SEOUL NATIONAL UNIV.
VISION & LEARNING

ACL 2025 VIENNA

Can LLMs Deceive CLIP? Benchmarking Adversarial Compositionality of Pre-trained Multimodal Representation via Text Updates

Jaewoo Ahn* Heeseung Yun* Dayoon Ko Gunhee Kim (* Equal contribution)

Jaewoo Heeseung Code

Yes, LLMs *can* deceive ANY "X-Language" Models ({image, video, audio}↔TEXT) and do so *even* better with diversity-promoting self-training!

Motivation

Pre-trained multimodal representations are everywhere, utilized in a wide range of downstream applications
e.g., CLIP, CLAP, VideoCLIP, LanguageBind, etc.

However, they are known to be considerably **brittle**:

Objective
Deceive {CLIP, LBVideo, LBAudio, ⋮} via text updates

A baby is sitting on a bed and reaching for the laptop keys.

Input: {image x, text t}

Generation Phase

(a) Sample 1 ($\tilde{t}^1$)
*A baby is sitting on a **vintage** bed and reaching for the laptop keys.*

(b) Sample 2 ($\tilde{t}^2$)
*A **bed** is sitting on a **baby** and reaching for the laptop keys.*

(c) Sample 3 ($\tilde{t}^3$)
*A baby is sitting on a bed and **accidentally typing an email**.*

Evaluation Phase

Deceive ✗ / Diversity ↓
- ✗ CLIP$(x, t) >$ CLIP$(x, \tilde{t}^1)$
- ✗ $(t, \tilde{t}^1)$ entailed
- ↓ word *vintage* overused

Deceive ✓ / Diversity ↓
- ✓ CLIP$(x, t) <$ CLIP$(x, \tilde{t}^2)$
- ✓ $(t, \tilde{t}^2)$ not entailed
- ↓ operation *swap* overused

Deceive ✓ / Diversity ↑
- ✓ CLIP$(x, t) <$ CLIP$(x, \tilde{t}^3)$
- ✓ $(t, \tilde{t}^3)$ not entailed
- ↑ diverse words & phrases

How to address such vulnerabilities in these embeddings in a modality-agnostic manner through the lens of compositionality+? (+ Structured relationship between words and elements)

→ **MAC** (**M**ultimodal **A**dversarial **C**ompositionality)

- **Comparison with Existing Frameworks & Benchmarks**

Method	Modality	Generation	Crossmodal	Diversity
FOIL[1]	Image	Rule-based	✓	
Winoground[2]	Image	Human	✓	
SugarCrepe[3]	Image	ChatGPT	✓	
VIOLIN[4]	Video	Human	✓	
VideoCon[5]	Video	PaLM-2	✓	
CompA[6]	Audio	GPT-4	✓	
MAC	Image, Video, Audio	Llama3-8B	✓	✓

Crossmodal: Evaluate whether a generated sample achieves the intended attack ($(x_i, t_i) \prec (x_i, \tilde{t}_i)$)
Diversity: Evaluate the diversity of a set of generated samples ($H = -\sum_j p_j \log p_j$)

Solution : [1])MAC & [2])Diversity-prompting Self-training

[1])Modality-agnostic comprehensive eval & [2])Self-train + Large-N distilled + Gibbs sampling-based diverse train data selection

Dataset $\mathcal{D}$

A baby is sitting on a bed and reaching for the laptop keys.
Text t_i

Input x_i

Target Model (*e.g.*, CLIP)

Language Model

	Cross	Uni	Dist	Aux	Div
$\tilde{t}_i^1$	✓	×	✓	✓	+.012
$\tilde{t}_i^2$	✓	✓	✓	✓	+.015
$\tilde{t}_i^3$	×	✓	✓	✓	−.008
⋮					
$\tilde{t}_i^N$	×	×	✓	×	+.002

Deception Criteria

Crossmodal	Unimodal	Distance	Auxiliary
$(x_i, t_i) \prec (x_i, \tilde{t}_i)$	$(t_i, \tilde{t}_i)$ not entailed	$d_E(t_i, \tilde{t}_i) < \frac{L}{2}$	Satisfy prompt $\mathcal{I}$

Prompt $\mathcal{I}$

$\{t_i, \{\tilde{t}_i^n\}_{n=1}^N\}_{i=1}^{M_\mathcal{D}}$
Original & generated texts

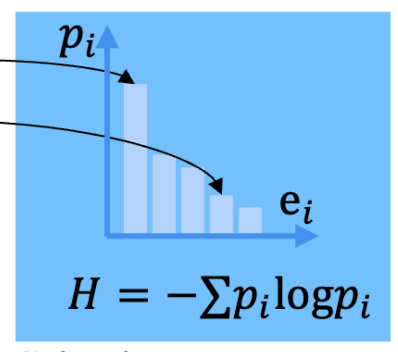
Filtering → Language Model (LoRA) → *A baby is sitting on a bed and **accidentally typing an email**.* Sample $\tilde{t}_i$

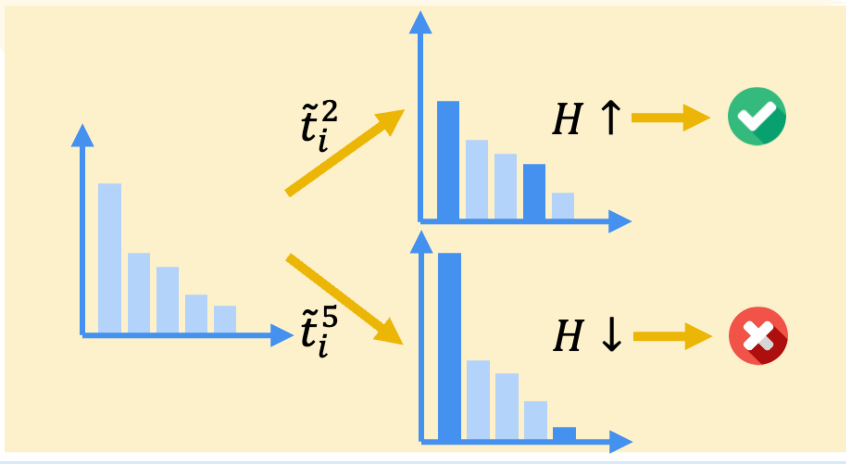
Multi-round

$\tilde{t}_i$

Diversity Criteria

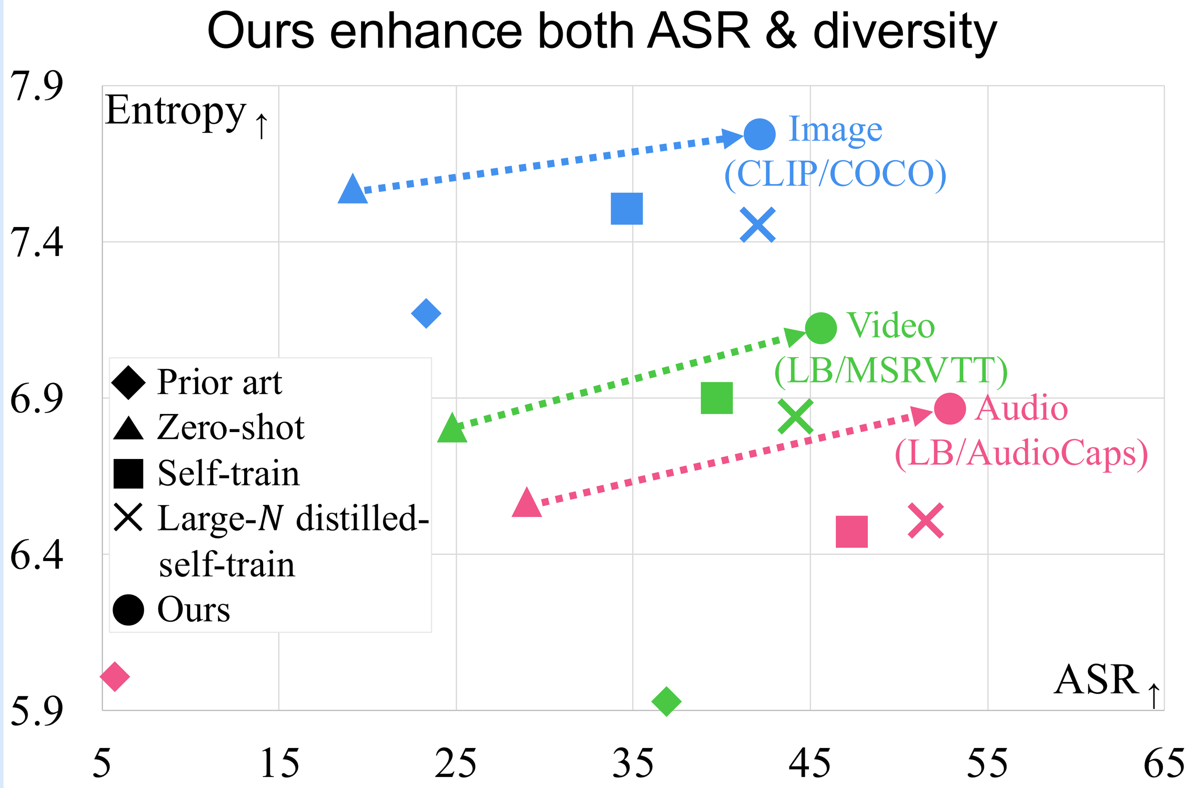
$\tilde{t}_1$ I_ADJ_vintage, …
$\tilde{t}_2$ I_NOUN_ski, …
⋮
$\tilde{t}_{M_\mathcal{D}}$ D_NOUN_man, I_ADJ_vintage, …



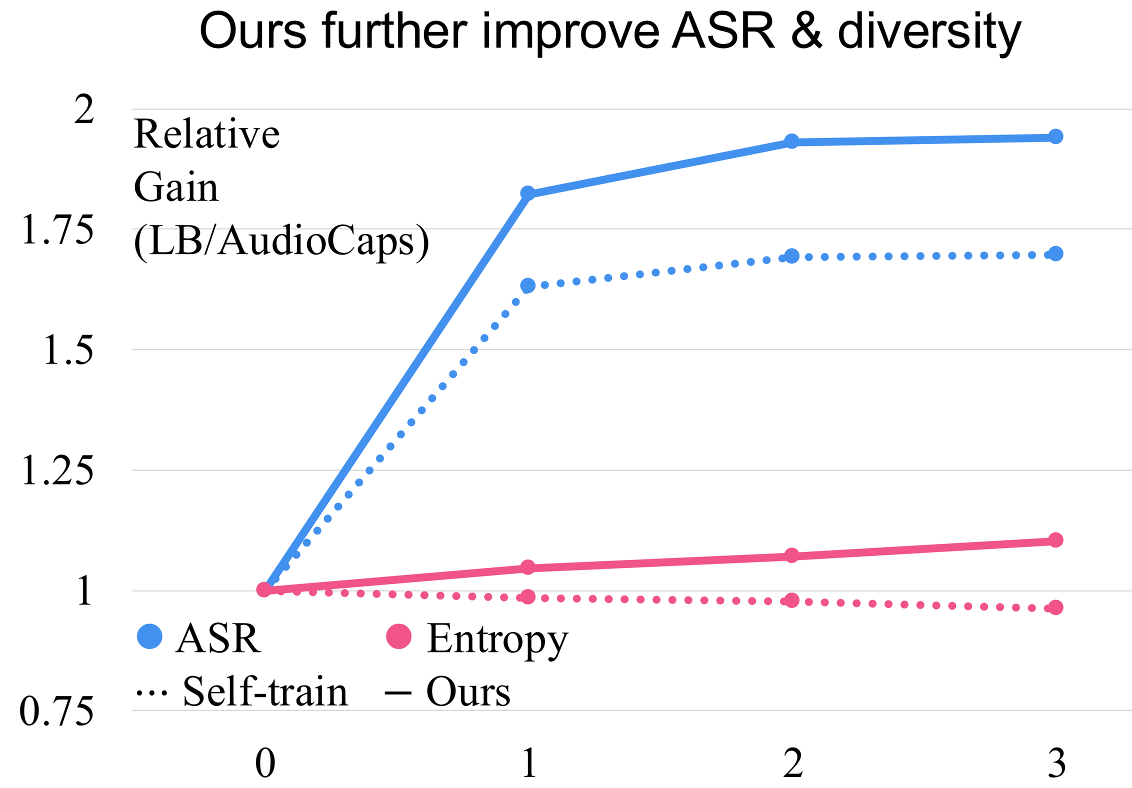


Experiments

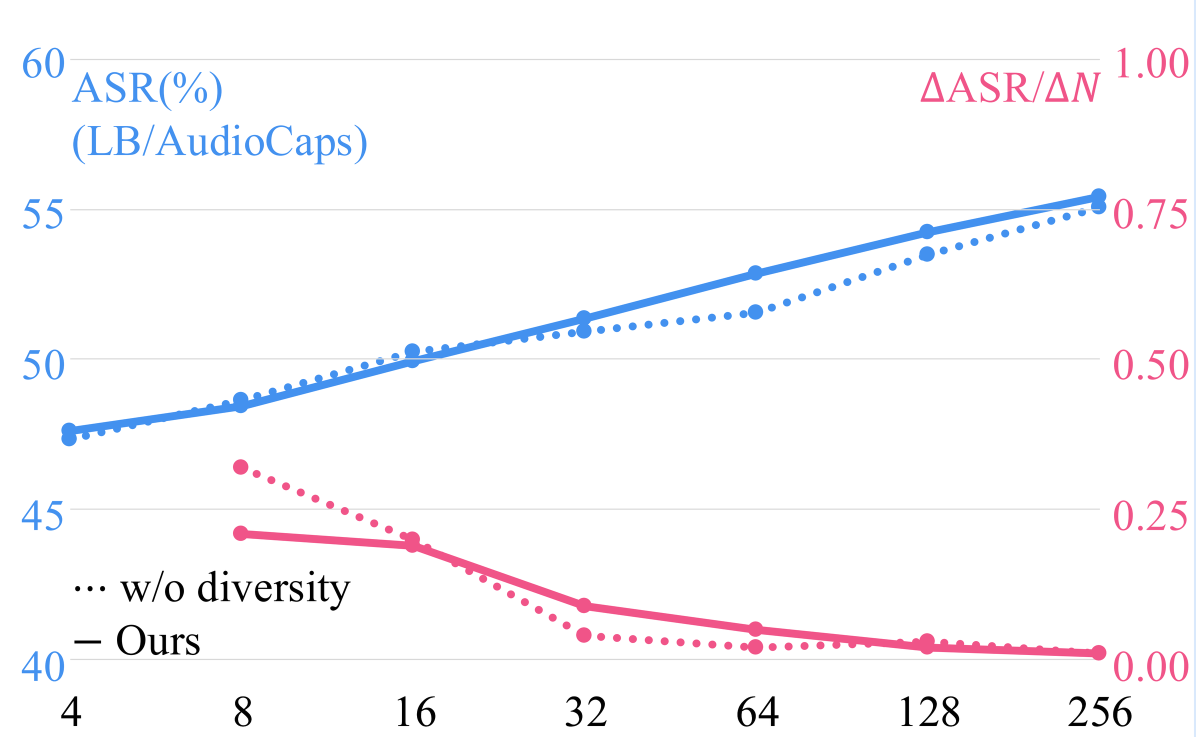
Comparison with Prior Arts:
Ours enhance both ASR & diversity



Influence of Self-training Iterations:
Ours further improve ASR & diversity



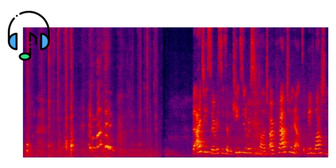
Influence of Self-training Sample N:
$N = 64$ offers a reasonable balance



Qualitative Examples

A person is looking at a camera during a wrestling event

Cross	Uni	Dist	
✗	✗	✓	**(Zero-shot)** A person is intensely staring at a camera during a dramatic wrestling event
✗	✓	✓	**(Self-train)** A person is smiling at a camera during a wrestling event
✓	✓	✓	**(Ours)** A person is looking directly at the referee during a wrestling event



A female speaking with some rustling followed by another female speaking

Cross	Uni	Dist	
✗	✓	✗	**(Zero-shot)** The female is speaking with some rustling but the other voice is a male
✗	✓	✓	**(Self-train)** A female speaking with some rustling, followed by a male speaking
✓	✓	✓	**(Ours)** A female speaking with some rustling followed by the same female speaking again

[1] Shekhar et al. Find one mismatch between image and language caption. ACL 2017.
[2] Thrush et al. Probing vision and language models for visio-linguistic compositionality. CVPR 2022.
[3] Hsieh et al. Fixing hackable benchmarks for vision-language compositionality. NeurIPS 2023.
[4] Liu et al. A large-scale dataset for video-and-language inference. CVPR 2020.
[5] Bansal et al. Robust video-language alignment via contrast captions. CVPR 2024.
[6] Ghosh et al. Addressing the gap in compositional reasoning in audio-language models. ICLR 2024.