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

ACL 2025
VIENNA

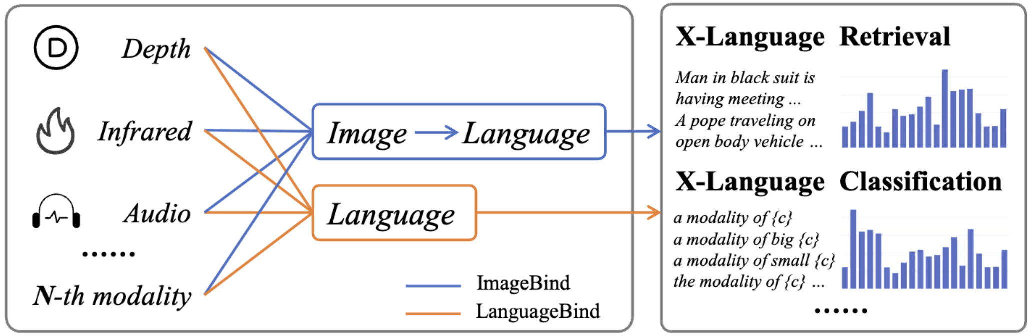
(* Equal Contribution)
Jaewoo Ahn*, Heeseung Yun*, Dayoon Ko, Gunhee Kim

SEOUL NATIONAL UNIV.
VISION & LEARNING

Intro: Pre-trained Multimodal Representations

- They encode rich information from **different modalities**
  - Cross-modality
    - *Image-Language*: CLIP, SigLIP, BLIP, ...
    - *Video-Language*: VideoCLIP, Frozen in Time, ...
    - *Audio-Language*: CLAP, ...
  - Multimodality
    - *{Language, Video, Audio, Depth, Thermal, IMU}-Image*: ImageBind, ...
    - *{Video, Audio, Depth, Thermal}-Language*: LanguageBind, ...
- Widespread applications across
  - *retrieval*
  - *generation*
  - *reward modeling*

  ...



(image credit: LanguageBind)

Intro: Compositional Vulnerability

- Contrary to the belief, these representations are known to be brittle
  - Intuitively exemplified by **compounding text elements**
  - *e.g.*, CLIP got confused by *simple negation* or *object swapping*
  - Usually explored in *vision-language compositionality* [1,2,3] domain

A man sitting on a bench next to a horse

(Negation) A man **not** sitting on a bench next to a horse
(Swap) A man sitting on a **horse** next to a **bench**
(Replace) A man standing on a bench next to a horse
(Add) A man sitting on a bench next to a horse, **while drinking a glass of water**
…

[1] Thrush et al., *Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality*, CVPR 2022
[2] Ma et al., *CREPE: Can Vision-Language Foundation Models Reason Compositionally?*, CVPR 2023
[3] Bansal et al., *VideoCon: Robust Video-Language Alignment via Contrast Captions*, CVPR 2024

Motivation: “Diverse” Compositional Vulnerabilities

- Existing studies are limited to specific modalities
  - (Mostly) *Image-Language compositionality* [1]
  - *Video-Language compositionality* [2]
  - (Few) *Audio-Language compositionality* [3]
- They usually assume specific scenarios (*negation*, *swap*, …)

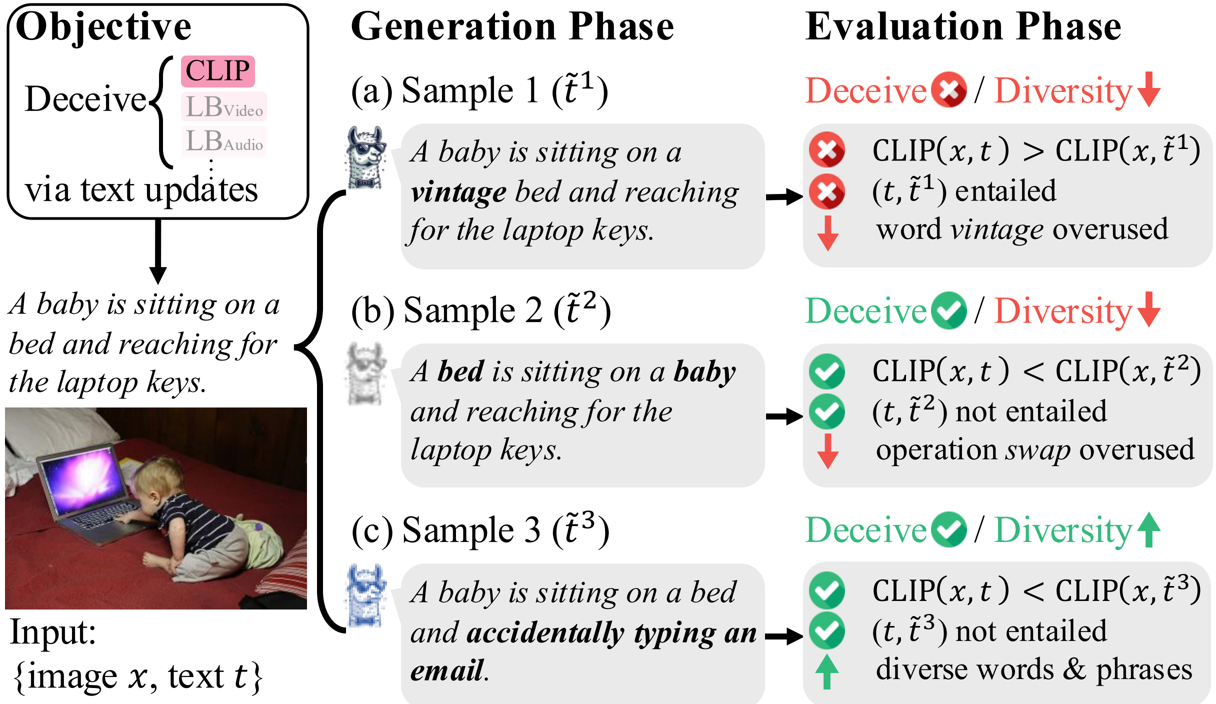
→ Comprehensive understanding of **diverse** compositional vulnerabilities, **without assuming specific scenarios**, remain an open challenge

[1] Yuksekgonul et al., *When and why vision-language models behave like bags-of-words, and what to do about it?*, ICLR 2023
[2] Park et al., *Exposing the Limits of Video-Text Models through Contrast Sets*, NAACL 2022
[3] Ghosh et al., *CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models*, ICLR 2024

The MAC Benchmark

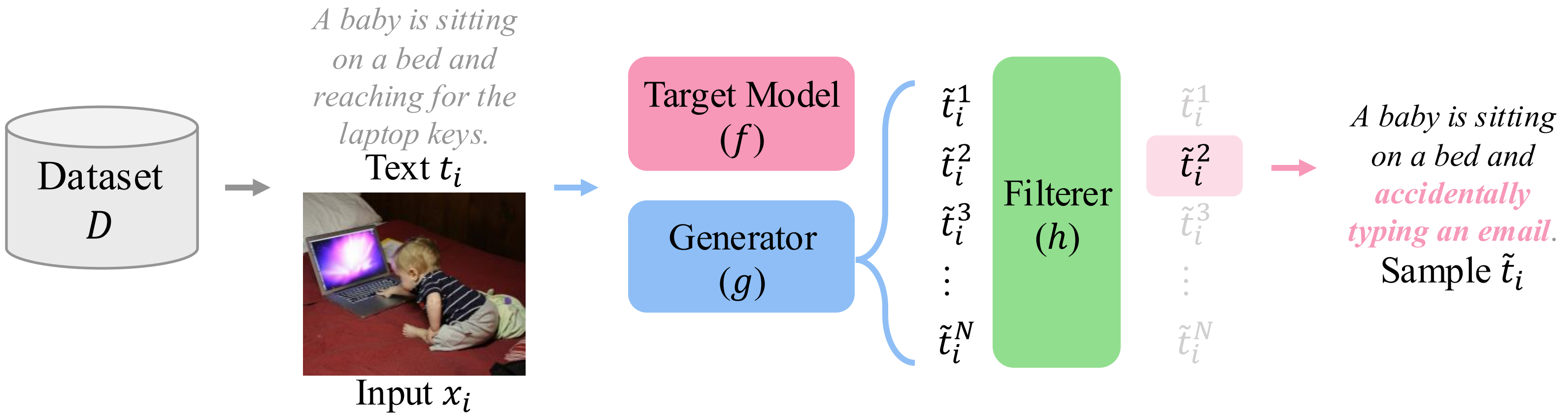
- We propose **M**ultimodal **A**dversarial **C**ompositionality (**MAC**) benchmark
  - Given multimodal data pairs (*e.g., image-caption*)

1. **Generate** deceptive captions (*via rule-based, LLM, etc*)
2. **Evaluate** whether generated captions successfully deceive target representations
   - Sample-wise Eval
   - Group-wise Eval



The MAC Benchmark: Problem Definition

1. Generation
   - We use **text updates** due to **modality-agnostic** assessment
   - Given a set of paired data $D = (t_i, x_i)_{i=1}^{M_D}$, we aim to generate a set of adversarial text $\{\widetilde{t_i}\}_{i=1}^{M_D}$ that deceives a target representation f, which encodes both t_i and x_i into embeddings $y_{t_i}, y_{x_i} = f(t_i, x_i) \in \mathbf{R}^d$
   - Two key components
     - Adversarial sample "generator" g produces N samples $\{\tilde{t}_i^n\}_{n=1}^N$
     - Sample "filterer" h identifies a single $\tilde{t}_i$ from N candidates



The MAC Benchmark: Problem Definition

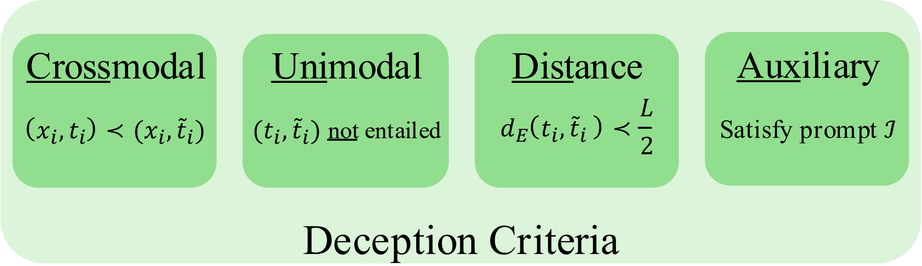
2. Evaluation

- **Sample-wise Deception Evaluation**

  1. Cross-modal criterion ($\boldsymbol{s_i^c}$): Generated sample should achieve the intended attack
  2. Uni-modal criterion ($\boldsymbol{s_i^u}$): Meaningful semantic distinction btw generated & original text
  3. Distance criterion ($\boldsymbol{s_i^d}$): Only limited lexical deviation from the original sample
  4. Auxiliary criterion ($\boldsymbol{s_i^a}$): Whether a generated sample follows a set of predefined rules

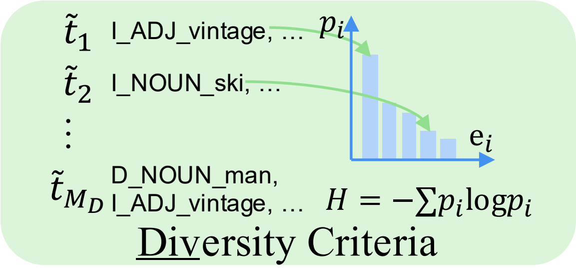
In total, the **attack success rate (ASR)** R is

$$R = \frac{1}{M_D} \sum_i (s_i^c, s_i^u, s_i^d, s_i^a)$$



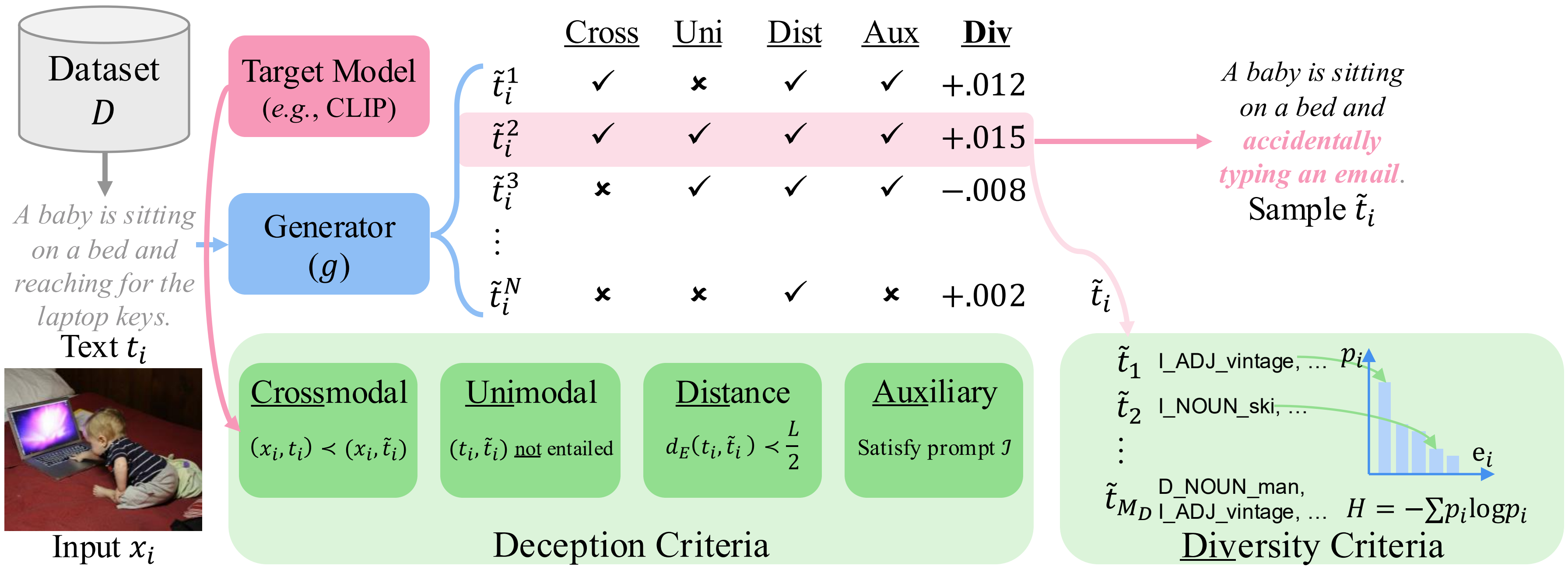
The MAC Benchmark: Problem Definition

2. Evaluation
   - **Group-wise Diversity Evaluation**
     - Another crucial criterion? $\longrightarrow$ **Diversity**
       - Repeated & similar attack is easily defendable & lacks generalizability
     - First, construct a set of "attribute-enhanced token" e_i^j, defined as OP_POS_LEMMA
     - Using a set of tokens, compute **entropy** $H = -\sum_j p_j \log p_j$
       - Additionally, we use **distinct-1**= $\frac{\#\,\text{unique attribute–enhanced tokens}}{\#\,\text{all attribute–enhanced tokens}}$



The MAC Benchmark: Overall

- Key advantages
  - **Modality-agnostic:** Applied to any formats (image, video, audio)
  - **Leaderboard:** Existing compositionality frameworks can be consistently compared
  - **Comprehensive eval:** Assess both deception and attack diversity



Approach: Motivation

- Among diverse **generators** $\boldsymbol{g}$, we prioritize **LLM-based** methods
  - Rule-based
    - produce nonsensical & non-fluent text
  - Human-based
    - difficult to scale
  - **LLM-based**
    - generate fluent text at scale
    - Recent studies adopted
      LLM-based > Rule & Human-based

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

[1] Park et al., *RoCOCO: Robustness Benchmark of MS-COCO to Stress-Test Image-Text Matching Models*, ECCV 2024 Workshop
[2] Thrush et al., *Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality*, CVPR 2022
[3] Hsieh et al., *SugarCrepe: Fixing Hackable Benchmarks for Vision-Language Compositionality*, NeurIPS 2023
[4] Liu et al., *Violin: A Large-Scale Dataset for Video-and-Language Inference*, CVPR 2020
[5] Park et al., *VideoCon: Robust Video-Language Alignment via Contrast Captions*, CVPR 2024
[6] Ghosh et al., *CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models*, ICLR 2024

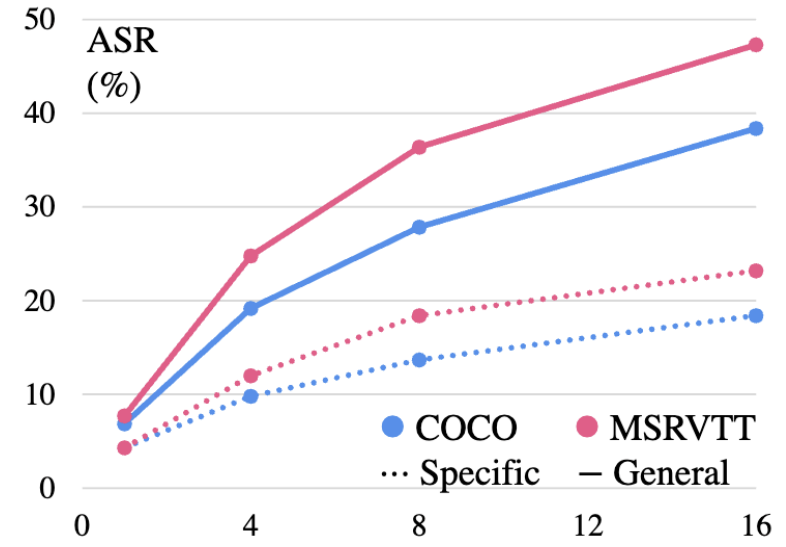
Approach: Preliminary

- Revealing Compositional Vulnerabilities via **Filtering $\boldsymbol{f}$**
  - Multiple attempt ($N > 1$): effective than $N = 1$
  - Best-of-N sampling
    - Given N samples $\{\tilde{t}_i^n\}_{n=1}^N$, sample deceptive one first; otherwise randomly sample

$$T_i = \left\{\tilde{t}_i^n \middle| \left(s_i^c \cdot s_i^u \cdot s_i^d \cdot s_i^a\right)(\tilde{t}_i^n, t_i, x_i) = 1\right\},$$

$$\tilde{t}_i \sim \begin{cases} \text{Uniform}(T_i), & if\ T_i \neq \emptyset, \\ \text{Uniform}(\{\tilde{t}_i^n\}_{n=1}^N), & \text{otherwise.} \end{cases}$$

- Pros: ASR ↑
- Cons:
  - computational cost scales linearly with N
  - Larger N masks true effectiveness of adversarial strategies (*i.e.*, brute-force)



[ASR (%) vs. Best-of-N sampling]

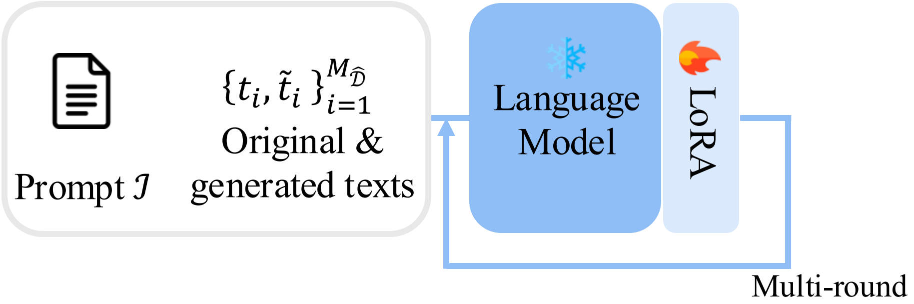
Approach: Self-training

- We propose a **learnable** method for the first time
  - Given the absence of ground-truth, we employ **self-training**
    - or **rejection sampling fine-tuning (RFT)**
  - From the training set $D_{\text{train}} = (t_i, x_i)_{i=1}^{M_{D_{\text{train}}}}$,
    - We first generate & filter samples $\{\tilde{t}_i^n\}_{n=1}^N$ using best-of-N sampling
    - Then only use $M_{\widehat{D}}$ successful adversarial samples to train the model

$$\{\tilde{t}_i\}_{n=1}^{M_{\widehat{D}}} = \{\tilde{t}_i^n \,|\, s_i^c \cdot s_i^u \cdot s_i^d \cdot s_i^a = 1\},$$

$$\mathcal{L} = -\frac{1}{M_{\widehat{D}}} \sum_i \sum_j \log g(\tilde{t}_{i,j} \,|\, \tilde{t}_{i,<j}, I, t_i; \Theta)$$

+ To further enhance ASR, we use large-N distilled self-training

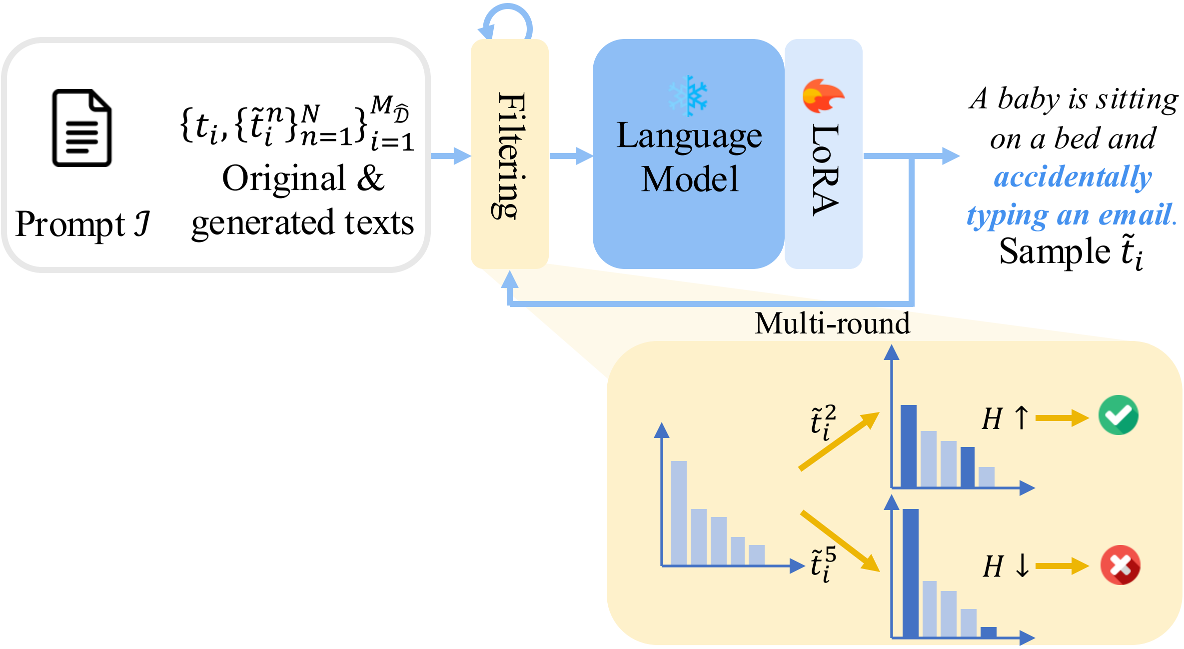


Approach: "Diversity-promoting" Self-training

- Despite high ASR, naïve self-training decreases diversity
- To enhance diversity:
  - Introduce **Gibbs sampling-based train data selection**
  - Motivation: Iteratively selects samples that maximizes diversity

Algorithm 1 Diversity-promoting Self-training Data Selection

Require: Set of N samples $\{\tilde{t}_i^n\}_{n=1}^N$ generated for each training instance $i \in [1, M_{\hat{\mathcal{D}}}]$, and diversity function H
Ensure: Diverse successful samples $\{\tilde{t}_i\}_{i=1}^{M_{\hat{\mathcal{D}}}}$
  Initialize $\{\tilde{t}_i\}_{i=1}^{M_{\hat{\mathcal{D}}}}$ randomly from $\{\tilde{t}_i^n | (s_i^c \cdot s_i^u \cdot s_i^d \cdot s_i^a)(\tilde{t}_i^n, t_i, x_i) = 1\}$
  for iteration $k = 1$ to K **do**
    for $i = 1$ to $M_{\hat{\mathcal{D}}}$ **do**
      $\mathcal{T}_i \leftarrow \{\tilde{t}_i^n | (s_i^c \cdot s_i^u \cdot s_i^d \cdot s_i^a)(\tilde{t}_i^n, t_i, x_i) = 1\}$
      $\tilde{t}_i \leftarrow \mathrm{argmax}_{\tilde{t}_i^n \in \mathcal{T}_i} H(\tilde{t}_1, ..., \tilde{t}_i^n, ..., \tilde{t}_{M_{\hat{\mathcal{D}}}})$
    end for
  end for
  return $\{\tilde{t}_i\}_{i=1}^{M_{\hat{\mathcal{D}}}}$



Experiments: Evaluation Protocol

- Target representations: CLIP, LanguageBind
  - + SigLIP, NegCLIP, BLIP, CLAP, LLaVA
- Source datasets: MS-COCO (image), MSRVTT (video), AudioCaps (audio)
- Generator LLM: Llama-3.1-8B
  - Prompt version
    - *deceptive-specific*: *replace*, *swap*, and *add* operations
    - *deceptive-general* (default): no constraints
- Evalulation metrics
  - Sample-wise: ASR (%)
  - Group-wise: Diversity (H)

Experiments: Results (1)

- Existing methods
  - Single modality
- ASR vs. Diversity
- ASR: $N = 4 > N = 1$
- Ablation
  - + Self-train
    - ASR ↑ (+68% on avg)
    - Reduce diversity
  - + Large-N Distilled
    - Further ASR ↑
    - Reduce diversity
  - + Diversity-Promoted
    - **Pareto front in ASR-diversity**

Method	Image (CLIP/COCO)		Video (LB/MSRVTT)		Audio (LB/AudioCaps)	
	ASR (%)	Diversity (H)	ASR (%)	Diversity (H)	ASR (%)	Diversity (H)
$N = 1$						
RoCOCO$_{\text{rand-voca}}$	1.99	<u>7.64</u>	-	-	-	-
RoCOCO$_{\text{Danger}}$	7.88	4.45	-	-	-	-
RoCOCO$_{\text{same-concept}}$	5.29	7.10	-	-	-	-
RoCOCO$_{\text{diff-concept}}$	2.75	7.13	-	-	-	-
SugarCrepe	2.40	7.31	-	-	-	-
VideoCon	-	-	7.10	6.70	-	-
Deceptive-General Prompt (zero-shot)	6.88	7.56	7.70	6.81	10.47	<u>6.57</u>
$N = 4$						
SeeTrue	23.33	7.17	-	-	-	-
VFC	-	-	36.90	5.93	-	-
CompA	-	-	-	-	5.76	6.01
(1) Deceptive-General Prompt (zero-shot)	19.19	7.57	24.80	6.81	29.02	6.57
(2): (1) + Self-Train	34.64	7.51	39.70	<u>6.90</u>	47.35	6.47
(3): (2) + Large-N Distilled	<u>42.03</u>	7.45	<u>44.20</u>	6.84	<u>51.57</u>	6.51
(4): (3) + Diversity-Promoted (ours)	**42.10**	**7.75**	**45.60**	**7.13**	**52.87**	**6.87**

Experiments: Results (2)

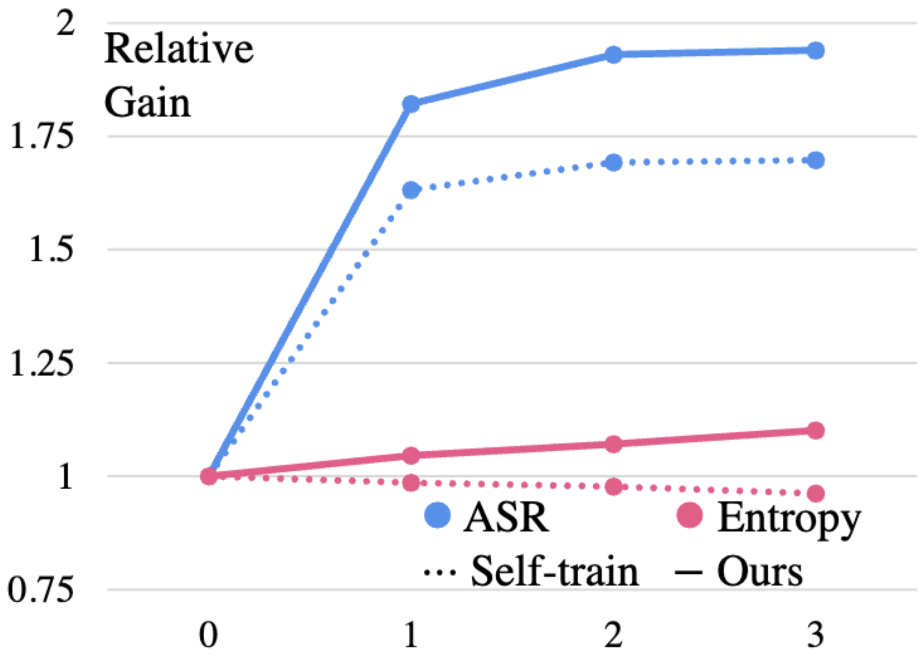
- Transferability across representations
  - High transferability, exceeding the best performing baseline (23.33)
  - Performance gains from self-training: 2.1x improvements

ASR (%)	CLIP	SigLIP	NegCLIP	BLIP
CLIP	42.10 (+22.91)	28.63 (+15.68)	24.84 (+12.71)	25.25 (+14.13)
SigLIP	29.37 (+16.13)	41.04 (+21.32)	23.84 (+12.17)	25.01 (+13.76)
NegCLIP	25.40 (+12.68)	23.63 (+11.47)	40.81 (+20.10)	23.77 (+12.33)
BLIP	19.84 (+10.60)	19.11 (+10.04)	18.02 (+8.94)	32.50 (+17.80)

[Columns: source models for filtering
Rows: target models for evaluation
Numbers in parentheses: gain from ours vs. zero-shot]

Experiments: Analysis (1)

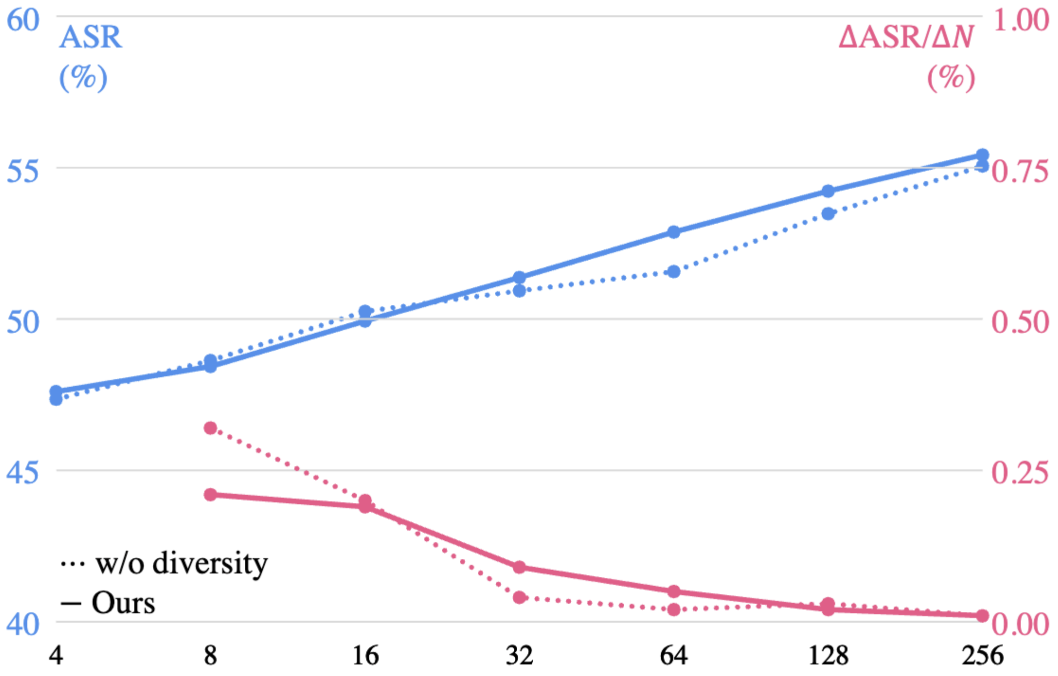
- Multi-round self-training
  - Further improves ASR, reaching saturation by 3rd round
  - Ours continuously improve diversity



[Influence of multi-round self-training]

Experiments: Analysis (2)

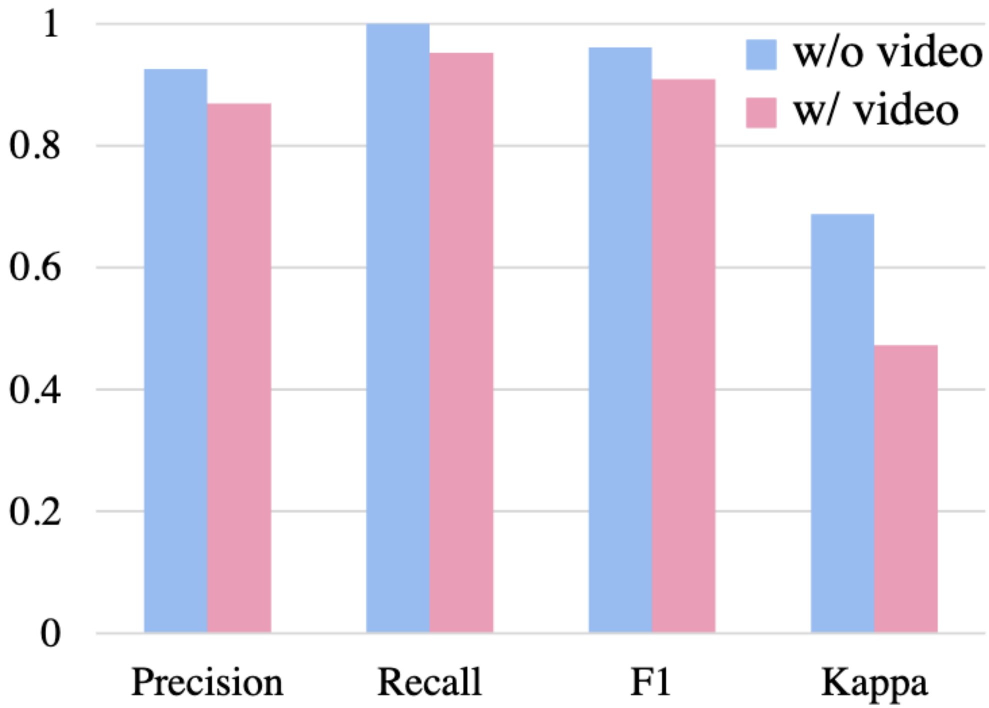
- Influence of N in large-N distilled self-training
  - Increasing N does not display a clear signal of saturation
  - Still, $\Delta ASR/\Delta N$ does $\rightarrow N = 64$ offers a reasonable balance



[Influence of N in large-N distilled self-training]

Experiments: Analysis (3)

- Human evaluation
  - Confirms reliability of evaluation of uni-modal criteria (s_i^u)



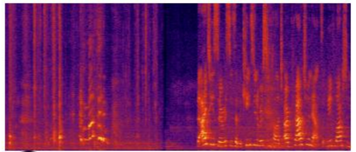
Experiments: Qualitative Examples

A lady walking in the rain carrying a pink umbrella

Cross	Uni	Dist	
✗	✓	✓	**(Zero-shot)** A lady dancing in the rain carrying a pink umbrella
✓	✓	✓	**(Self-train)** A lady walking in the rain under a broken pink umbrella
✓	✓	✓	**(Ours)** A lady walking in the rain with her pink umbrella left behind

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

Concluding Remarks

- MAC: Comprehensive testbed for evaluating compositional vulnerabilities
  - Evaluate ASR & diversity of LLM-generated outputs
  - Modality-agnostic assessment
- Diversity-promoted self-training
  - LLM-based self-training for MAC
  - Iterative RFT w/ diversity-promoting filtering: improve both ASR & diversity
- Potential extension of vulnerability analysis
  - Less-explored modalities (IMU, tactile, …)

Thank you

Code https://github.com/ahnjaewoo/MAC

Paper https://arxiv.org/abs/2505.22943

Contact jaewoo.ahn@vision.snu.ac.kr,

heeseung.yun@vision.snu.ac.kr