Université de Montréal

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

# Contrasting Intra-Modal and Ranking Cross-Modal Hard-Negatives to Enhance Visio-Linguistic Fine-grained Understanding


Le Zhang

Rabiul Awal

Aishwarya Agrawal


Mila

under review at neurips2023

# Visio-Linguistic fine-grained understanding

## Visual Genome Relation

**Assessing relational understanding (23,937 test cases)**

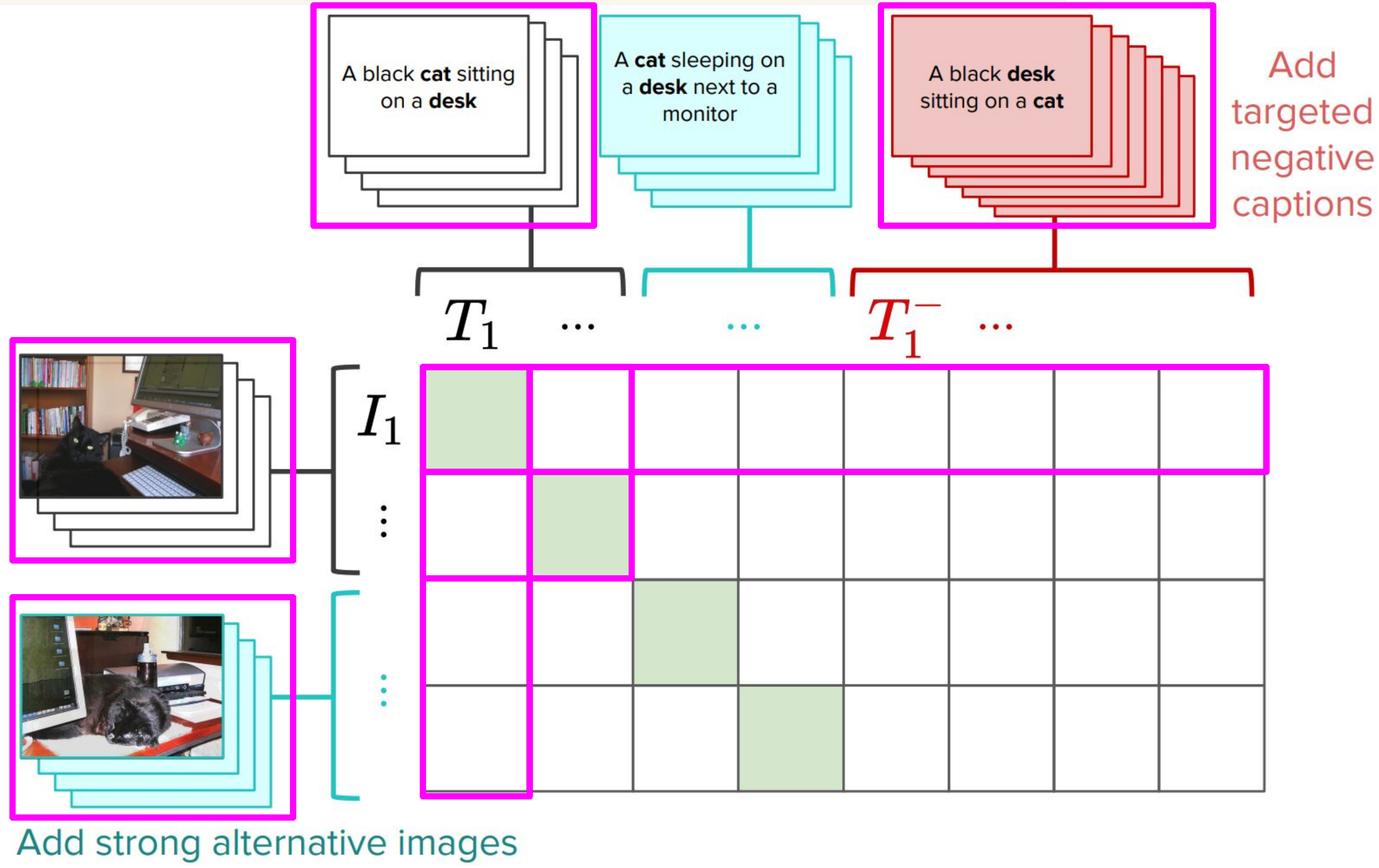
✔ *the horse is eating the grass*
**X** *the grass is eating the horse*

[Yuksekgonul et al. ICLR 2023]

# What existing approaches do

1. **Create** hard-negative sentences and retrieve similar images

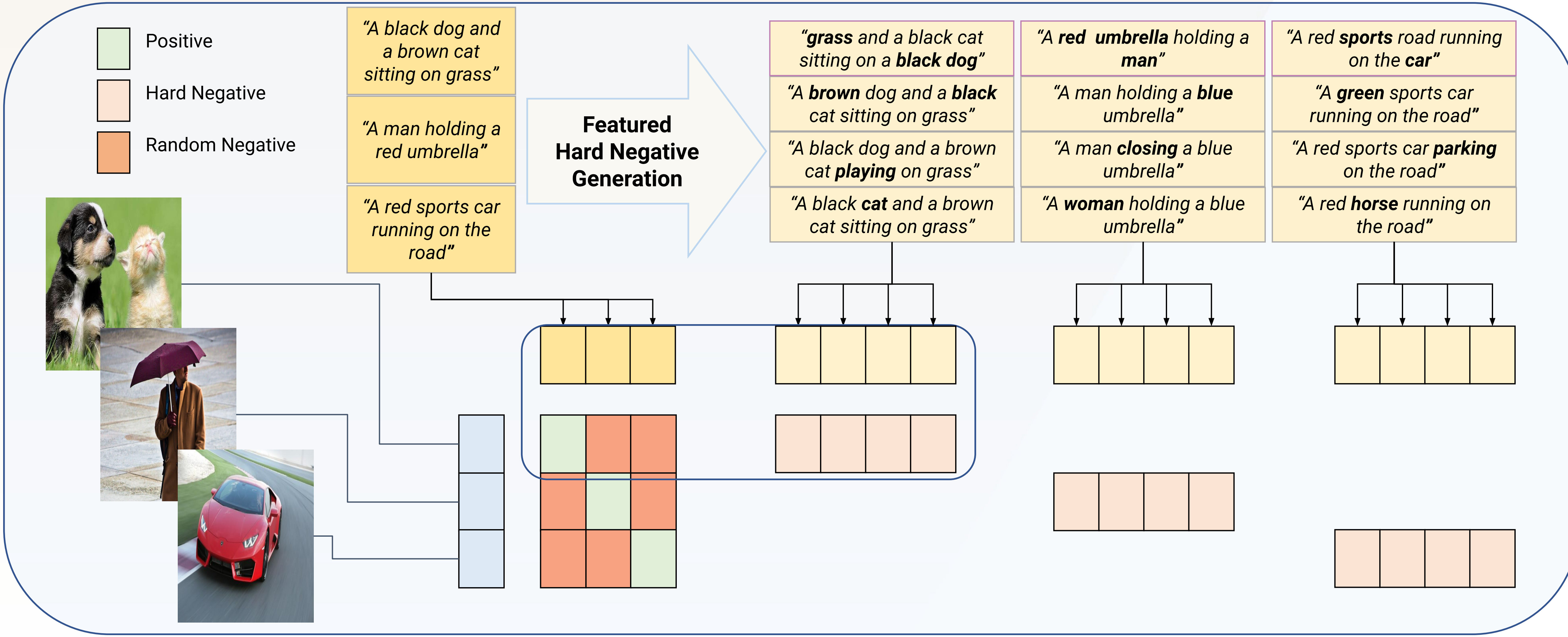
2. **Contrast** them against correct image-caption pairs



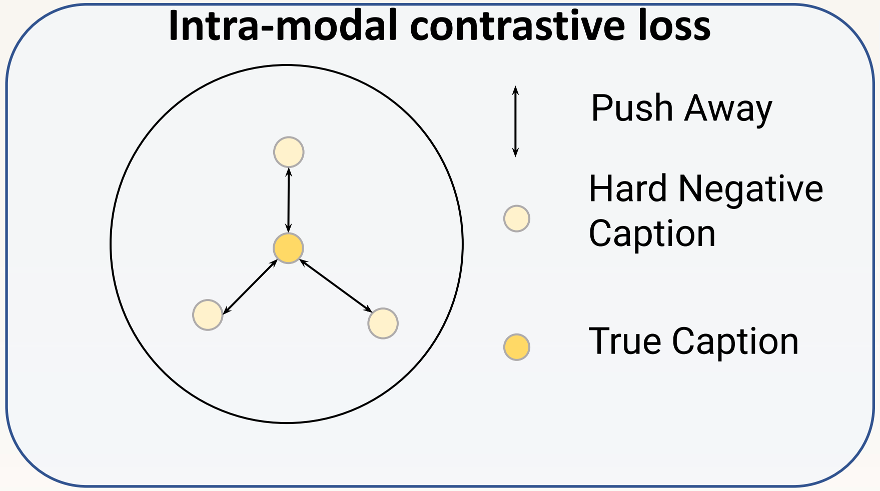

NegCLIP approach from
[Yuksekgonul et al. ICLR 2023]

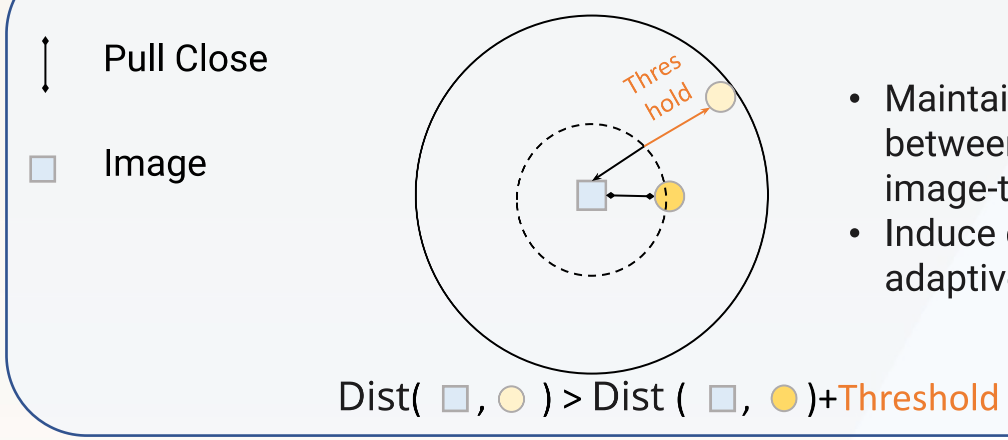
# Our approach

1. **Create** featured hard-negative sentences and images



# Our approach




**Motivation**

- Maintain a minimum distance between positive and hard-negative image-text similarities.
- Induce curriculum learning by adaptively increasing the threshold.

2. Additionally **contrast** hard-negative **sentences** against correct **sentences**

3. Add **rank loss** between correct and hard-negative image-text pairs

$$\mathcal{L}_{imc} = \sum_{(I,T)\in\mathcal{B}} -\log \frac{\exp^{S(I,T)}}{\sum\limits_{T_k\in\mathcal{T}_{hn}} \exp^{S(T,T_k)}}$$

$$\mathcal{L}_{cmr} = \sum_{(I,T)\in\mathcal{B}} \sum_{T_k\in\mathcal{T}_{hn}} max(0, S(I,T_k) - S(I,T) + Th_k)$$

# Our approach

4. Use **adaptive margin** for the rank loss – **curriculum learning**

$$\mathcal{L}_{cmr} = \sum_{(I,T)\in\mathcal{B}} \sum_{T_k\in\mathcal{T}_{hn}} max(0, S(I,T_k) - S(I,T) + Th_k^t)$$

$$Th_k^t = min\left(u, \frac{1}{N}\sum_{(I,T)\in\mathcal{B}} (S^{t-1}(I,T) - S^{t-1}(I,T_k))\right)$$

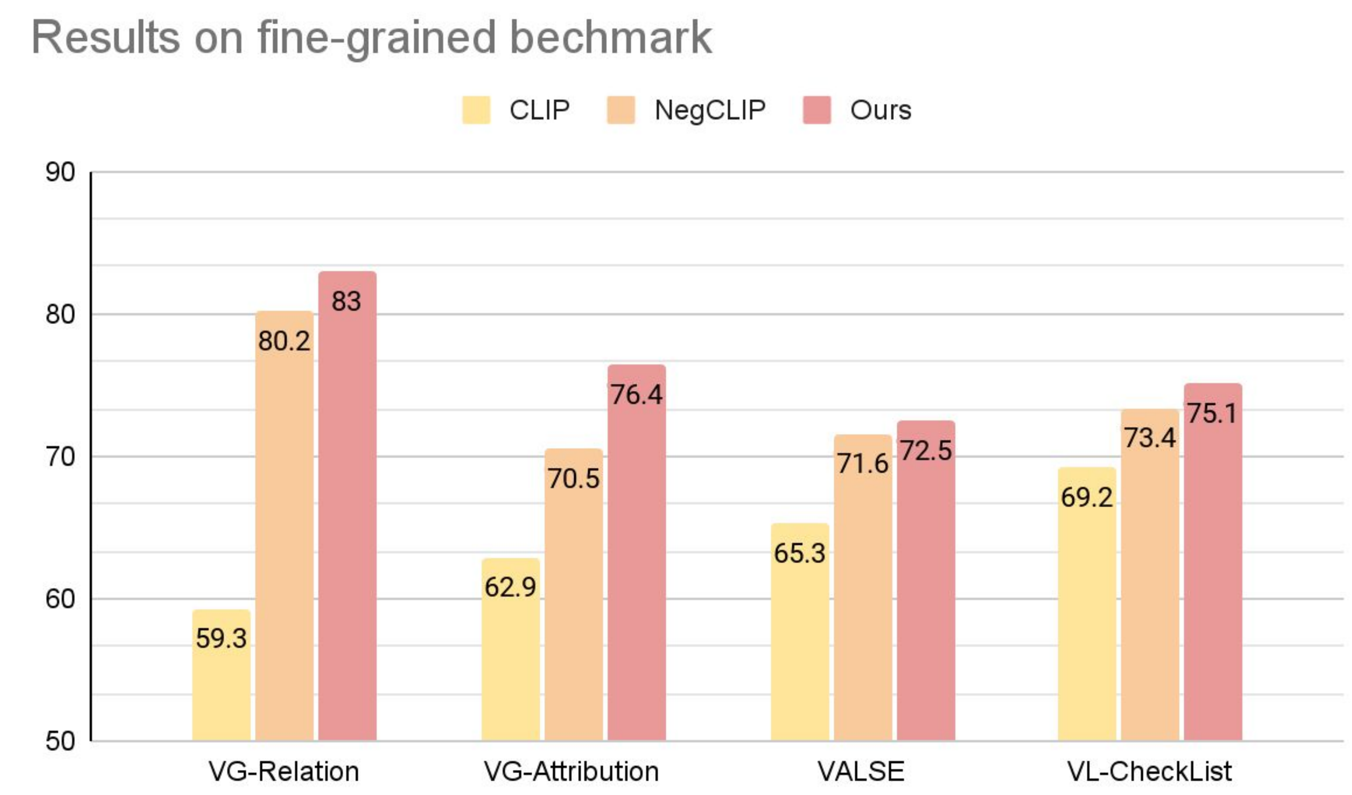
u = upper bound to stabilize training

$$\mathcal{L} = \mathcal{L}_{itm(hn)} + \alpha \cdot \mathcal{L}_{imc} + \beta \cdot \mathcal{L}_{cmr}$$

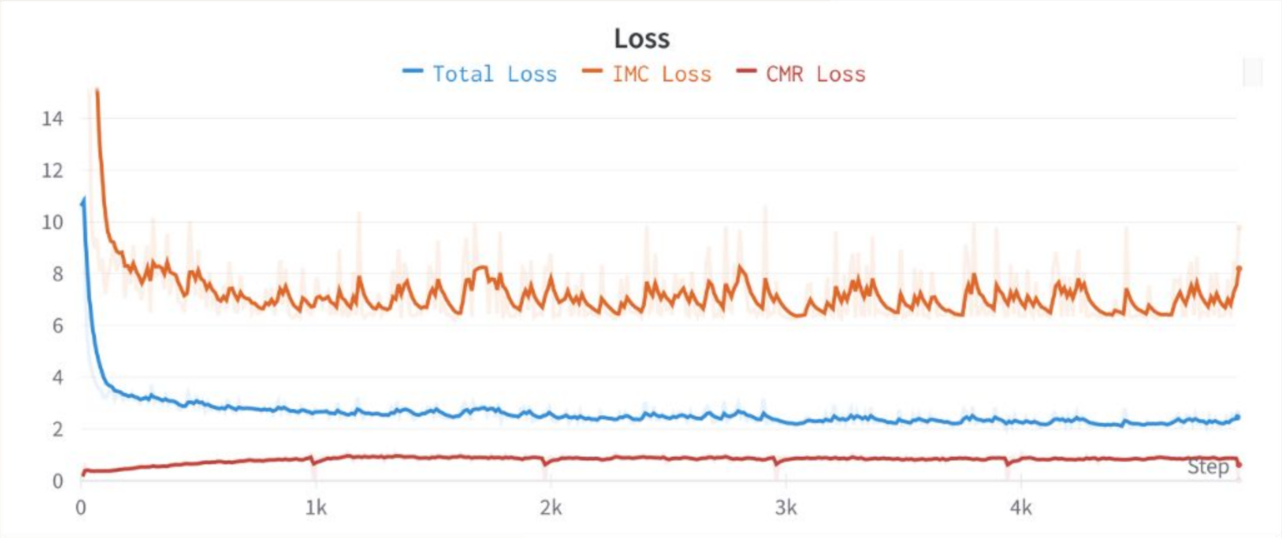
# Experimental Results

- We **outperform** existing methods **significantly** in both relation and attribution understanding.

| Benchmark | Task | # image-text pairs |
|---|---|---|
| | *Fine-grained Tasks* | |
| ARO | Relation, Attributes | 24k |
| VALSE | Linguistic Phenomena | 6.8k |
| VL-CheckList | Objects, Attributes and Relations | 410k |



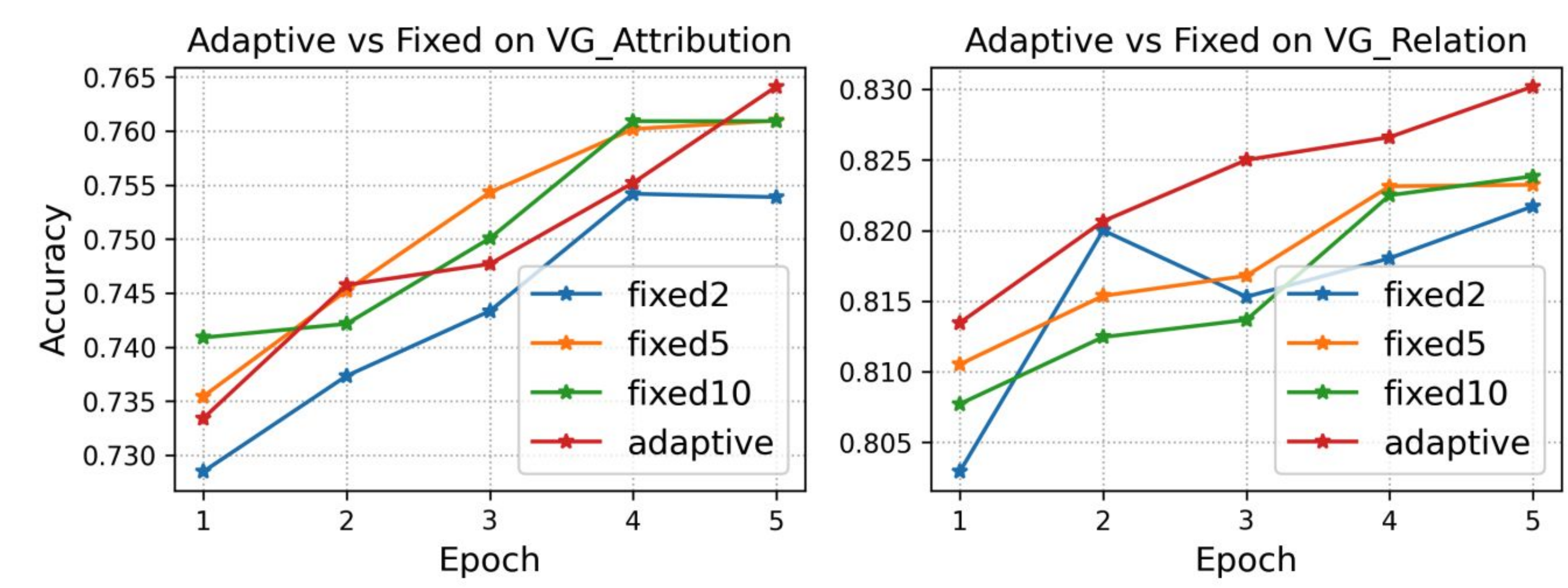
# Learning Dynamics


1. Growing threshold indicate growing ability, underlining curriculum learning

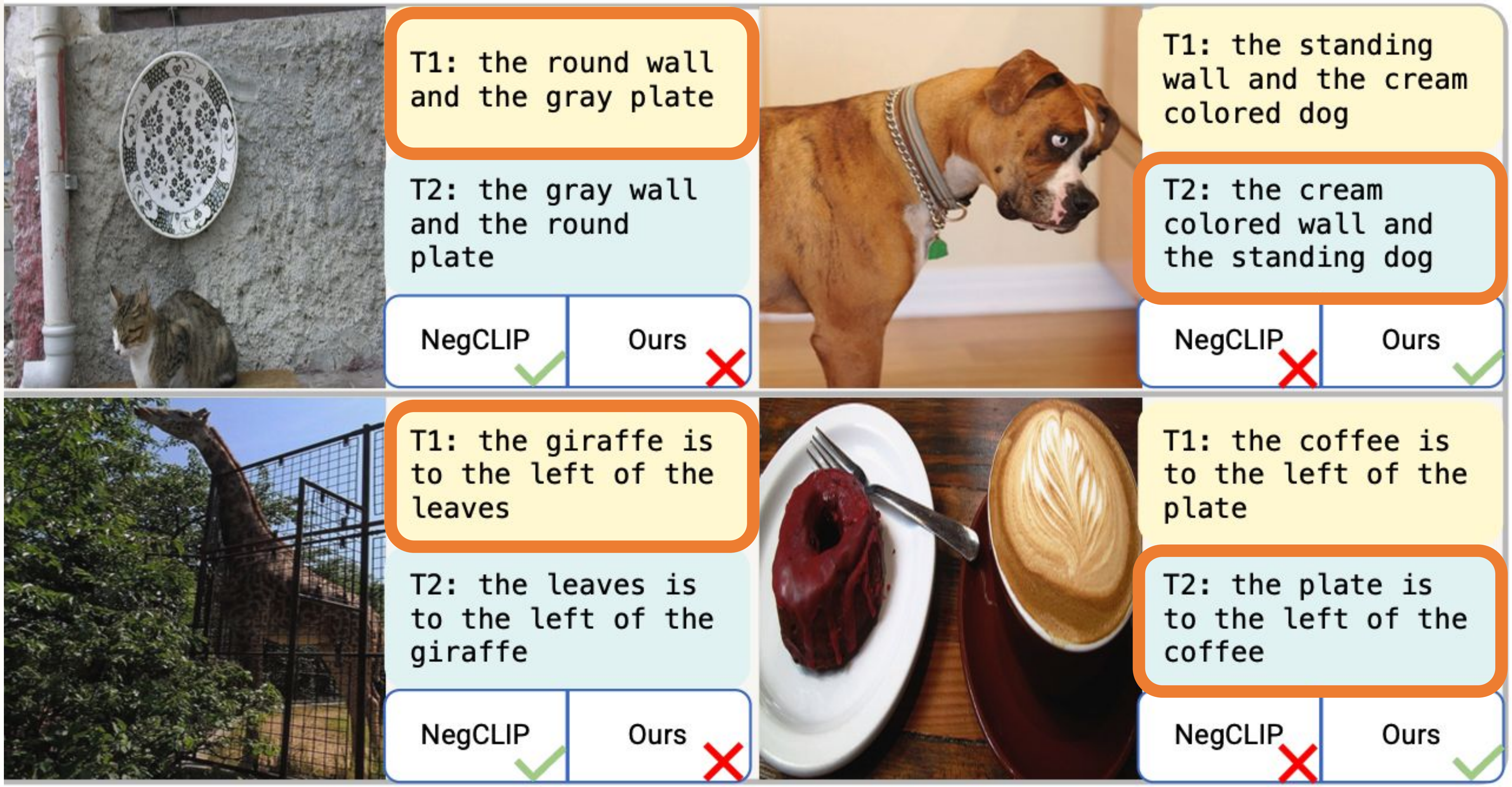
2. CMR loss remain stable, indicating balance between growing ability and task difficulty

# Adaptive vs. Fixed Thresholds




Adaptive threshold yields better results, without the need for complex hyper-parameters tuning

# Qualitative Examples



# Experimental Results

- We **outperform** existing methods **significantly** in both relation and attribution understanding.

- Ablation studies show that **both proposed losses are effective** for learning compositionality

**Le Zhang**

# Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic fine-grained understanding

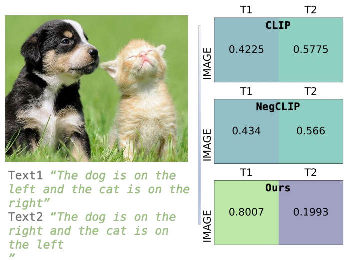
Le Zhang, Rabiul Awal, Aishwarya Agrawal

Mila - Quebec AI Institute, Université de Montréal

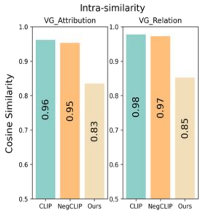
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

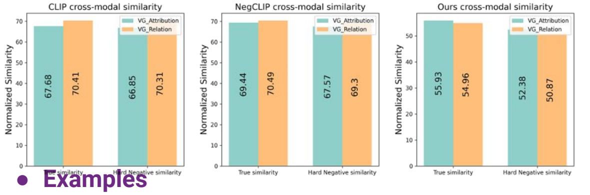
## Introduction

- **Task: fine-grained understanding (relation, attribution, object existence)**



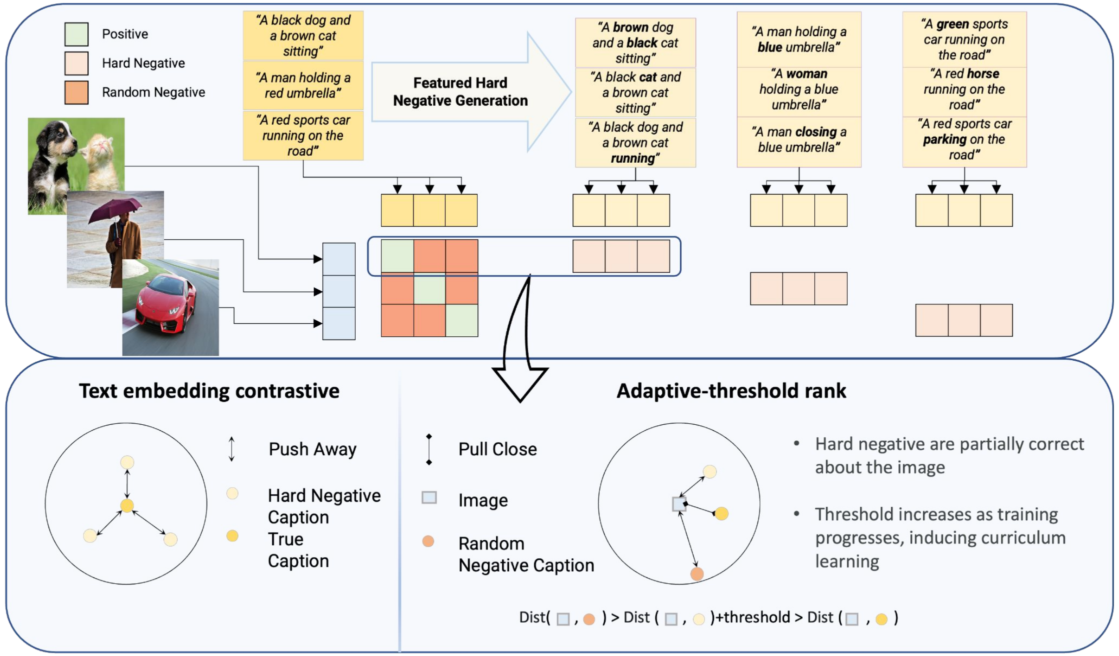

- **Limitation of current models**
  - **High intro-modal similarity**
  - **Small gap between true and hard negative pairs**




- **Examples**


## Method




- **Featured Hard Negative Generation**
- **Hard Negative ITC**

$$\mathcal{L}_{itc(hn)} = \sum_{(I,T)\in\mathcal{B}} -\left(\log \frac{\exp^{S(I,T)}}{\sum_{T_i\in\mathcal{B}} \exp^{S(I,T_i)} + \sum_{T_k\in\mathcal{T}_{hn}} \exp^{S(I,T_k)}} + \log \frac{\exp^{S(I,T)}}{\sum_{I_j\in\mathcal{B}} \exp^{S(I_j,T)}}\right)$$

- **Intro-Modal Contrastive**

$$\mathcal{L}_{imc} = \sum_{(I,T)\in\mathcal{B}} -\log \frac{\exp^{S(I,T)}}{\sum_{T_k\in\mathcal{T}_{hn}} \exp^{S(T,T_k)}}$$

- **Cross-Modal Rank with adaptive threshold**

$$\mathcal{L}_{cmr} = \sum_{(I,T)\in\mathcal{B}} \sum_{T_k\in\mathcal{T}_{hn}} max(0, S(I,T_k) - S(I,T) + Th_k^t)$$

$$Th_k^t = \frac{1}{|\mathcal{B}|} \sum_{(I,T)\in\mathcal{B}} (S^{t-1}(I,T) - S^{t-1}(I,T_k))$$

$$\mathcal{L} = \mathcal{L}_{itm(hn)} + \alpha \cdot \mathcal{L}_{imc} + \beta \cdot \mathcal{L}_{cmr}$$

## Experiments

| Model | ARO | | VALSE | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Relation | Attribution | Existence | Plurality | Counting | Relations | Actions | Coreference | Foil-it | Avg |
| Random | | | | | 50 | | | | | |
| BLIP | 59.0 | 88.0 | 86.3 | 73.2 | 68.1 | 71.5 | 69.1 | 51.0 | 93.8 | 69.96 |
| LXMERT† | - | - | 78.6 | 64.4 | 58.0 | 60.2 | 50.3 | 45.5 | 87.1 | 59.6 |
| CLIP | 59.3 | 62.9 | 68.7 | 57.1 | 61.0 | 65.4 | 74.8 | 52.5 | 89.8 | 65.3 |
| NegCLIP | 80.2 | 70.5 | 76.8 | 71.7 | **65.0** | 72.9 | 83.2 | **56.2** | 91.9 | 71.6 |
| CLIP *Ours* | **83.0** | **76.4** | **78.6** | **77.7** | 64.4 | **74.4** | **84.9** | 54.7 | **93.7** | **72.5** |
| XVLM-coco | 73.4 | 86.8 | 83.0 | **75.6** | 67.5 | 69.8 | 71.2 | 48.0 | **94.8** | 69.5 |
| XVLM *Ours* | **73.9** | **89.3** | **83.3** | 73.8 | **69.8** | **70.0** | **71.5** | **48.4** | 93.3 | **70.8** |

Table 2: **Results (%) of ARO and VALSE**, the best scores for each section emphasized in boldface. † represents scores extracted from papers.

| Model | VL-CheckList | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Attribute | | | | | Object | | Relation | | Avg |
| | Action | Color | Material | Size | State | Location | Size | Action | Spatial | |
| Random Chance | | | | | 50 | | | | | |
| BLIP† | 79.5 | 83.2 | 84.7 | 59.8 | 68.8 | 83.0 | 81.3 | 81.5 | 59.5 | 75.7 |
| CLIP-SVLC† | 69.4 | 77.5 | 77.4 | 73.4 | 62.3 | - | - | 74.7 | 63.2 | - |
| CLIP | 70.5 | 69.4 | 69.5 | 60.7 | 67 | 80.2 | 79.7 | 72.2 | 53.8 | 69.2 |
| NegCLIP | 72.1 | **75.7** | 78.1 | 61.3 | 67.3 | 84.4 | 83.8 | **80.7** | 57.1 | 73.4 |
| CLIP *Ours* | **75.6** | 72.7 | **79.7** | **65.3** | **69.8** | **84.8** | **84.5** | 78.5 | **65.0** | **75.1** |
| XVLM-coco | 80.4 | **81.1** | **83.1** | 60.3 | **70.8** | 86.3 | 85.3 | 79.0 | 61.8 | 76.5 |
| XVLM *Ours* | **80.5** | 76.0 | 80.6 | **67.2** | 69.8 | **87.3** | **86.6** | **80.8** | **78.6** | **78.6** |

Table 3: **Results (%) of VL-CheckList.** † represents scores are extracted from papers.

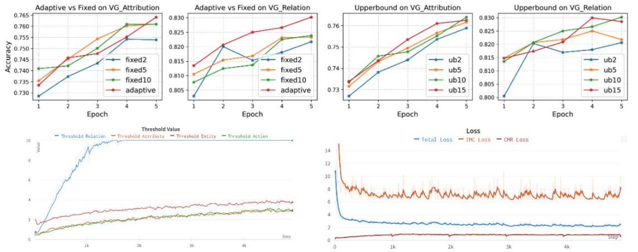



Figure 4: **Ablation study and analysis on threshold** (Top Left) Adaptive threshold vs Fixed threshold; (Top Right) Performance with different upper bound values.; (Bottom Left) Curves showing how the thresholds evolve over time ; (Bottom Right) Proposed loss curves change over time

## Conclusion

- **Hard-negatives can largely improve fine-grained understanding of VLMs**
- **Teach model to distinguish intro-modal hard negatives improve cross-modal fine-grained understanding**
- **Cross-modal rank encourage model to distinguish hard negatives, adaptive threshold entails curriculum learning**