SQQA↑	SF↑	Story↑	SAC↑
w/o Activation	19.7	22.0	0.30	0.19 (0.29)	0.33 (0.77)	7.77 (0.00)	0.02 (0.04)
Reference Value	16.5	15.6	0.31	0.77 (1.00)	0.46 (1.00)	-	-

(b) Results of the level 2 and level 3 tasks.

实验设计



	Task	Test Data	Eval Metrics	Reference Value
Level 1 微调的任务内	ASR	LibriSpeech test-clean/-other,	%WER	Whisper
	ASR	GigaSpeech test	%WER	Whisper
	En2Zh	CoVoST2-En2Zh	BLEU4	Whisper + Vicuna
	AAC	AudioCaps	METEOR | SPIDEr	SOTA (Mei et al., 2023)
	PR	LibriSpeech test-clean	%PER	WavLM (Chen et al., 2022)
	ER	IEMOCAP Session 5	Accuracy	(Wu et al., 2021)
	MC	MusicCaps	BLEU4, RougeL	SOTA (Doh et al., 2023)
	OSR	LibriMix	%WER	(Huang et al., 2023c)
	SV	Voxceleb1	Accuracy	-
Level 2 分布外任务	En2De	CoVoST2-En2De	BLEU4	Whisper + Vicuna
	En2Ja	CoVoST2-En2Ja	BLEU4	Whisper + Vicuna
	KE	Inspec (Hulth, 2003)	Accuracy	Whisper + Vicuna
	SQQA	WikiQA (Yang et al., 2015)	Accuracy (FR)	Whisper + Vicuna
	SF	SLURP (Bastianelli et al., 2020)	Accuracy (FR)	Whisper + Vicuna
Level 3 逻辑推理题	Story	AudioCaps	Diversity (FR)	–
	SAC	In-house Data	Accuracy (FR)	–

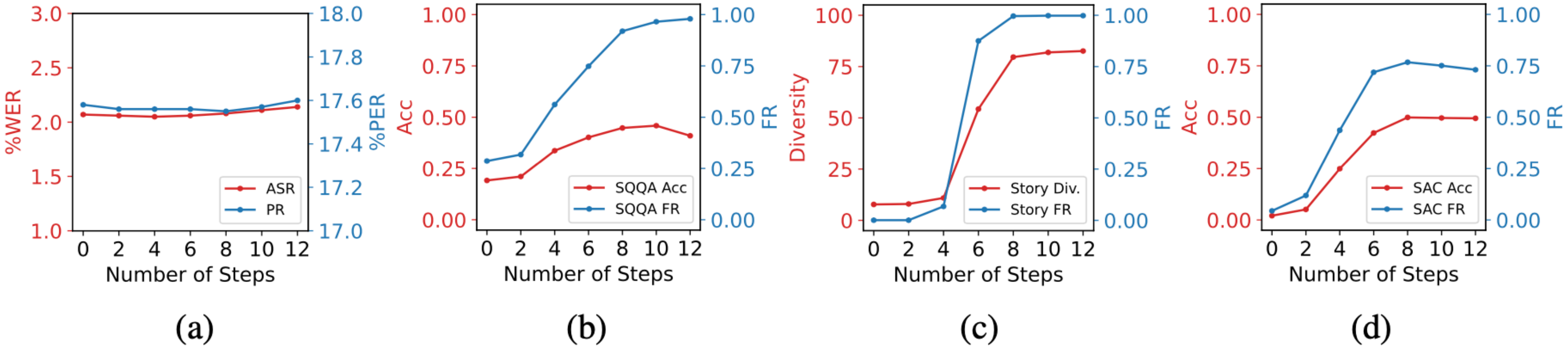


Figure 2: Performance changes on ASR & PR (a), SQQA (b), Story (c) and SAC (d) along with the FR of the emergent abilities against the number of training steps during activation tuning.

Level 1微调的任务内

Level 2 分布外任务

Level 3 逻辑推理题

SALMONN: Towards Generic Hearing Abilities for Large Language Models

① 当前语音大模型的基本架构变化较为统一

② 机器同传也会面临机器翻译领域相同的问题（多语言/低资源）

③ 机器同传具有独有的特点（单调性/源端不完整），具体较大的挖坑空间

④ 机器同传具有多模态的属性，多模态的同传场景工作不多，但每年都有

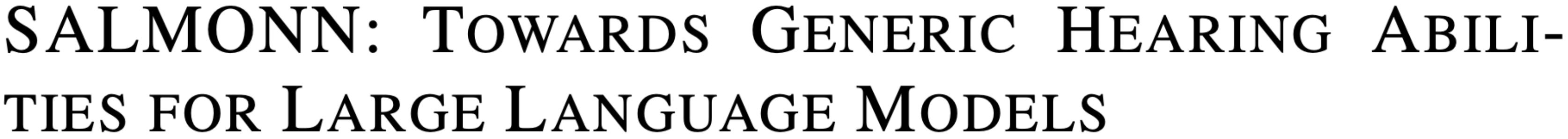
SALMONN: Towards Generic Hearing Abilities for Large Language Models

① 当前语音大模型的基本架构变化较为统一

② 发展趋势使得我们对多模态的组件需要有进一步的认识

③ 机器同传具有独有的特点（单调性/源端不完整），具体较大的挖坑空间

④ 机器同传具有多模态的属性，多模态的同传场景工作不多，但每年都有



SALMONN: Towards Generic Hearing Abilities for Large Language Models

① 当前语音大模型的基本架构变化较为统一

② 发展趋势使得我们对多模态的组件需要有进一步的认识

③ 大模型在多模态应用使得模态对齐工作成为一些工作的切入点