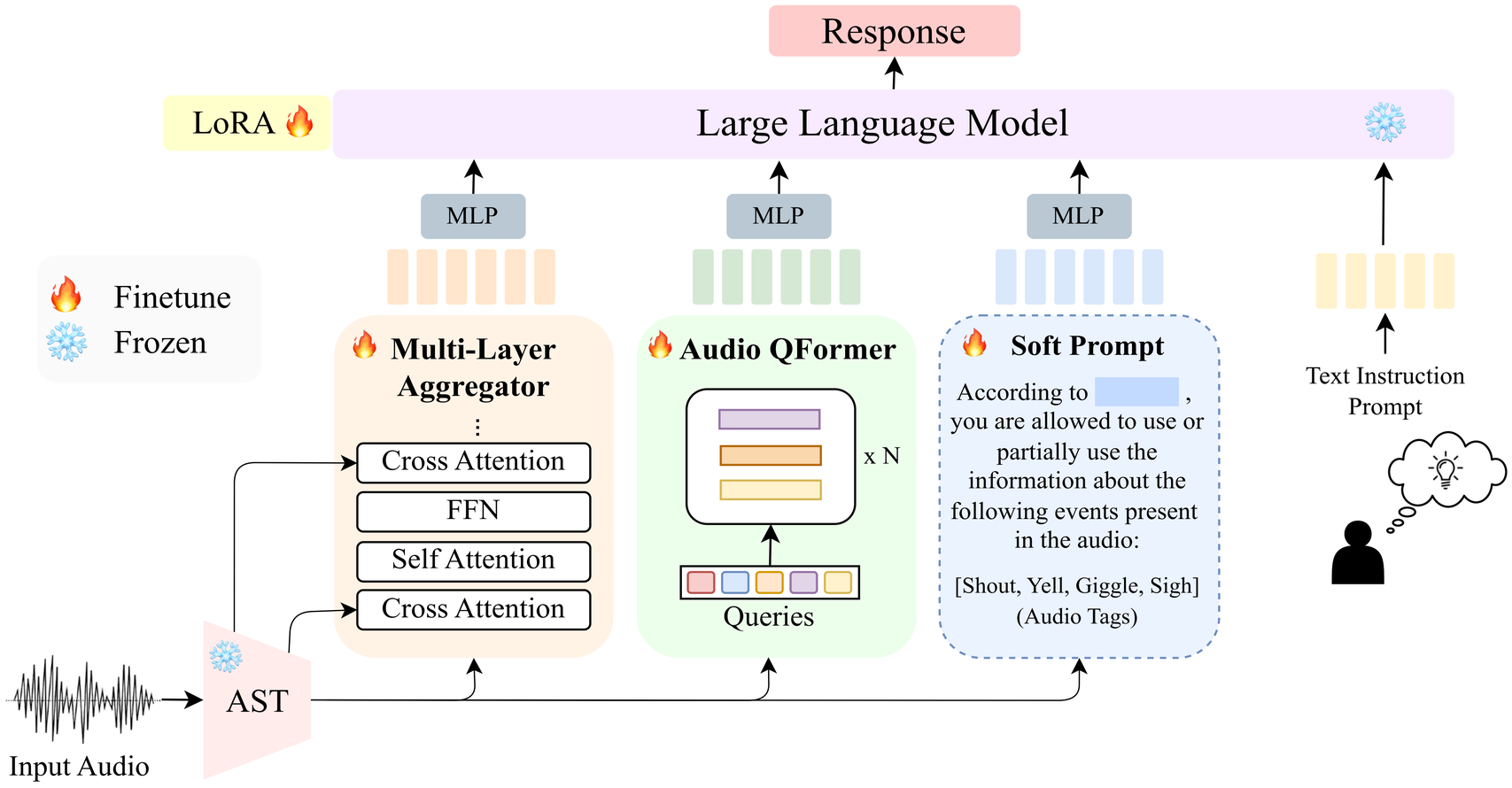
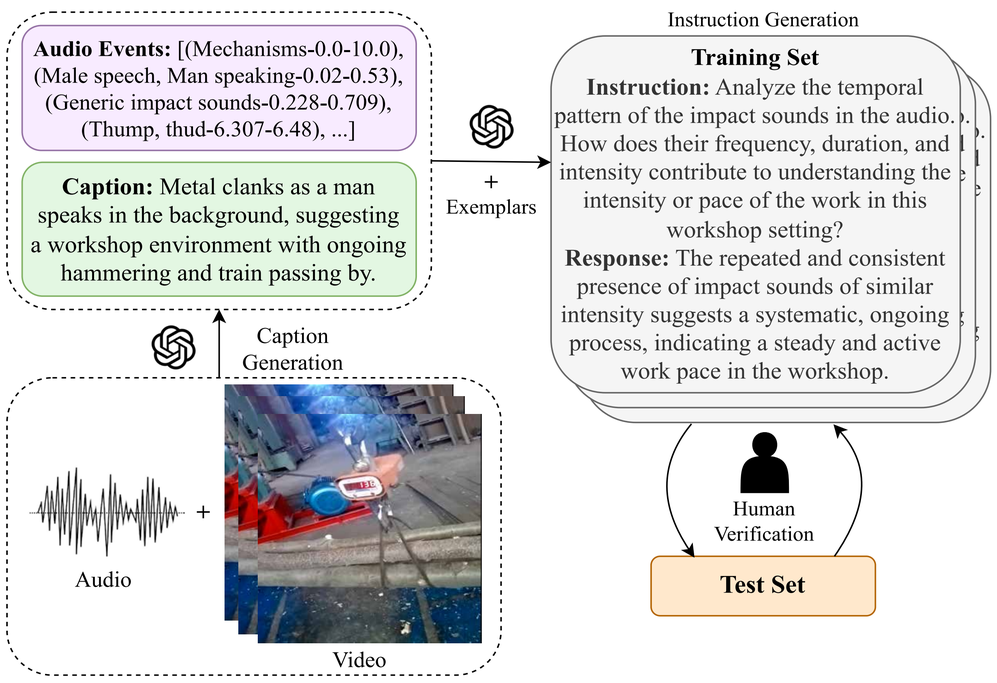
GAMA: Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities

Sreyan Ghosh[1*], Sonal Kumar[1*], Ashish Seth[1], Chandra Kiran Reddy Evuru[1],
Utkarsh Tyagi[1], S Sakshi[1], Oriol Nieto[2], Ramani Duraiswami[1], Dinesh Manocha[1]
[1]University of Maryland, College Park, USA [2]Adobe, USA

Paper and Code



CompA - R



Pipeline for synthesizing CompA-R

Results

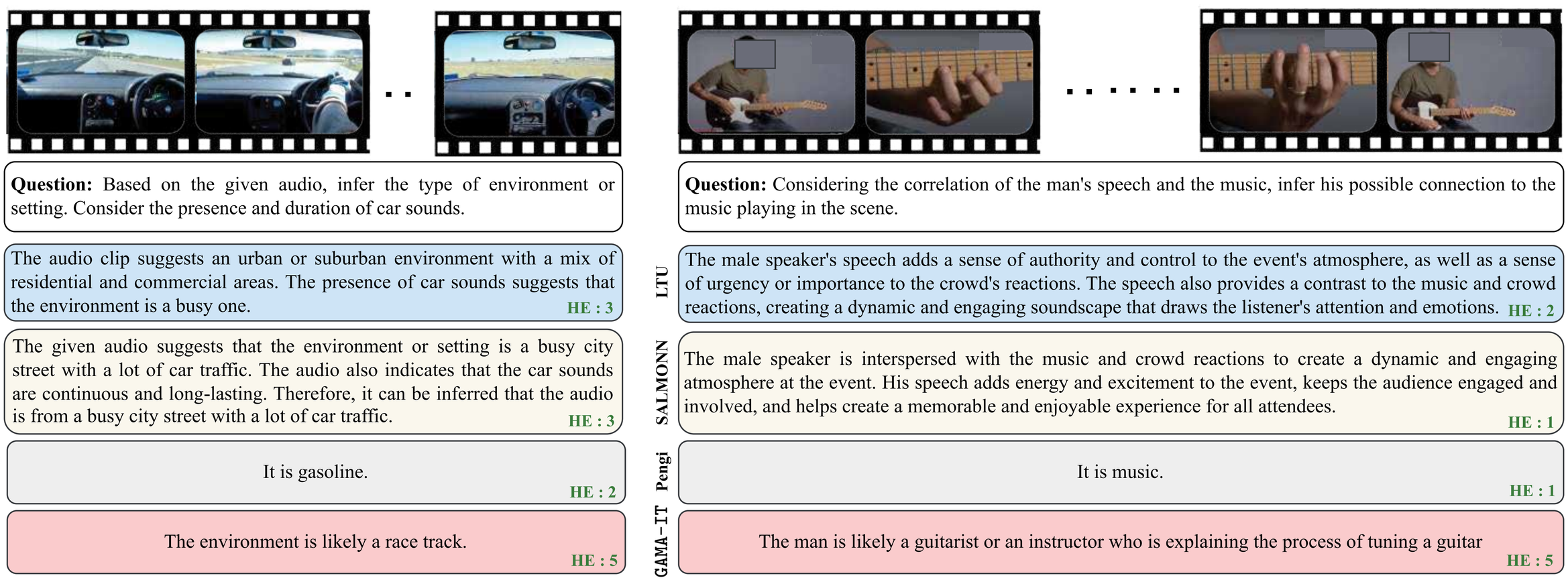
Model	ESC50# (Acc)	DCASE# (Mi-F1)	VS† (Acc)	TUT† (Acc)	BJO† (Acc)	VGG (Acc)	FSD (mAP)	$NS_{ins.}$ (ACC)	$NS_{src.}$ (ACC)	GTZAN† (ACC)	MSD† (ACC)	AudioSet (mAP)	Classif. Avg.	AudioCaps (SPICE)	Clotho (SPICE)	Cap. Avg.	ClothoAQA (ACC)
Audio-Language encoder-based models. They are generalizable to unseen labels, but a pre-defined label set is required for inference.																	
AudioCLIP	69.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
CLAP (Elizalde et al., 2023a)	82.6	30.0	48.4	29.6	47.5	24.0	30.2	22.7	16.4	25.0	44.0	5.8	29.4	-	-	-	-
CLAP (Wu* et al., 2023a)	89.1	31.3	47.1	35.6	48.0	26.3	30.8	25.2	18.9	26.3	46.9	6.2	36.0	-	-	-	-
CompA-CLAP	90.1	30.6	49.5	35.8	48.2	29.5	31.5	24.9	17.0	26.1	46.2	6.2	36.3	-	-	-	-
Audio-Language generation-based models. They directly output label names and do not need a pre-defined label set is needed at inference.																	
Qwen-Audio-Chat	71.7	32.4	74.2	16.9	50.8	17.5	39.8	30.2	41.3	41.6	69.1	13.4	41.1	14.7	9.8	12.3	32.3
LTU	81.7	37.5	**53.3**	19.9	67.8	50.3	43.9	28.0	41.8	9.9	74.2	18.3	42.4	16.9	11.7	15.8	25.1
SALMONN	16.4†	18.0†	16.9†	7.8†	25.0†	23.3†	22.1†	16.2†	33.7†	10.1†	28.8†	13.4†	17.9	8.3	7.6	8.0	23.1†
Pengi	80.8†	29.6†	46.4†	18.4†	47.3†	16.6†	35.8	39.2	46.0	11.9	**93.0**	11.5	39.7	12.7	7.0	9.9	63.6
AudioGPT	41.3	20.9	35.8	14.9	21.6	5.6	18.8	40.9	15.6	11.9	28.5	12.7	22.4	6.9	6.2	6.6	33.4
GAMA *(ours)*	**82.6**	**38.4**	52.4	21.5	**69.5**	**52.2**	**47.8**	**63.9**	**99.5**	**13.8**	85.6	**19.2**	**53.9**	**18.5**	**13.5**	**16.0**	**71.6**

Comparison of GAMA with baselines on evaluation datasets described on close-ended general audio and music understanding benchmarks.

Models	CompA-R-*test* (GPT-4/Human) Clarity	Correctness	Engagement	Avg.	OpenAQA Clarity	Correctness	Engagement	Avg.	Dense Captioning AudioCaps	Clotho	Avg.
Qwen-Audio-Chat	3.5 / 3.4	3.3 / 3.4	3.6 / 3.7	3.5 / 3.5	3.6	3.6	3.5	3.6	3.8	3.6	3.7
LTU	3.5 / 4.0	3.2 / 3.3	3.4 / 3.5	3.4 / 3.6	3.5	3.7	3.5	3.6	3.5	3.6	3.5
SALMONN	2.6 / 2.8	2.4 / 2.3	2.0 / 2.2	2.3 / 2.4	2.4	2.5	2.7	2.5	2.8	3.1	2.9
Pengi	1.8 / 1.6	1.5 / 1.4	1.3 / 1.2	1.5 / 1.4	1.7	1.5	1.4	1.5	2.6	2.8	2.7
AudioGPT	1.3 / 1.4	1.6 / 1.5	1.4 / 1.7	1.4 / 1.5	1.6	1.5	1.5	1.5	2.7	2.9	2.8
LTU w/ CompA-R	3.5 / 4.0	3.2 / 3.3	3.4 / 3.5	3.6 /3.6	3.5	3.7	3.5	3.6	3.7	3.8	3.8
GAMA *(ours)*	3.6 / 3.9	3.8 / 4.0	3.3 / 3.6	3.6 / 3.8	3.7	3.6	3.5	3.6	3.6	3.8	3.7
GAMA-IT *(ours)*	**4.3 / 4.5**	**3.9 / 4.1**	**3.9 / 4.3**	**4.0 / 4.3**	**4.0**	**4.2**	**3.8**	**4.0**	**4.3**	**4.1**	**4.2**

Comparison of GAMA with other baselines on open-ended AQA on OpenAQA, complex open-ended AQA on CompA-R-test and Dense Captioning on 500 instances from AudioCaps and Clotho

Evaluation examples of complex reasoning



Qualitative comparison of GAMA with other baselines on instances from CompA-R-test.