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

Sreyan Ghosh♠♦* Ashish Seth♠* Sonal Kumar♠* Utkarsh Tyagi♠* Chandra Kiran Evuru♠*
S.Ramaneswaran♥ S. Sakshi♠ Oriol Nieto♦ Ramani Duraiswami♠ Dinesh Manocha♠
♠University of Maryland, College Park, ♥NVIDIA, Bangalore, India, ♦Adobe, USA

## Understanding Compositional Reasoning in ALMs

- **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, which allows them to solve audio tasks through a language interface.
- **What is Compositional Reasoning?** Compositional Reasoning, characterized as the ALM's capability 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.

*The extent to which ALMs can perform compositional reasoning is largely under-explored. Our work aims to bridge this gap by evaluating and improving compositional reasoning in ALMs.*

## Motivation: Why are current benchmarks insufficient for evaluating compositional reasoning in ALMs?

**Rethinking Evaluation of Compositional Reasoning in ALMs**

- Current retrieval benchmarks are insufficient in evaluating the compositional reasoning of ALMs.
- Figure 1 shows CLAP undergoes only minor degradation in retrieval performance when the word order in captions is shuffled.
- Previous studies also show that ALMs often act as a bag of words and lack natural language comprehension.

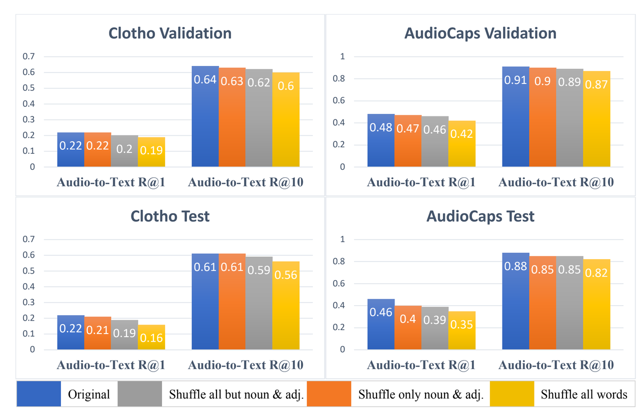
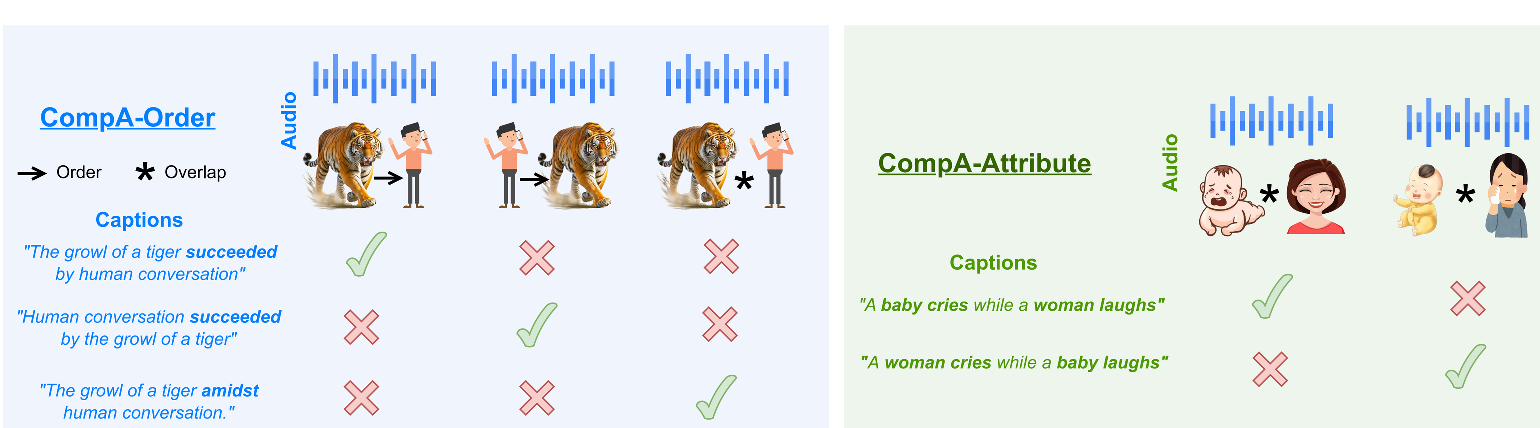



Figure 1. Performance on common retrieval evaluation datasets with shuffling.

## CompA-order/attribute: A Novel Benchmark for evaluating Compositional Reasoning in ALMs




In this work, we perform the first systematic study for understanding compositional reasoning capability in ALMs. We propose two expert-annotated benchmarks, **CompA-order** and **CompA-attribute**. While **CompA-order** is used to evaluate the ALMs ability to understand the order of occurrence between two acoustic events in a audio, **CompA-attribute** is used to evaluate the models' ability to understand attribute-binding for acoustic events.

## CompA-661K: A balance dataset for learning Compositional Reasoning in ALMs

- There is an acute scarcity of compositional audios in large audio-text pre-training datasets.
- To address this issue, We introduce a **CompA-661k** dataset, with ≈661k unique audio-caption pairs, which have a uniform distribution of audios with a number of unique acoustic events as compared to the previously used training datasets.

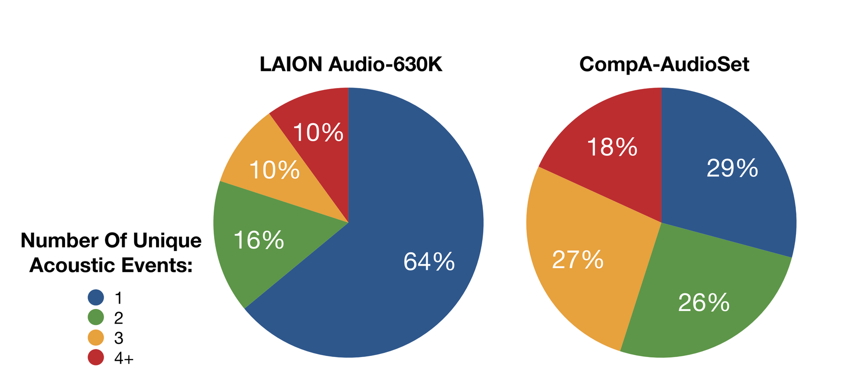



Figure 2. Distribution of audios with number of unique acoustic events: LAION-Audio-630k Vs CompA-AudioSet

## CompA-CLAP: Contrastive Pre-Training with Compositional Aware Hard Negatives

**Motivation:** To teach the ALMs compositional reasoning, we modify the vanilla contrastive learning objective and introduce compositionally aware hard negative captions for each audio in the batch.

- Each audio sample in the training batch is paired with hard negative captions (generated using GPT4) 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.

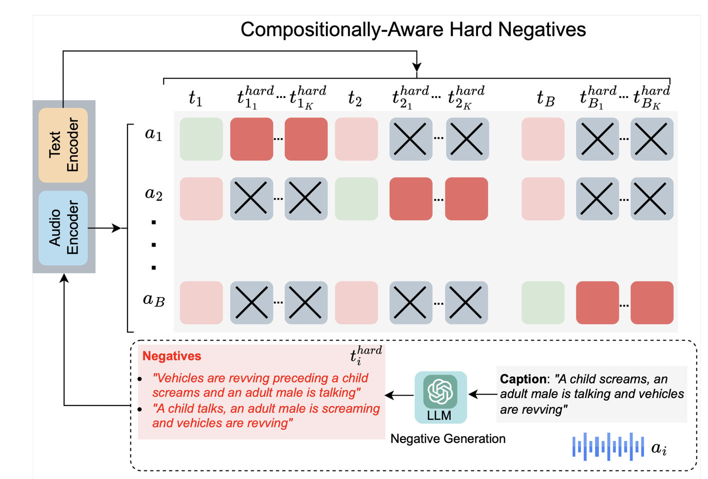



Figure 3. Contrastive training with hard negatives

Training Objective Function:

$$\mathcal{L}^{S_1} = \frac{1}{2B}\sum_{i=1}^{B}\left(\alpha_1 \ell_i^{t2a} + \alpha_2 \ell_i^{a2t}\right)$$

$$\ell_i^{t2a} = -t_i^\top a_i/\sigma + \log\sum_{j=1}^{B}\exp\left(t_i^\top a_j/\sigma\right)$$

$$\ell_i^{a2t} = -a_i^\top t_i/\sigma + \log\left(\sum_{j=1}^{B}\exp\left(a_i^\top t_j/\sigma\right) + \sum_{k=1}^{K}\exp\left(a_i^\top t_{i_k}^{\text{hard}}/\sigma\right)\right)$$

## Results: Zero-Shot evaluation on standard benchmarks

| Model | T-A Retrieval R@1 | T-A Retrieval R@5 | T-A Retrieval R@10 | A-T Retrieval R@1 | A-T Retrieval R@5 | A-T Retrieval 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 |

Table 1. Result comparison on retrieval benchmarks (AudioCap/Clotho)

Table 1, 2 shows the performance comparison of CompA-CLAP with baselines on benchmark datasets. While our CLAP achieves SoTA performance in almost all cases, CompA-CLAP retains its performance even after fine-tuning for compositionality.

| | 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-CompA-661k *(ours)* | **90.2** | **86.1** | 29.1 | **77.8** |
| CompA-CLAP *(ours)* | 89.1 | 85.7 | **29.5** | 77.4 |

Table 2. Result comparison on audio classification benchmarks.

## Results: Evaluation on CompA-order/attribute benchmarks

Table 3 compares the results of CompA-CLAP on CompA-order/attribute benchmarks. Our **vanilla CLAP** performs better than all other baselines from literature, outperforming CLAP-LAION by **≈6%-33%** over both benchmarks. **CompA-CLAP**, which is CLAP trained consecutively with hard negatives and modular contrastive learning, improves performance on both benchmarks by **≈10%-28%** over CLAP.

| Model | CompA-order Text | CompA-order Audio | CompA-order Group | CompA-attribute Text | CompA-attribute Audio | CompA-attribute 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 | 6.85 | 3.90 | 29.59 | 4.69 | 3.12 |
| ML-ACT | 21.85 | 8.00 | 4.35 | 31.63 | 5.11 | 3.75 |
| CLAP | 22.80 | 8.35 | 4.70 | 33.27 | 6.14 | 4.66 |
| CLAP-LAION | 24.0 | 9.25 | 5.50 | 34.78 | 6.52 | 5.07 |
| CompA-CLAP *(ours)* | **40.70** | **35.60** | **33.85** | **44.28** | **22.52** | **15.13** |
| - Hard Negative | 36.25 | 31.45 | 20.20 | 39.27 | 17.71 | 11.35 |
| - Modular Contrastive | 38.0 | 33.50 | 21.25 | 43.48 | 19.57 | 13.04 |
| CLAP *(ours)* | 33.75 | 15.75 | 11.50 | 42.40 | 20.50 | 14.75 |

Table 3. Result comparison on our proposed CompA benchmarks

## CompA-CLAP: Modular Contrastive Learning for Fine-grained Understanding

**Motivation:** Contrastive Pretraining with hard negatives still requires compositional audios and their corresponding captions. Further, an audio with a large number of acoustic events makes fine-grained learning difficult. To overcome these issues, we propose a Template-based algorithm for creating compositionally rich audio-caption creation. Next, we propose Modular Contrastive training for fine-grained understanding

**Template-based synthetic creation of audio-caption pairs**

- We propose a simple and scalable template-based approach to create compositional audio
- An LLM first generates a scene from a pool of available acoustic events from which we perform simple operations to generate compositional audio and their captions.

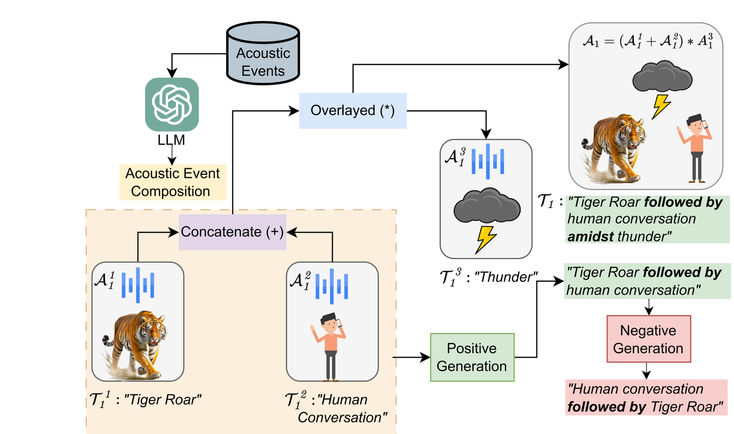



Figure 4. Illustration of template-based audio synthesis

**Modular Contrastive Learning**

- Our proposed Modular Contrastive training employs multiple positives and negatives for each audio, generated using a **template-based algorithm**.
- Each positive describes compositional relationships of various granularities in the audio and this helps the model learn fine-grained order and attribute binding.

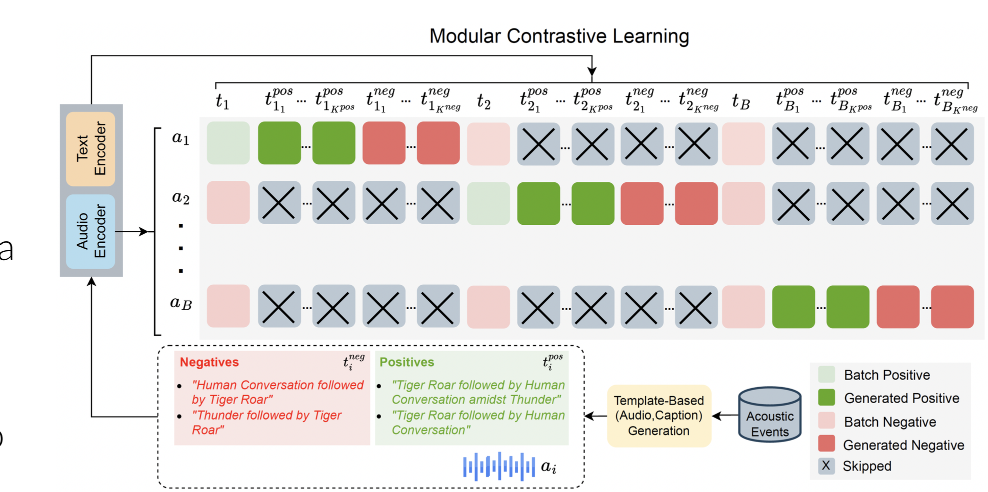



Figure 5. Illustration of Modular Contrastive training with multiple positive and negative caption

Training Objective Function:

$$\mathcal{L}^{S_2} = \frac{1}{2B}\sum_{i=1}^{B}\left(\beta_1 \ell_i^{t2a} + \beta_2 \ell_i^{a2t}\right)$$

$$\ell_i^{t2a} = -\left(\frac{1}{K^{pos}}\sum_{k=1}^{K^{pos}}\left(t_{i_k}^{pos}\right)^\top a_i/\sigma\right) + \log\sum_{j=1}^{B}\exp\left(t_i^\top a_j/\sigma\right)$$

$$\ell_i^{a2t} = -\left(\frac{1}{K^{pos}}\sum_{k=1}^{K^{pos}} a_i^\top t_{i_k}^{pos}/\sigma\right) + \log\left(\sum_{j=1}^{B}\exp\left(a_i^\top t_j/\sigma\right) + \sum_{k=1}^{K^{neg}}\exp\left(a_i^\top t_{i_k}^{neg}/\sigma\right)\right)$$

**Notations:** $\ell^{t2a}$, $\ell^{a2t}$ is the contrastive losses for text and audio respectively. $(t_{i_k}^{hard})_{k\in[1,K]}$ is the $k^{th}$ negative caption for audio sample $a_i$. $(t_{i_k}^{pos})_{k\in[1,K^{pos}]}$ and $(t_{i_k}^{neg})_{k\in[1,K^{neg}]}$ are $k^{th}$ generated fine-grained positive and negative caption for audio sample $a_i$. $\beta_1$ and $\beta_2$ are scaling parameters.

## Evaluation Metric: For evaluating CompA-order/attribute

Given two audios $A_0$ and $A_1$ and their corresponding captions $C_0$ and $C_1$, we define a text score $f(\cdot)$ and an audio score $g(\cdot)$ w.r.t the ALMs capability to select texts given audio and audios given text respectively. We also define a group score $h(\cdot)$, combining text and audio scores.

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

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

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

## Future Work

- **Expand the CompA Benchmarks:** Introduce more complex compositional scenarios to further push ALMs capabilities.
- **Refine Training Techniques:** Continue to develop 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