CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models

Sreyan Ghosh[1 2*] Ashish Seth[1*] Sonal Kumar[1*] Utkarsh Tyagi[1*]

Chandra Kiran Reddy Evuru[1*] Ramaneswaran S[3] S Sakshi[1*] Oriol Nieto[2]

Ramani Duraiswami[1] Dinesh Manocha[1]

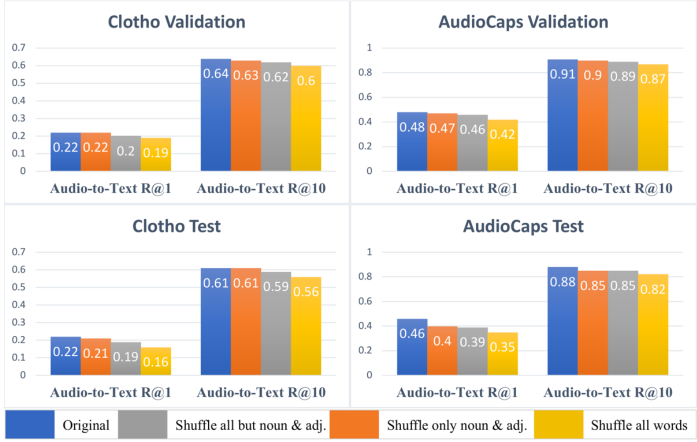
[1]University of Maryland, College Park, USA

[2]Adobe USA, [3]NVIDIA India

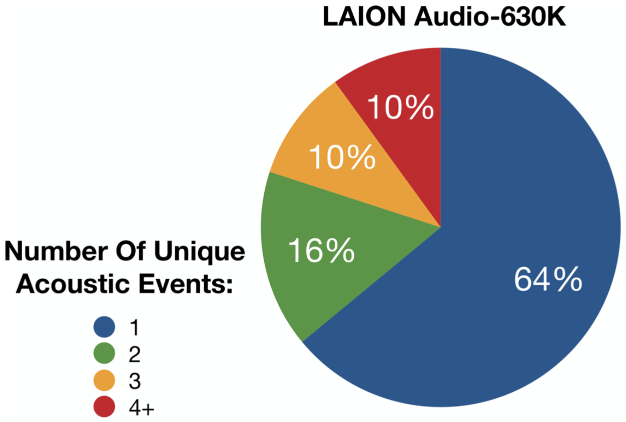
Introduction

- **What are Audio-Language Models?:** Audio-Language Models (ALMs) like Contrastive Language-Audio Pre-training (CLAP) learn a shared space between the audio and language modalities. This allows them to solve audio tasks through a language interface. Recently CLAP-like models have achieved SOTA in various downstream tasks, including zero-shot audio classification, audio retrieval, etc.
- **What is compositional reasoning?:** Understanding the relationship between text in captions and the corresponding content of the audio is a fundamental goal of audio processing, and the fact that different word orders correspond to differently perceived audio should be reflected in the capabilities of the ALMs. This phenomenon, also known as compositional reasoning, may be characterized as the ALM's capacity to understand the interrelationships among multiple discrete acoustic events in audio, such as order of occurrence and attribute-binding, as conveyed through the words in the caption

Despite their success, the extent to which ALMs can perform compositional reasoning is largely under-explored. Our paper aims at bridging this gap by evaluating and improving compositional reasoning in ALMs



Performance on common retrieval evaluation datasets with shuffling.



Distribution of audios with the number of unique acoustic events in LAION-Audio-630k, the largest open-source audio-caption training dataset.

Primary Motivation

- **Rethinking Evaluation of Compositional Reasoning in ALMs:** Current retrieval benchmarks are insufficient in evaluating compositional reasoning of ALMs. Wu et al. (2023) also show that ALMs often act as bag of words and lack natural language comprehension.

- **Lack of sufficient training data to learning Compositional Reasoning:** There is an acute scarcity of compositional audios in large audio-text pre-training benchmarks.

Main Contributions

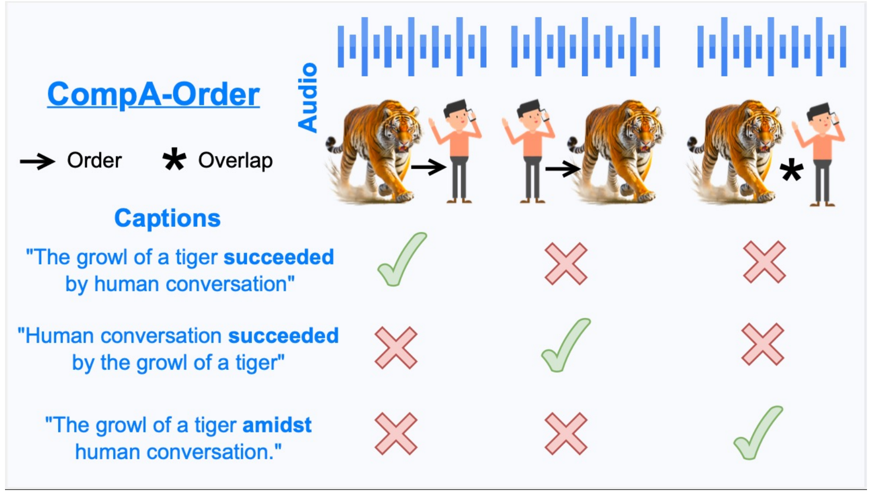
Evaluating and Improving Compositional Reasoning in Audio-Language Models

- **CompA Benchmarks:**
  - We develop two expert-annotated benchmarks, ***CompA-order*** and ***CompA-attribute***, to assess compositional reasoning in ALMs.
  - CompA-order tests the models' understanding of the order of audio events, while CompA-attribute focuses on how attributes are associated with specific events.
  - These benchmarks include a diverse set of real-world audio samples, making them robust tools for evaluating ALM capabilities.
- **Improved Contrastive Learning Techniques:**
  - We introduce ***CompA-CLAP***, trained on a novel mixture of freely available datasets, including CompA-AudioSet, that is rich in compositional audio-caption pairs.
  - To improve compositional reasoning, we first propose training with composition-aware hard negatives.
  - Finally, to overcome the acute scarcity of open-source compositional audio-caption pairs, we propose modular contrastive learning to scale CompA-CLAP training.

CompA-Order

Evaluating Order Understanding in Audio

- We build CompA-order to evaluate an ALM’s ability to understand the order of occurrence between multiple acoustic events.
- CompA-order has 400 test instances. Each instance includes pairs of audio caption pairs (and a maximum of 3), where each audio has the same events but the order of their occurrence in different sequences.
- The captions for the audios in an instance with two pairs have the exact same words but in a different order, except for the i nstances with three pairs, where only a single word that defines the preposition between the events is changed.



Captions	→	→	*
"The growl of a tiger **succeeded** by human conversation"	✓	✗	✗
"Human conversation **succeeded** by the growl of a tiger"	✗	✓	✗
"The growl of a tiger **amidst** human conversation."	✗	✗	✓

CompA-order evaluates an ALMs’ capability to understand the *order of occurrence* between multiple acoustic events in an audio.

CompA-Order examples

🔊

A woman's chatter followed
by the pouring of a liquid.

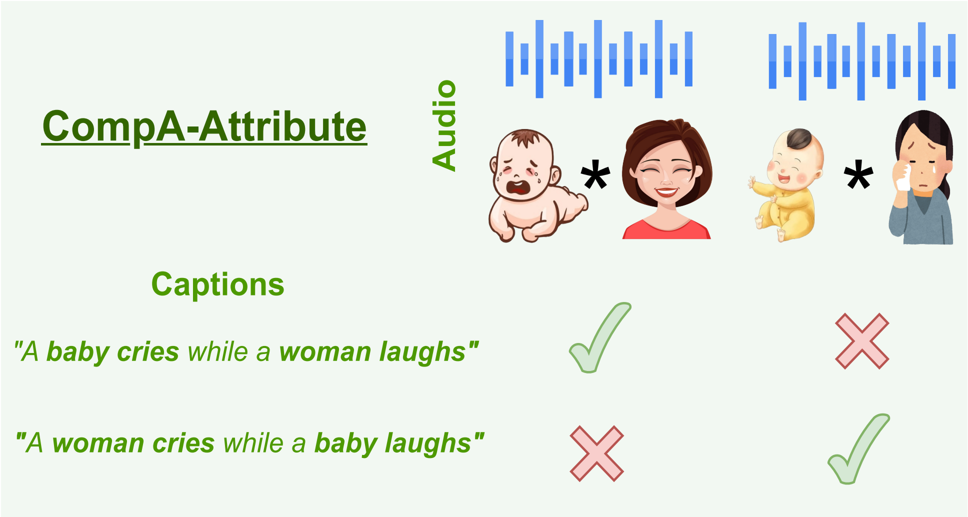
🔊

The pouring of a liquid
followed by a woman's chatter.

CompA-Attribute

Evaluating Attribute Understanding in Audio

- We build CompA-attribute to evaluate an audio-language model's ability to link attributes to specific acoustic events.
- CompA order has 200 test instances, where each instance includes pairs of audio clips and captions. The audio clips feature the same events but with differing attributes.
- Models are challenged to match each audio clip with a caption that accurately describes the attributes of the events.


CompA-Attribute
Audio
*
*
Captions
"A baby cries while a woman laughs"
✓
✗
"A woman cries while a baby laughs"
✗
✓

CompA-attribute evaluates an ALM's capability to understand *attribute-binding* for multiple acoustic events in an audio.

CompA-Attribute examples

🔊

A child sneezes and
an adult laughs.

🔊

A child laughs and an
adult sneezes.

Evaluation Setup

Given two audios A_0 and A_1 and their corresponding captions C_0 and C_1, from an instance in either benchmark, we define the *text score* that measures whether an ALM can select the correct caption, given an audio. We define *text score* as:

$$f(C_0, A_0, C_1, A_1) = \begin{cases} 1 & \text{if } s(C_0, A_0) > s(C_1, A_0) \\ & \text{and } s(C_1, A_1) > s(C_0, A_1) \\ 0 & \text{otherwise} \end{cases}$$

where s(·) is the cosine similarity between the audio-caption pair. The second metric is the audio score, which measures whether an ALM can select the correct audio, given a caption. We define *audio score* as:

$$h(C_0, A_0, C_1, A_1) = \begin{cases} 1 & \text{if } f(C_0, A_0, C_1, A_1) \\ & \text{and } g(C_0, A_0, C_1, A_1) \\ 0 & \text{otherwise} \end{cases}$$

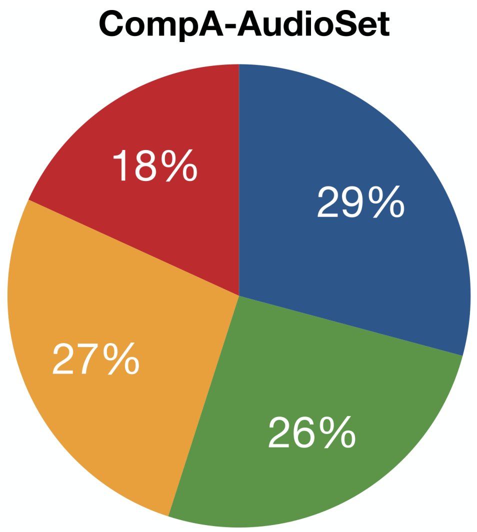
Finally, we define a *group score* combining the audio and text scores defined above as follows:

$$g(C_0, A_0, C_1, A_1) = \begin{cases} 1 & \text{if } s(C_0, A_0) > s(C_0, A_1) \\ & \text{and } s(C_1, A_1) > s(C_1, A_0) \\ 0 & \text{otherwise} \end{cases}$$

Improving Vanilla Contrastive Pre-training

- To compensate for the lack of compositional audio-caption pairs, we synthesize captions using GPT-4 for AudioSet-strong and with it build CompA-661k, a novel mixture of open-source audio-caption pairs. This is used to train CLAP from scratch.

- **Result:** CLAP trained on CompA-661k outperforms the best CLAP model available and improves on our compositional understanding benchmarks.

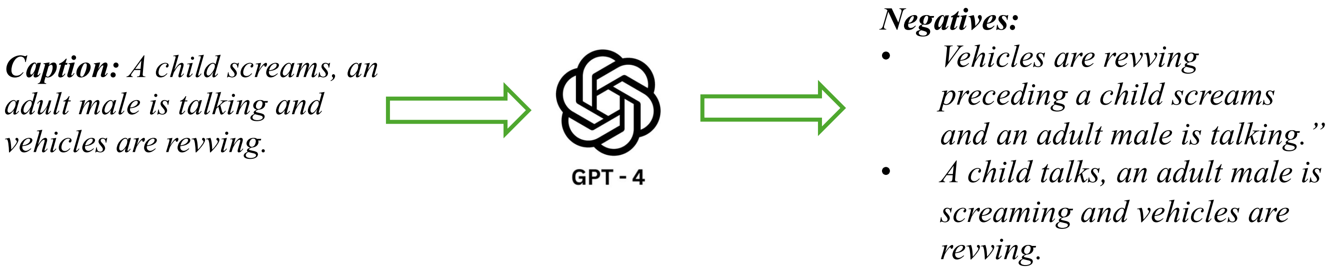


We build CompA-661k with a significant portion of compositional audios.

Contrastive Pre-Training with Compositionally-Aware Hard Negatives

Improving CLAPs Compositional Reasoning Through a Modified Contrastive Learning Formulation

- Each audio sample in the training batch is paired with hard negative captions that are ignored by other samples, ensuring targeted and effective learning.
- This training approach significantly improves the model's ability to differentiate subtle differences and relationships between audio events, crucial for complex audio understanding.

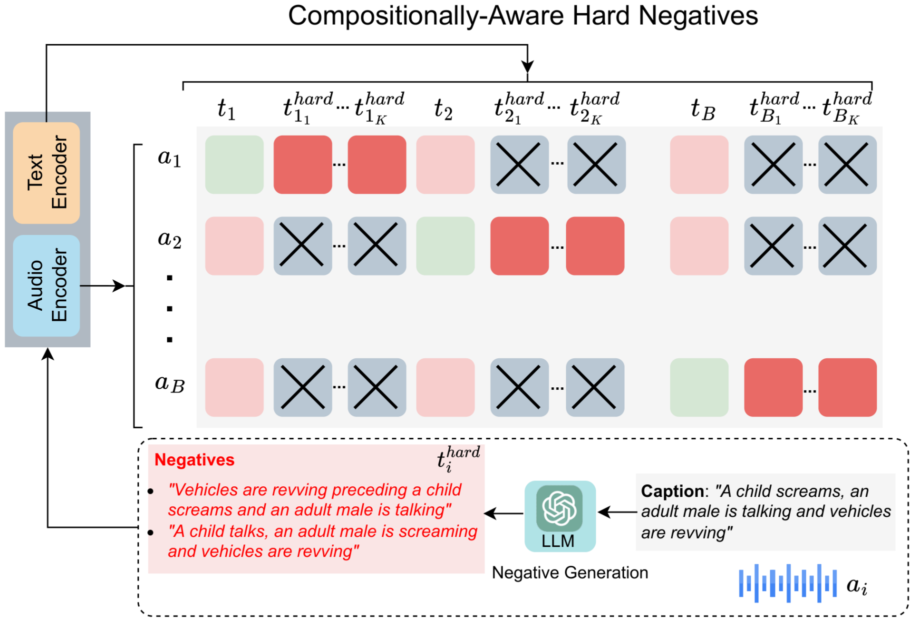


Caption: *A child screams, an adult male is talking and vehicles are revving.*

Negatives:

- *Vehicles are revving preceding a child screams and an adult male is talking."*
- *A child talks, an adult male is screaming and vehicles are revving.*

Synthesis of hard negatives using GPT-4.



Compositionally-Aware Hard Negatives
Text Encoder
Audio Encoder
t_1
$t_{1_1}^{hard}$ ... $t_{1_K}^{hard}$
t_2
$t_{2_1}^{hard}$ ... $t_{2_K}^{hard}$
t_B
$t_{B_1}^{hard}$ ... $t_{B_K}^{hard}$
a_1
a_2
a_B
Negatives
t_i^{hard}
LLM
Negative Generation
a_i

Negatives

- *"Vehicles are revving preceding a child screams and an adult male is talking"*
- *"A child talks, an adult male is screaming and vehicles are revving"*

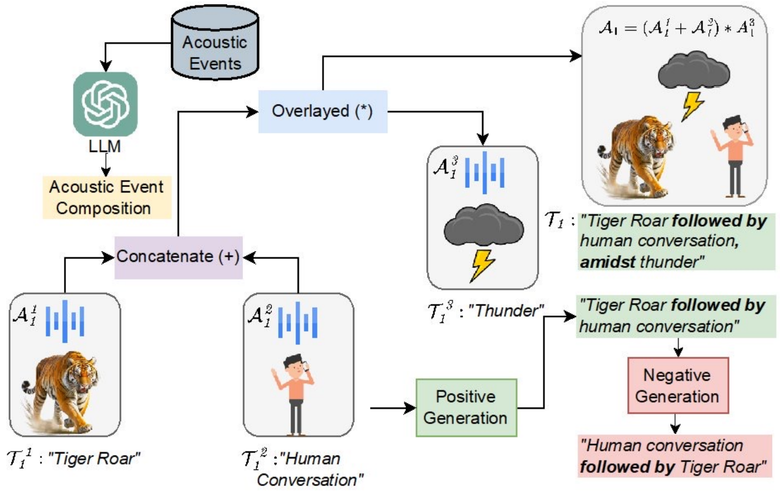
Caption: *"A child screams, an adult male is talking and vehicles are revving"*

Contrastive training with compositionally-aware hard negatives where each audio has K hard negative captions generated using an LLM, and each audio in the batch ignores negatives of other audios in the batch for more focused training.

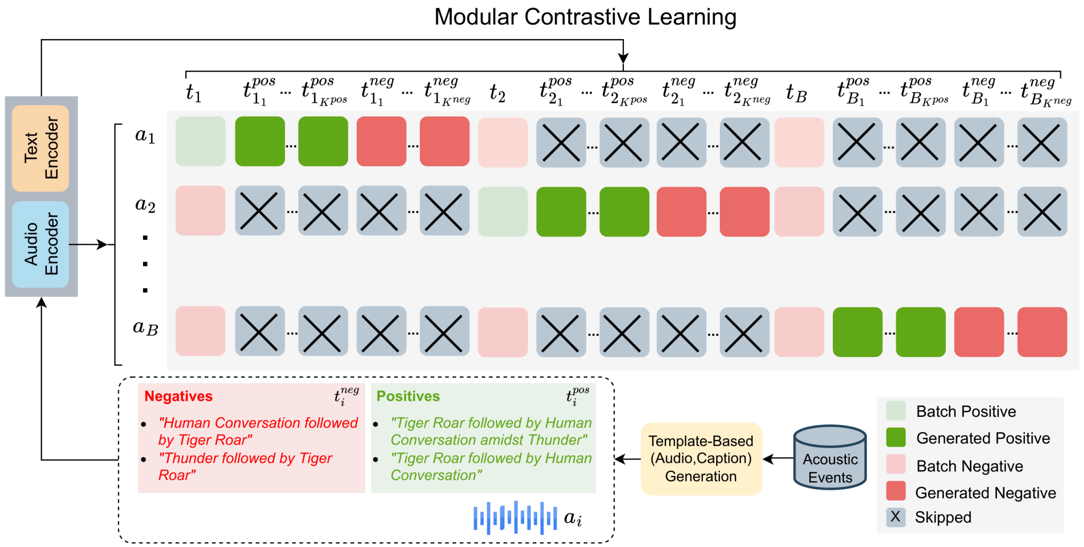
Modular Contrastive Learning

Compositional Audio is still Low-Resource!

- Contrastive Pre-training with hard negatives still requires compositional audios and their corresponding captions! Thus, we should find a way to overcome this requirement!
- Audios with a large number of audio events makes fine-grained learning difficult. For example, a single hard negative for an audio with multiple difficult-to-distinguish acoustic events can be too complicated for the model to understand effectively.

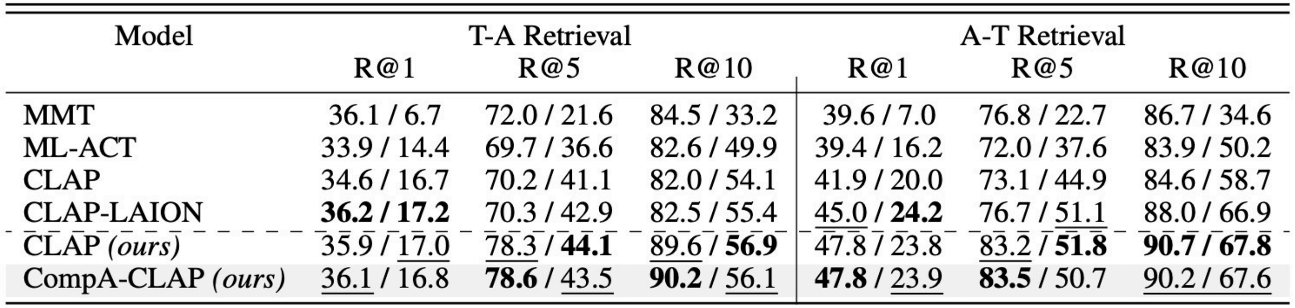

Acoustic Events
LLM
Acoustic Event Composition
Overlayed (*)
Concatenate (+)
$\mathcal{A}_1 = (\mathcal{A}_1^1 + \mathcal{A}_1^2) * \mathcal{A}_1^3$
$\mathcal{A}_1^3$
$\mathcal{T}_1^3$: "Thunder"
$\mathcal{T}_1$: "Tiger Roar **followed by** human conversation**,** **amidst** thunder"
$\mathcal{A}_1^1$
$\mathcal{A}_1^2$
$\mathcal{T}_1^1$: "Tiger Roar"
$\mathcal{T}_1^2$:"Human Conversation"
Positive Generation
"Tiger Roar **followed by** human conversation"
Negative Generation
"Human conversation **followed by** Tiger Roar"

A simple technique to generate unlimited compositional audio-caption pairs! 1. 1. Ask GPT to generate plausible real-world scenarios from a pre-defined label space. 2. Select the audios corresponding to the labels from a pool and either concatenate or overlay these audios. 3. Build the corresponding caption.


Modular Contrastive Learning
Text Encoder
Audio Encoder
t_1 $t_{1_1}^{pos}$ ... $t_{1_{K^{pos}}}^{pos}$ $t_{1_1}^{neg}$ ... $t_{1_{K^{neg}}}^{neg}$ t_2 $t_{2_1}^{pos}$ ... $t_{2_{K^{pos}}}^{pos}$ $t_{2_1}^{neg}$ ... $t_{2_{K^{neg}}}^{neg}$ t_B $t_{B_1}^{pos}$ ... $t_{B_{K^{pos}}}^{pos}$ $t_{B_1}^{neg}$ ... $t_{B_{K^{neg}}}^{neg}$
a_1
a_2
a_B
Negatives t_i^{neg}
"Human Conversation followed by Tiger Roar"
"Thunder followed by Tiger Roar"
Positives t_i^{pos}
"Tiger Roar followed by Human Conversation amidst Thunder"
"Tiger Roar followed by Human Conversation"
a_i
Template-Based (Audio,Caption) Generation
Acoustic Events
Batch Positive
Generated Positive
Batch Negative
Generated Negative
X Skipped

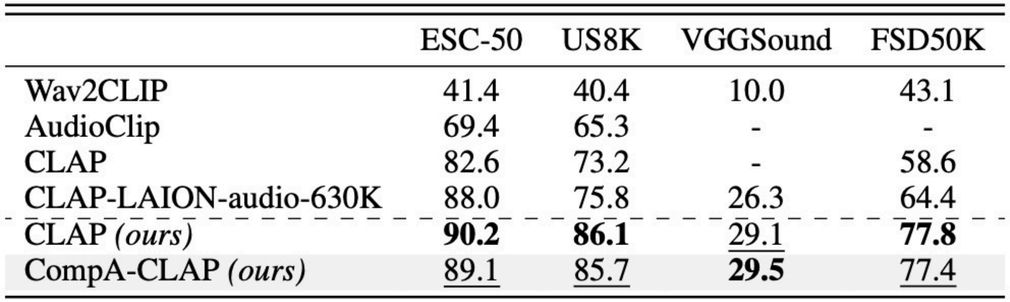
Each positive describes compositional relationships of various granularities in the audio, and this helps the model learn fine-grained order and attribute-binding. An audio in the batch ignores the positives and negatives of other audios.

Results



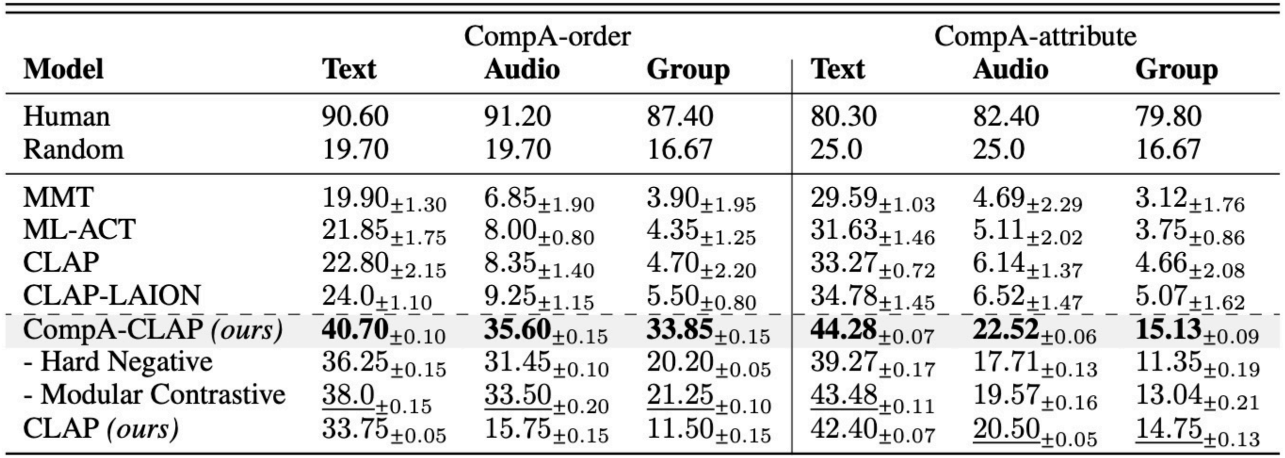
Model	T-A Retrieval			A-T Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
MMT	36.1 / 6.7	72.0 / 21.6	84.5 / 33.2	39.6 / 7.0	76.8 / 22.7	86.7 / 34.6
ML-ACT	33.9 / 14.4	69.7 / 36.6	82.6 / 49.9	39.4 / 16.2	72.0 / 37.6	83.9 / 50.2
CLAP	34.6 / 16.7	70.2 / 41.1	82.0 / 54.1	41.9 / 20.0	73.1 / 44.9	84.6 / 58.7
CLAP-LAION	**36.2 / 17.2**	70.3 / 42.9	82.5 / 55.4	45.0 / **24.2**	76.7 / 51.1	88.0 / 66.9
CLAP *(ours)*	35.9 / 17.0	78.3 / **44.1**	89.6 / **56.9**	47.8 / 23.8	83.2 / **51.8**	**90.7 / 67.8**
CompA-CLAP *(ours)*	36.1 / 16.8	**78.6** / 43.5	**90.2** / 56.1	**47.8** / 23.9	**83.5** / 50.7	90.2 / 67.6

Result comparison on retrieval benchmarks.



	ESC-50	US8K	VGGSound	FSD50K
Wav2CLIP	41.4	40.4	10.0	43.1
AudioClip	69.4	65.3	-	-
CLAP	82.6	73.2	-	58.6
CLAP-LAION-audio-630K	88.0	75.8	26.3	64.4
CLAP *(ours)*	**90.2**	**86.1**	29.1	**77.8**
CompA-CLAP *(ours)*	89.1	85.7	**29.5**	77.4

Result comparison on zero-shot audio classification benchmarks.



	CompA-order			CompA-attribute		
Model	**Text**	**Audio**	**Group**	**Text**	**Audio**	**Group**
Human	90.60	91.20	87.40	80.30	82.40	79.80
Random	19.70	19.70	16.67	25.0	25.0	16.67
MMT	$19.90_{\pm1.30}$	$6.85_{\pm1.90}$	$3.90_{\pm1.95}$	$29.59_{\pm1.03}$	$4.69_{\pm2.29}$	$3.12_{\pm1.76}$
ML-ACT	$21.85_{\pm1.75}$	$8.00_{\pm0.80}$	$4.35_{\pm1.25}$	$31.63_{\pm1.46}$	$5.11_{\pm2.02}$	$3.75_{\pm0.86}$
CLAP	$22.80_{\pm2.15}$	$8.35_{\pm1.40}$	$4.70_{\pm2.20}$	$33.27_{\pm0.72}$	$6.14_{\pm1.37}$	$4.66_{\pm2.08}$
CLAP-LAION	$24.0_{\pm1.10}$	$9.25_{\pm1.15}$	$5.50_{\pm0.80}$	$34.78_{\pm1.45}$	$6.52_{\pm1.47}$	$5.07_{\pm1.62}$
CompA-CLAP *(ours)*	$\mathbf{40.70}_{\pm0.10}$	$\mathbf{35.60}_{\pm0.15}$	$\mathbf{33.85}_{\pm0.15}$	$\mathbf{44.28}_{\pm0.07}$	$\mathbf{22.52}_{\pm0.06}$	$\mathbf{15.13}_{\pm0.09}$
- Hard Negative	$36.25_{\pm0.15}$	$31.45_{\pm0.10}$	$20.20_{\pm0.05}$	$39.27_{\pm0.17}$	$17.71_{\pm0.13}$	$11.35_{\pm0.19}$
- Modular Contrastive	$38.0_{\pm0.15}$	$33.50_{\pm0.20}$	$21.25_{\pm0.10}$	$43.48_{\pm0.11}$	$19.57_{\pm0.16}$	$13.04_{\pm0.21}$
CLAP *(ours)*	$33.75_{\pm0.05}$	$15.75_{\pm0.15}$	$11.50_{\pm0.15}$	$42.40_{\pm0.07}$	$20.50_{\pm0.05}$	$14.75_{\pm0.13}$

Result comparison on our proposed CompA benchmarks.

Conclusion and Future Work

Evaluating and Advancing Compositional Reasoning in Audio-Language Models

Key Takeaways:

- We demonstrate that current Audio-Language Models (ALMs) lack robust compositional reasoning capabilities, emphasizing the importance of compositional understanding in audio processing.
- CompA benchmarks and the CompA-CLAP model, which significantly enhance the compositional reasoning skills of ALMs.

Future Directions

- **Expand the CompA Benchmarks:** Introduce more complex scenarios and a greater variety of compositional challenges to further push the capabilities of ALMs.
- **Refine Training Techniques:** Continue to develop and refine training methodologies to include more nuanced compositional aspects and real-world variability.
- **Cross-Modal Applications:** Explore the application of compositional reasoning skills in other modalities, such as video and text, to foster cross-modal learning and understanding.

Code, Data and Checkpoints

https://sreyan88.github.io/compa_iclr/