JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Teaching Structured Vision & Language Concepts to Vision & Language Models

Sivan Doveh[1,2], Assaf Arbelle[1], Sivan Harary[1],
Rameswar Panda[1], Roei Herzig[1], Eli Schwartz, Donghyun Kim[1], Raja Giryes[3], Rogerio Feris[1],
Shimon Ullman[2], Leonid Karlinsky[1]

1. IBM Research, 2. Weizmann Institute, 3. Tel Aviv University

WEIZMANN INSTITUTE OF SCIENCE

TUE-AM-253

Current SOTA VL Models logic:

The problem: VL Models Struggle with SVLC

- Current VL models focus on the object.
- VL models ignore relations between objects.
- VL models ignores object attributes and states.
- Called an “object bias” in recent literature [winoground, vl checklist].

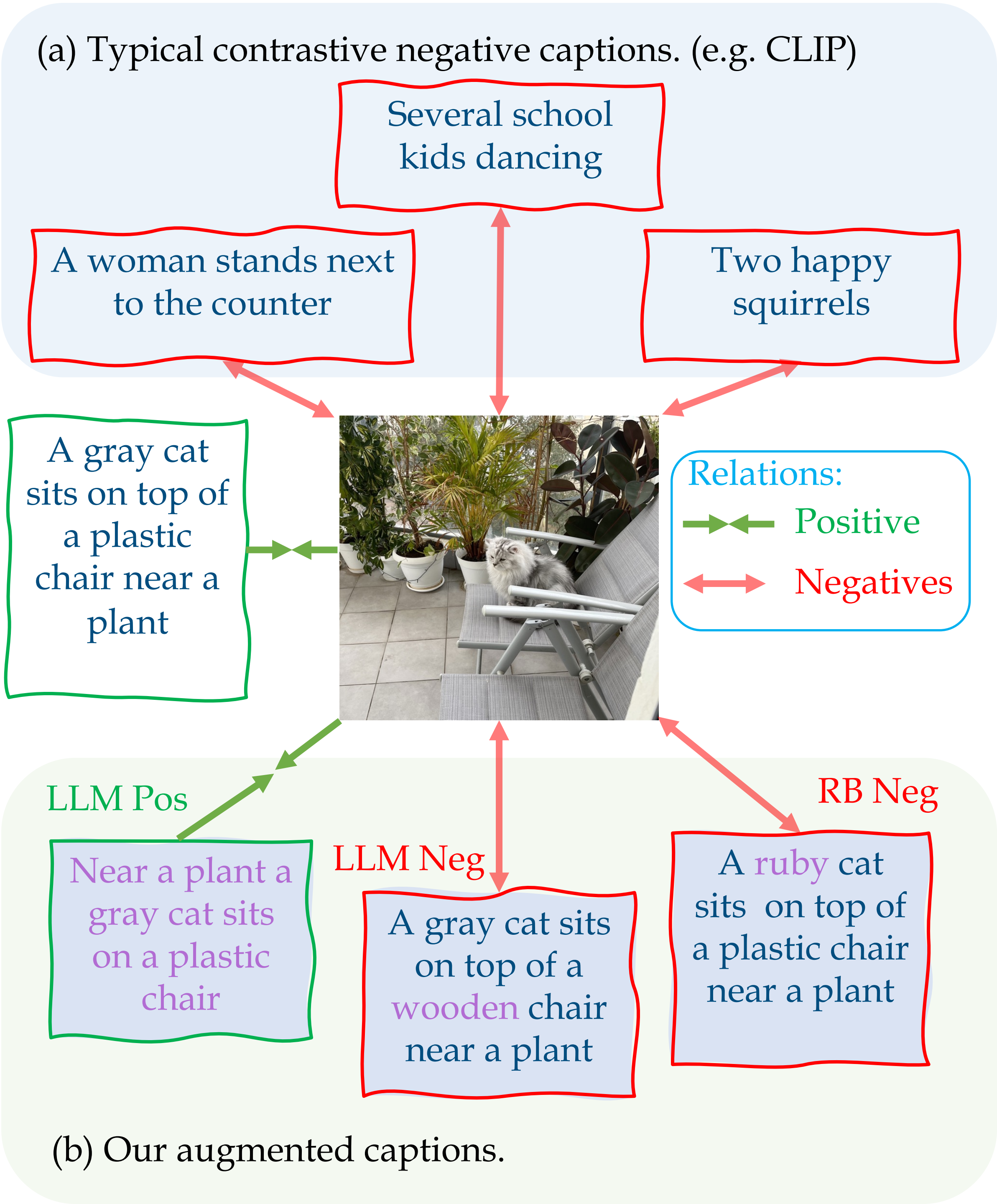
Query	Score
Cat sits on a chair.	83%
Chair sits on a cat.	81%

The Solution: specialized losses, augmented captions

- Current state: CLIP's negative captions are completely unrelated to the image.
- Our method:

➢ Positive captions augmentation.

➢ Negative captions augmentation by minor changes to positive captions.



What are "Structured Vision & Language Concepts" (SVLC)?

SVLC - Structured Vision & Language Concepts.

Characteristics from both image and caption:

- Object attributes.
- Inter-object relations.
- Object states.

"A gray cat sits on top of a plastic chair near a plant"

What are "Structured Vision & Language Concepts" (SVLC)?

SVLC - Structured Vision & Language Concepts.

Characteristics from both image and caption:

- Object attributes.
- Inter-object relations.
- Object states.

"A gray cat sits on top of a plastic chair near a plant"

- Gray cat

What are "Structured Vision & Language Concepts" (SVLC)?

SVLC - Structured Vision & Language Concepts.

Characteristics from both image and caption:

- Object attributes.
- Inter-object relations.
- Object states.

"A gray cat sits on top of a plastic chair near a plant"

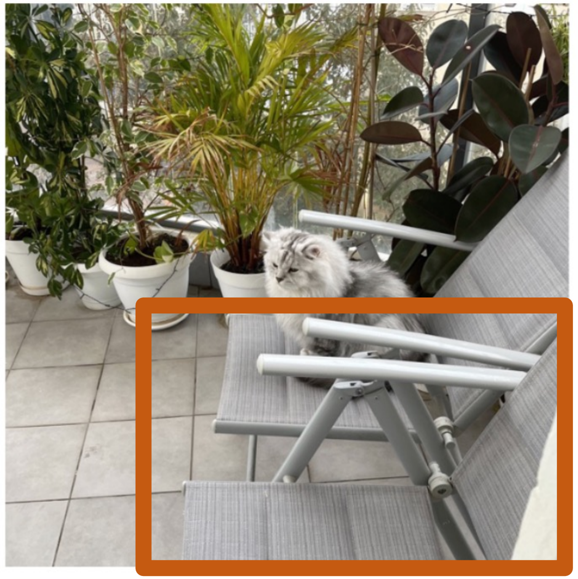
- Gray cat
- Cat sits on top of a chair

What are "Structured Vision & Language Concepts" (SVLC)?

SVLC - Structured Vision & Language Concepts.

Characteristics from both image and caption:

- Object attributes.
- Inter-object relations.
- Object states.



"A gray cat sits on top of a plastic chair near a plant"

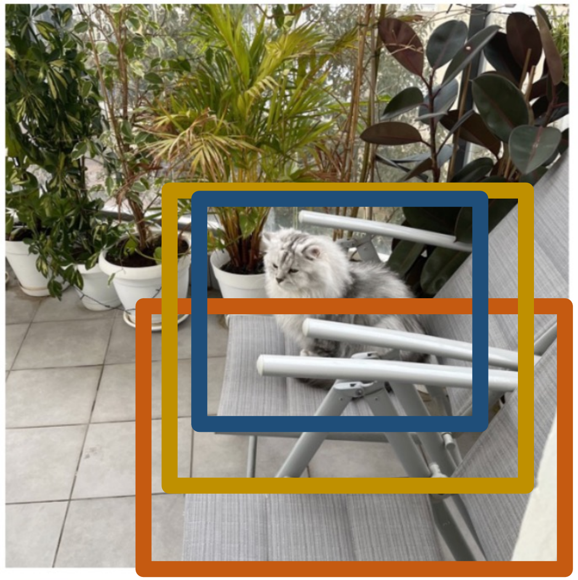
- Gray cat
- Cat sits on top of a chair
- Plastic chair

What are “Structured Vision & Language Concepts” (SVLC)?

SVLC - Structured Vision & Language Concepts.

Characteristics from both image and caption:

- Object attributes.
- Inter-object relations.
- Object states.



“A gray cat sits on top of a plastic chair near a plant”

- Gray cat
- Cat sits on top of a chair
- Plastic chair

The problem: VL Models Struggle with SVLC

- Current VL models focus on the object.
- VL models ignore relations between objects.

VL model

	Query	Score
✗	Cat sits on a chair.	83%
✓	Chair sits on a cat.	81%

VL model ignores who does the action on who

The problem: VL Models Struggle with SVLC

- Current VL models focus on the object.
- VL models ignores relations between objects.
- VL models ignores object attributes and states.
- Called an "object bias" in recent literature [winoground, vl checklist].

VL model

	Query	Score
✗	A black cat	75%
✓	A gray cat	72%

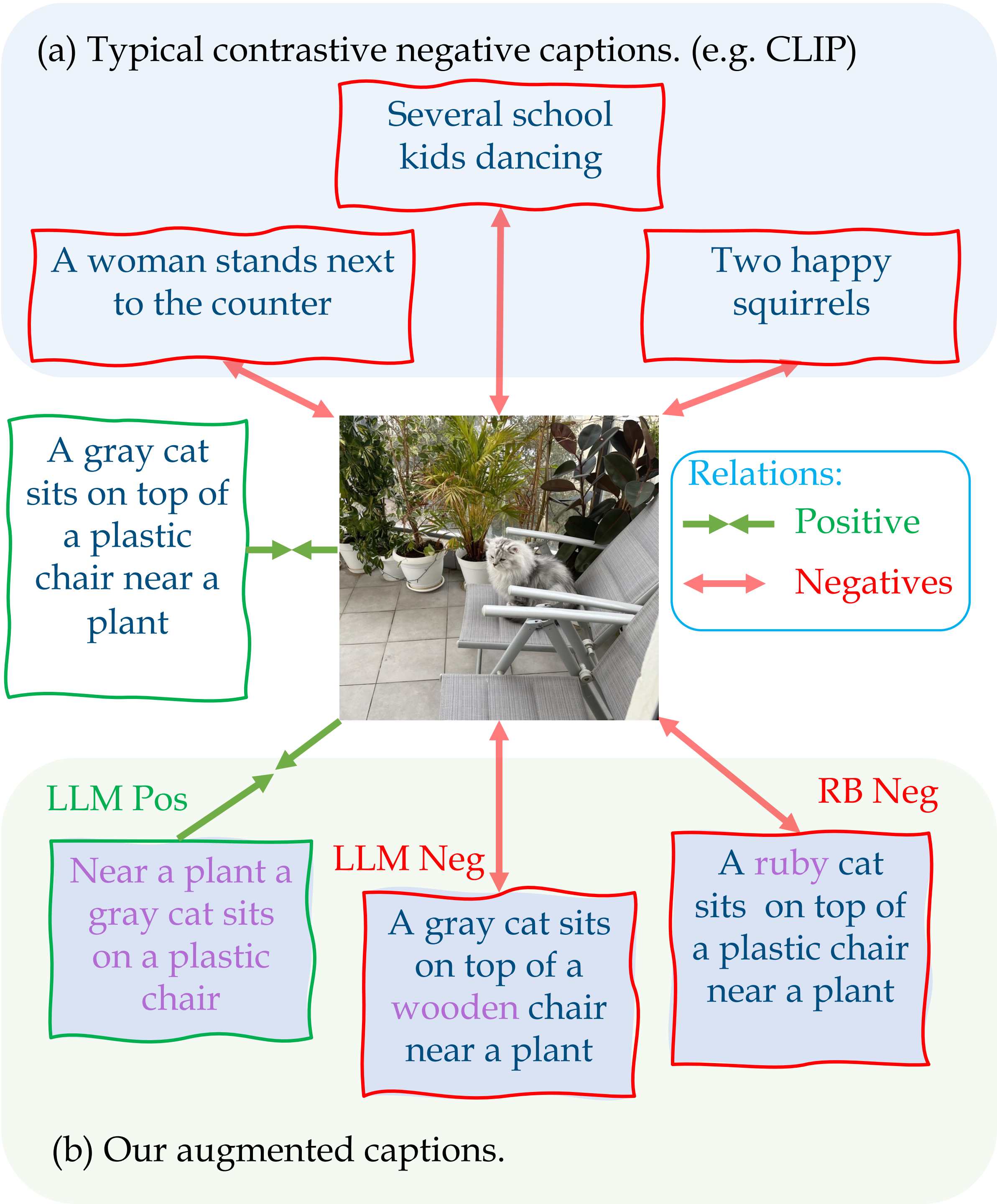
VL model ignores the cat's color

The Solution: specialized losses, augmented captions

- Current state: CLIP's negative captions are completely unrelated to the image.
- Our method:

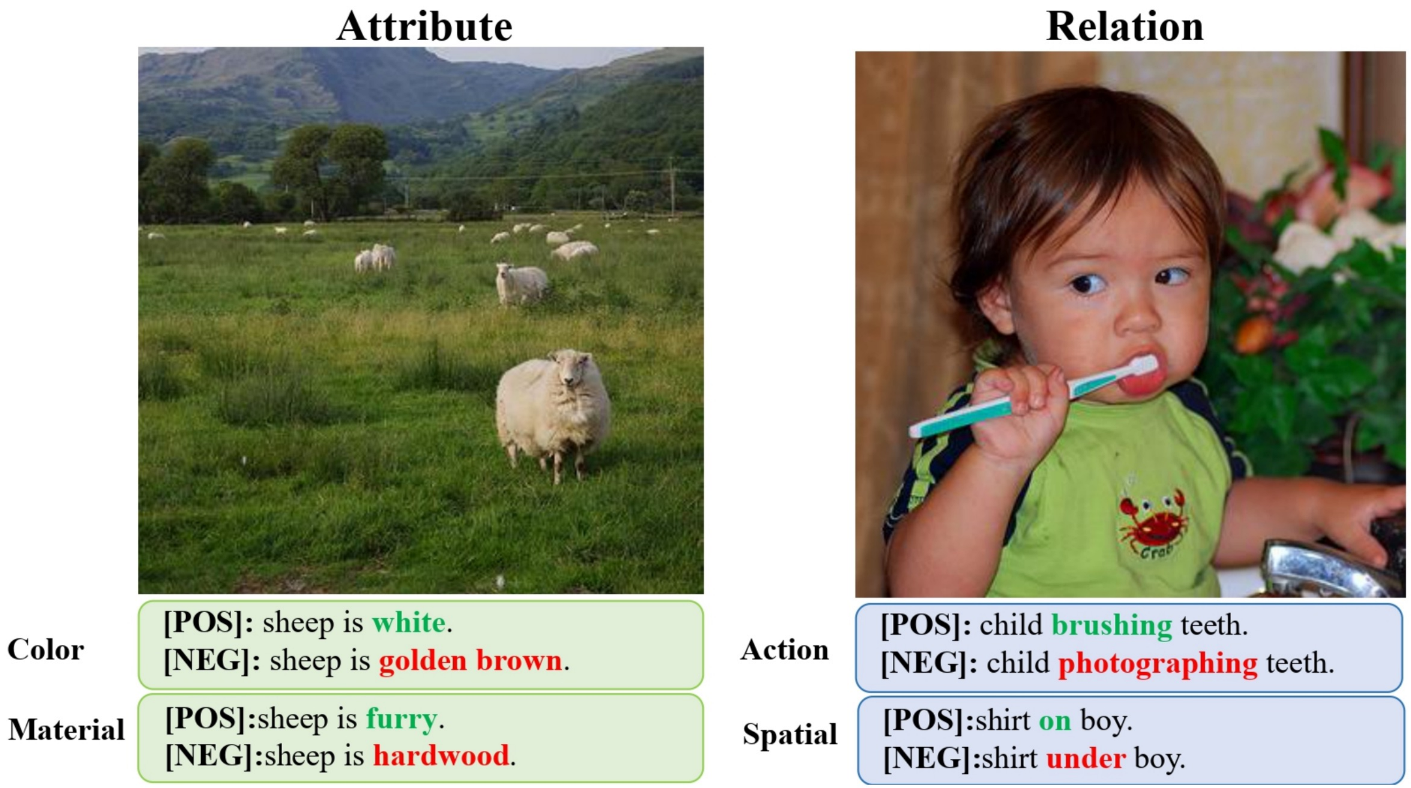
➢ Positive captions augmentation.

➢ Negative captions augmentation by minor changes to positive captions.



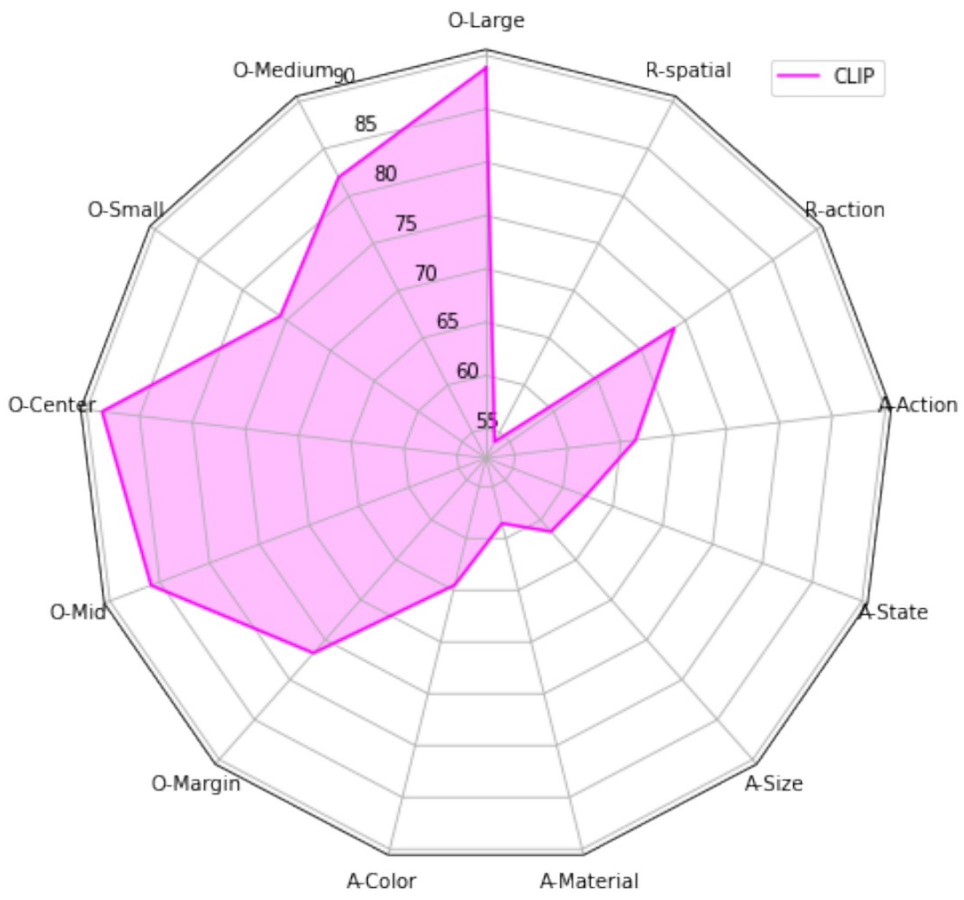
VL-Checklist – SVLC Benchmark

- Match the correct caption to the image.
- The captions only differ in one word:
    - Color
    - Material
    - Action
    - Size
    - etc.



VL-Checklist – VL models struggle

- CLIP excels in objects
- Struggles with relations and attributes



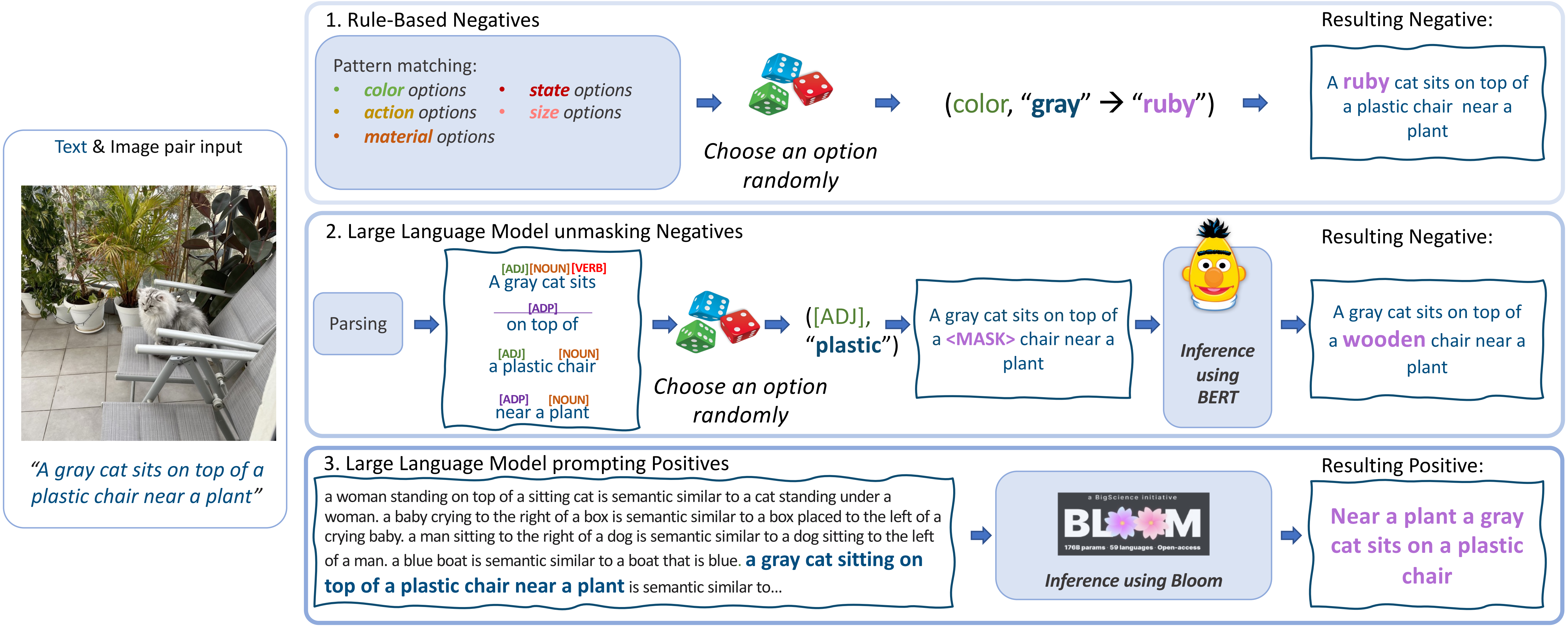
Large Vision and Language Datasets

- Conceptual Captions 3M (CC3M)
  - ~ 3 million images-Text pairs
  - harvested from the web
- LAION-400M (LAION)
  - ~ 400 Million Image-Text Pairs
  - CLIP-Filtered open dataset

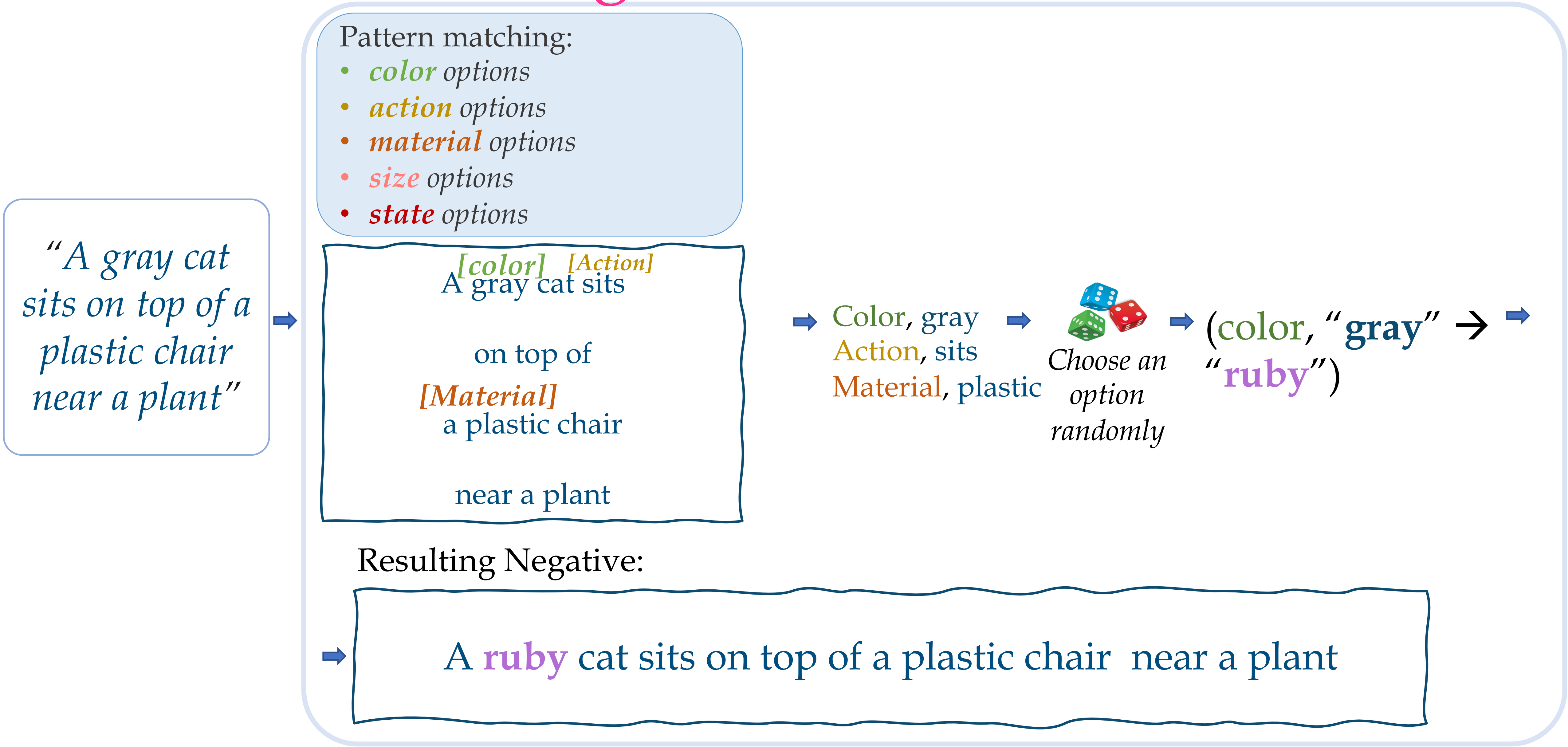
The man at bat readies to swing at the pitch while the umpire looks on.

A horse carrying a large load of hay and two people sitting on it.

Our Approach



Rule-Based Negatives



Large Language Model unmasking Negatives

"A gray cat sits on top of a plastic chair near a plant"

Parsing

[ADJ] [NOUN] [VERB]
A gray cat sits

[ADP]
on top of

[ADJ] [NOUN]
a plastic chair

[ADP] [NOUN]
near a plant

Choose an option randomly

([ADJ], "**plastic**")

A gray cat sits on top of a <MASK> chair near a plant

Inference using BERT

Resulting Negative:

A gray cat sits on top of a **wooden** chair near a plant

Large Language Model prompting Positives

"A gray cat sits on top of a plastic chair near a plant"

a woman standing on top of a table is semantically similar to a table placed under a woman; … ; … ; … ; … ; **a gray cat sitting on top of a plastic chair near a plant** is semantically similar to…

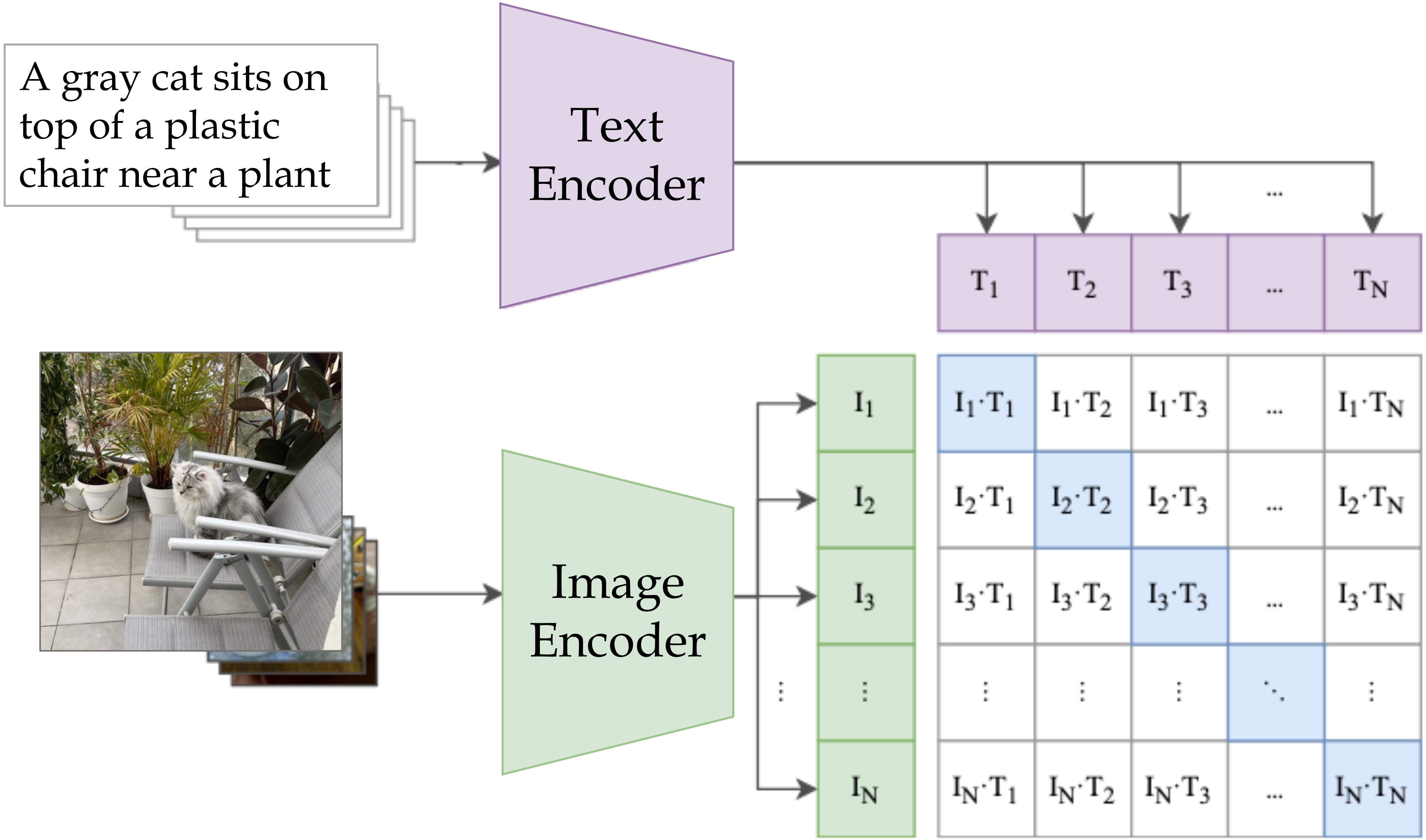
a BigScience initiative
BLOOM
176B params · 59 languages · Open-access

Inference using Bloom

Resulting Positive:

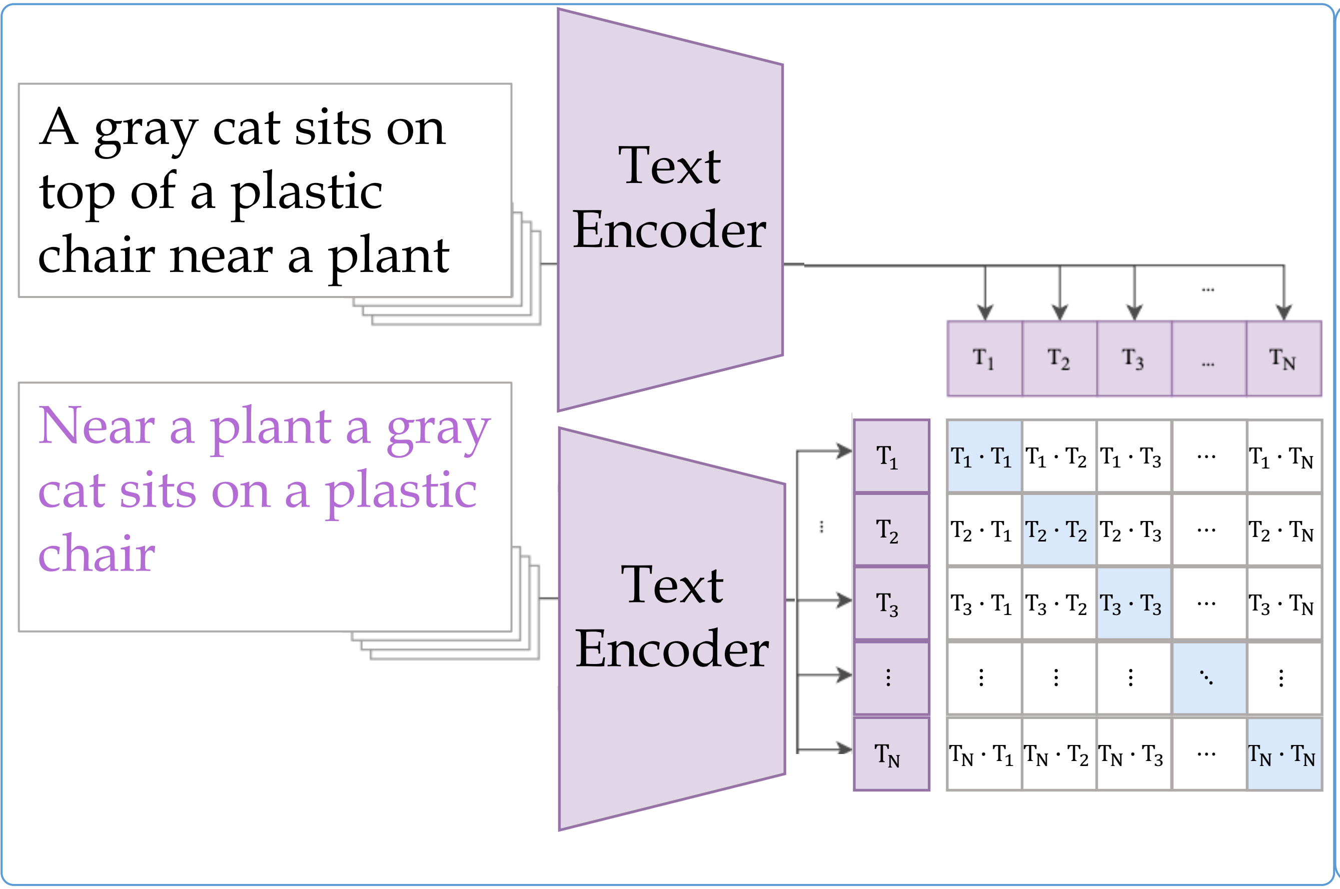
Near a plant a gray cat sits on a plastic chair

Loss Modification

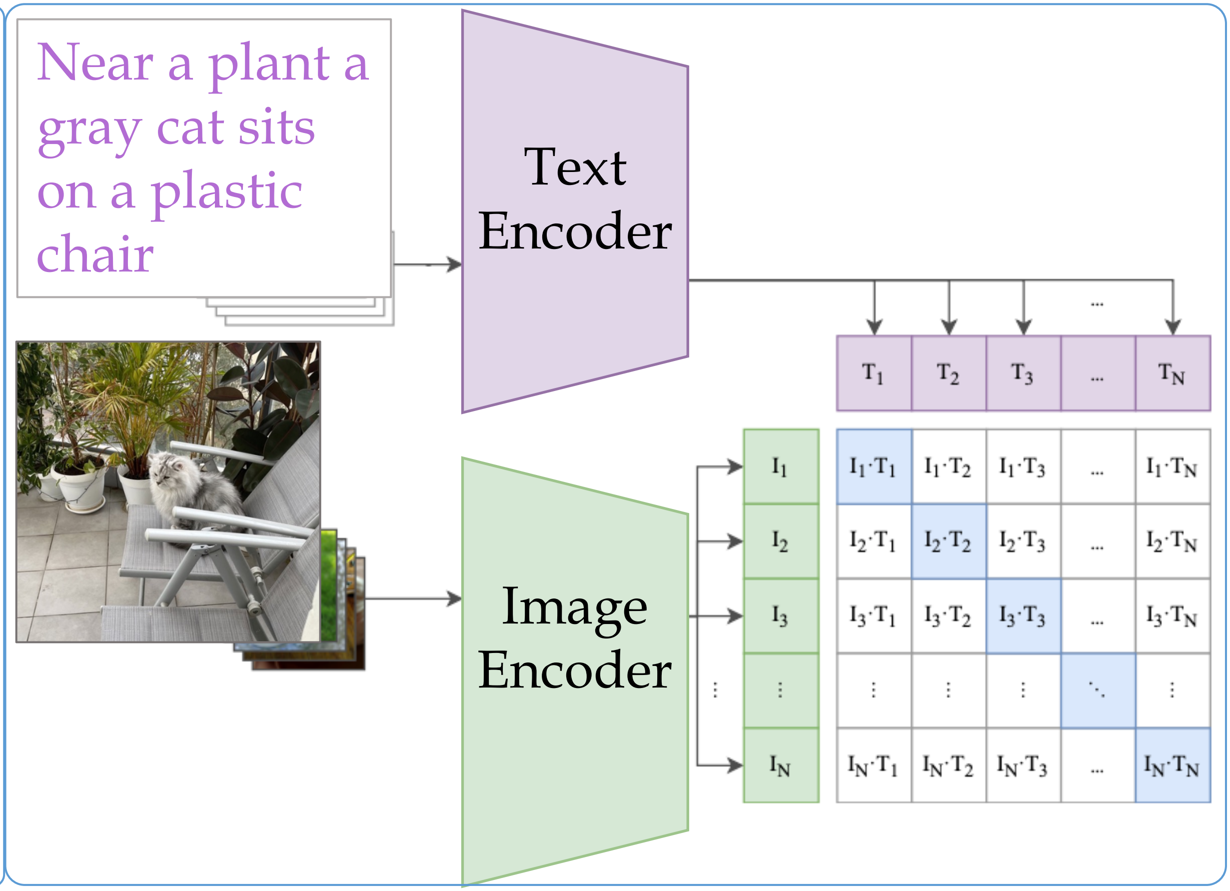


Loss Modification - Positive

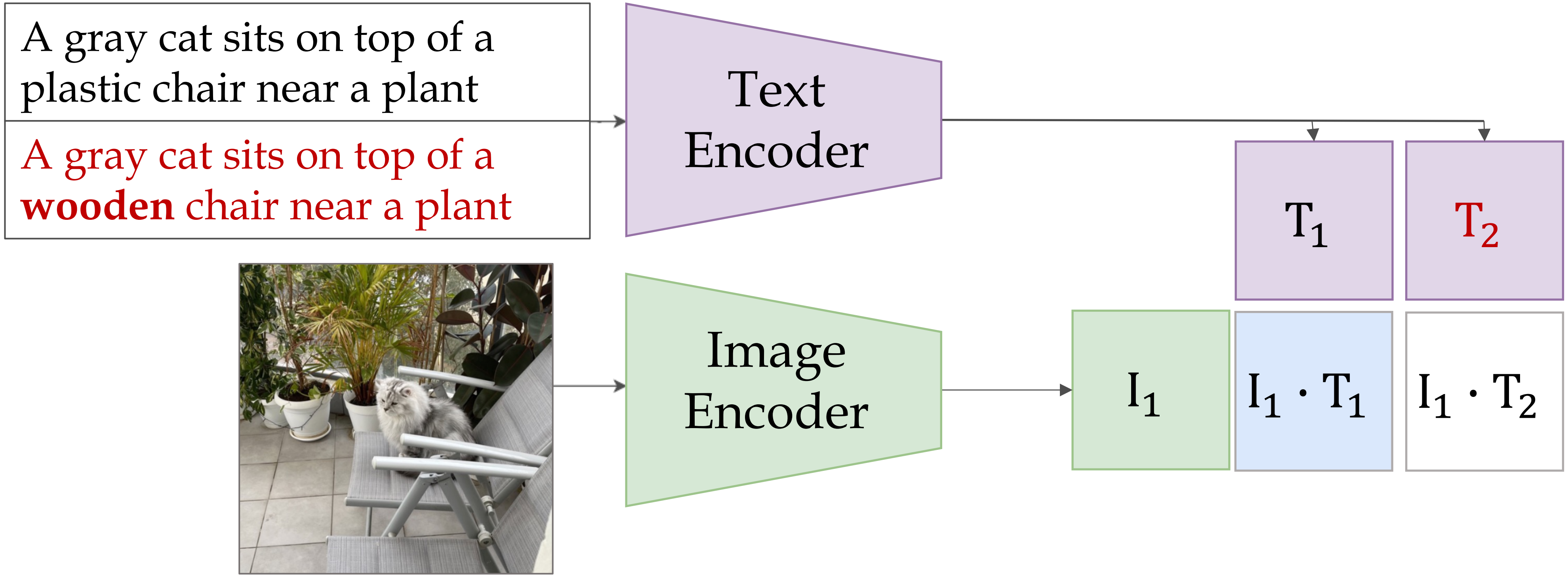
Text 2 Text:



Text 2 Image:

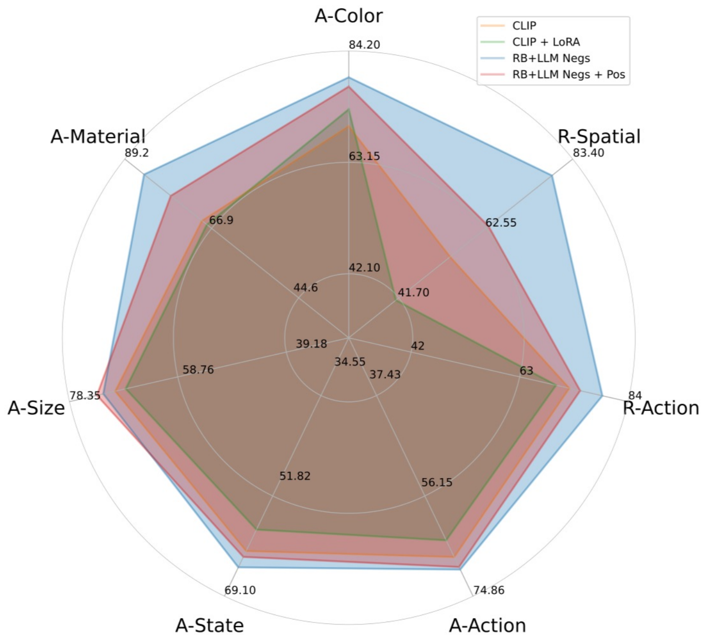


Loss Modification - Negative



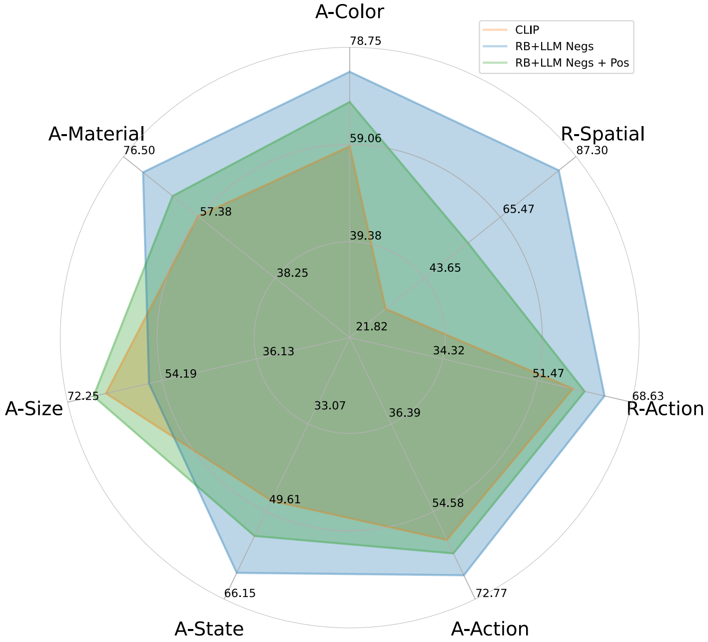
Total loss: $\mathcal{L} = \mathcal{L}_{cont} + \alpha \cdot \mathcal{L}_{neg} + \beta \cdot (\mathcal{L}_{sim}^{text} + \mathcal{L}_{sim}^{img})$

Finetuning Using CC3M Data



	VL-Checklist Object	Attribute	Relation	21 Zero-Shot Tasks Average
CLIP [59]	81.58%	67.60%	63.05%	56.37%
CLIP +LoRA	80.93% (-0.66%)	66.28% (-1.32%)	55.52% (-7.53%)	56.41%(+0.04%)
Ours RB Neg	83.89% (+2.30%)	73.35% (+5.75%)	75.33% (+12.28%)	54.32% (-2.05%))
Ours LLM Neg	84.44% (+2.85%)	71.63% (+4.03%)	74.82% (+11.77%)	55.60% (-0.77%))
Ours RB+LLM Negs	85.09% (+3.50%)	73.90% (+6.30%)	78.72% (+15.67%)	54.66% (-1.71%))
Ours Combined	85.00% (+3.42%)	71.97% (+4.37%)	68.95% (+5.90%)	54.77% (-1.60%))

Training from Scratch



	VL-Checklist Object	VL-Checklist Attribute	VL-Checklist Relation	21 Zero-Shot Tasks Average
CLIP	71.17%	57.86%	45.20%	21.96%
CLIP + Ours Combined	71.79% (+0.62%)	63.29% (+5.43%)	58.13% (+12.93%)	20.96% (-1.00%)
CyCLIP	69.41%	57.59%	53.70%	21.02%
CyCLIP + Ours Combined	71.50% (+2.09%)	65.69% (+8.10%)	70.20% (+16.50%)	20.44% (-0.42%)

Finetuning Using LAION400M Data

	VL-Checklist Object	Attribute	Relation	21 Zero-Shot Tasks Average
CLIP [59]	0.8158	0.676	0.6305	56.37%
CLIP + LoRA	82.18% (+0.60%)	68.48% (+0.88%)	62.72% (-0.33%)	57.15% (+0.78%)
Ours Combined	82.54% (+0.96%)	69.64% (+2.04%)	66.05% (+3.00%)	56.71% (+0.34%)

Summary

- Current V&L models mainly focus on objects and disregard detailed information in the text
- By manipulating just the textual descriptions and slightly modifying the loss function, we found that these models can greatly improve SVLC tasks.
- We are still a long way from having good SVLC in Vision & Language models. Further research in this direction is welcome and necessary