TUE-PM-261

GeneCIS: A Benchmark for General Conditional Image Similarity


Sagar Vaze


Nicolas Carion


Ishan Misra

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Meta UNIVERSITY OF OXFORD VGG UNIVERSITY OF OXFORD

Conditional Image Similarity

Humans understand many notions of ‘similarity’, and **choose one** for a given task

However, most image representations are **fixed**

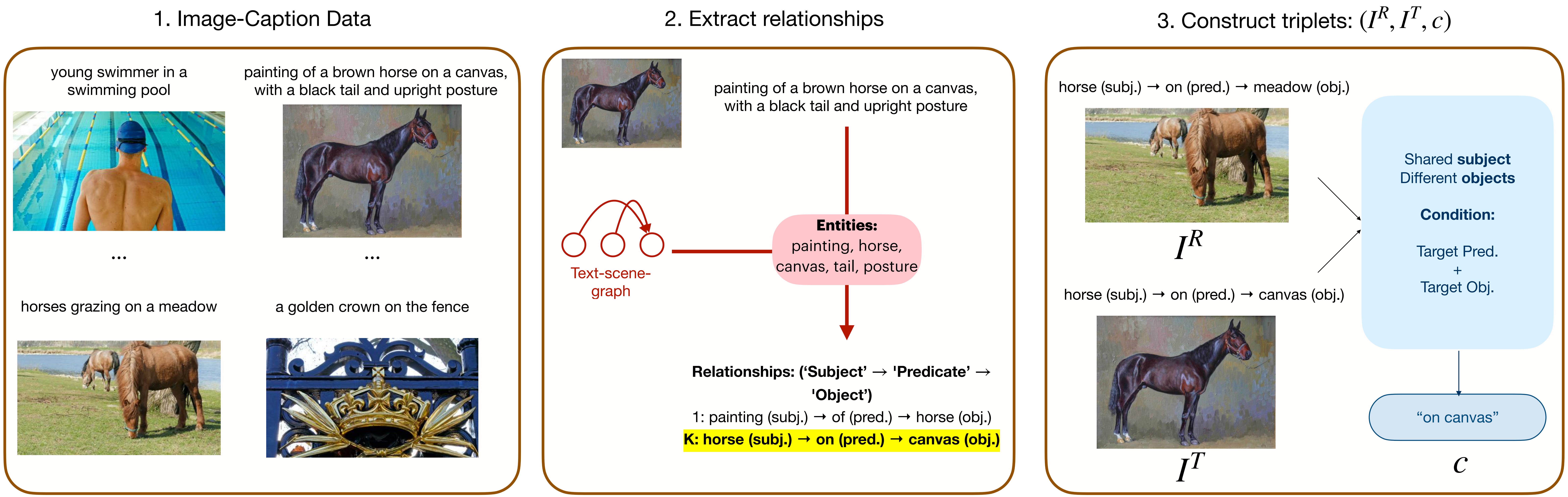
We present a way to **train** and **evaluate** models which can **adapt** to different notions of similarity

The GeneCIS Benchmark

- **Four conditional retrieval tasks** for zero-shot evaluation

Method

- We automatically (**scalably**) mine training data from **image-caption datasets**



Conditional Image Similarity

Key Challenge: The set of possible conditions is **infinite**

How do we **train** and **evaluate** such models?

Prior work focusses on constrained domains like **fashion** or **birds** [1, 2]

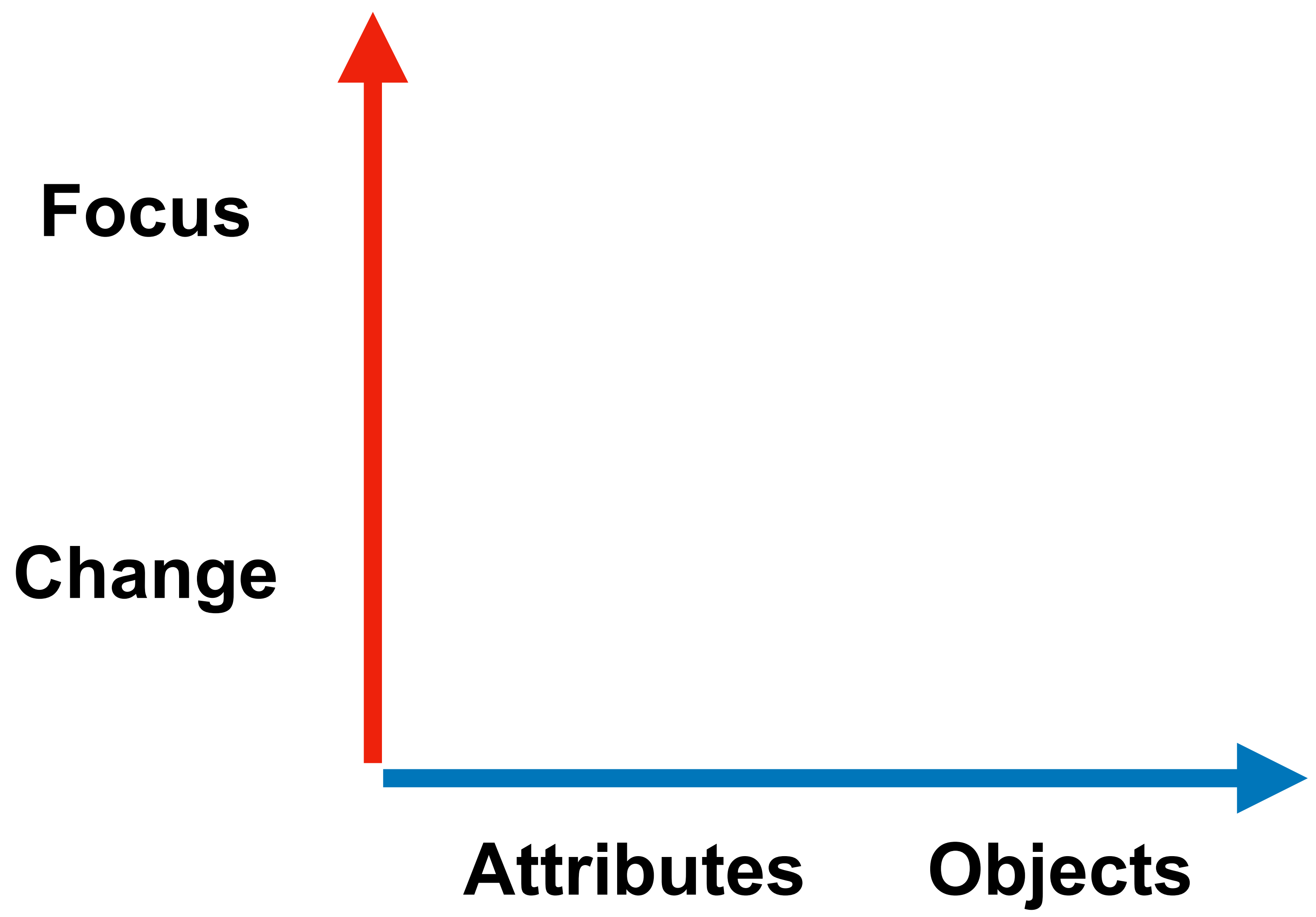
[1] Effectively Leveraging Attributes for Image Similarity, Mishra et. al, ICCV 2021
[2] Conditional Similarity Networks, Veit et al., CVPR 2017

Conditional Image Similarity

Solution: Evaluate **zero-shot** on an **open-set** of conditions

Models which perform well on a range of conditions understand general conditional similarity

Consider conditions along two **axes**

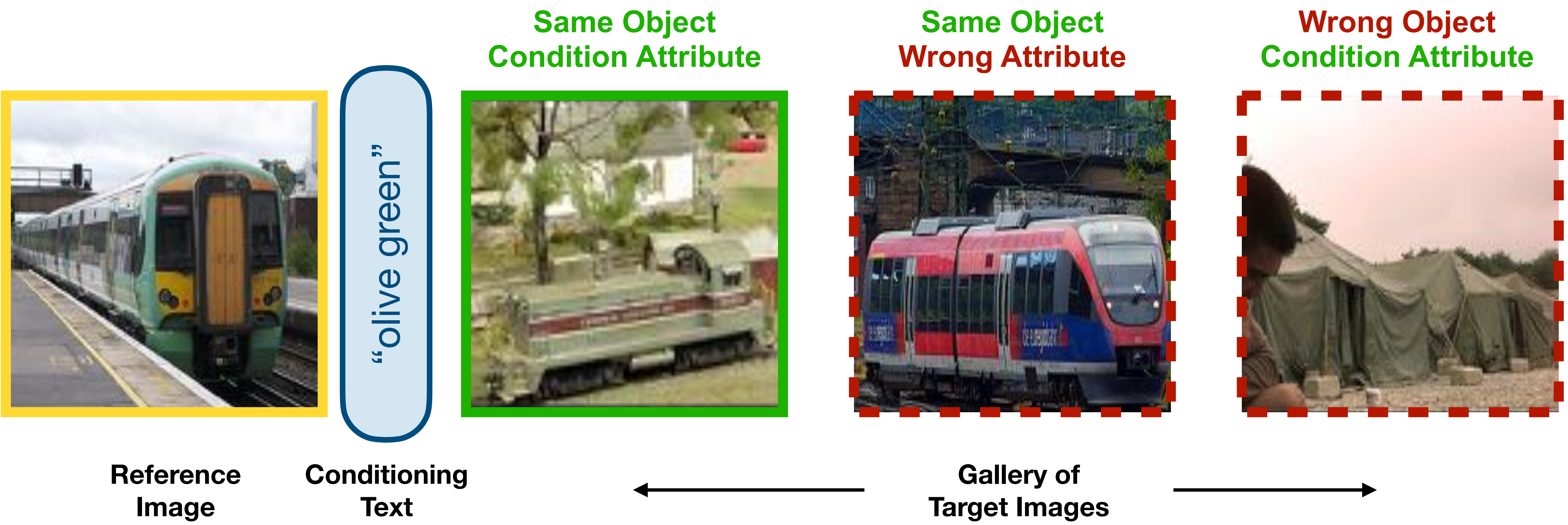


GeneCIS

- GeneCIS contains **four conditional retrieval tasks** for zero-shot evaluation
- Dataset is constructed from COCO and VAW (Visual Genome)
  - 2k samples per task and a long tail of conditions. Full details in the paper.

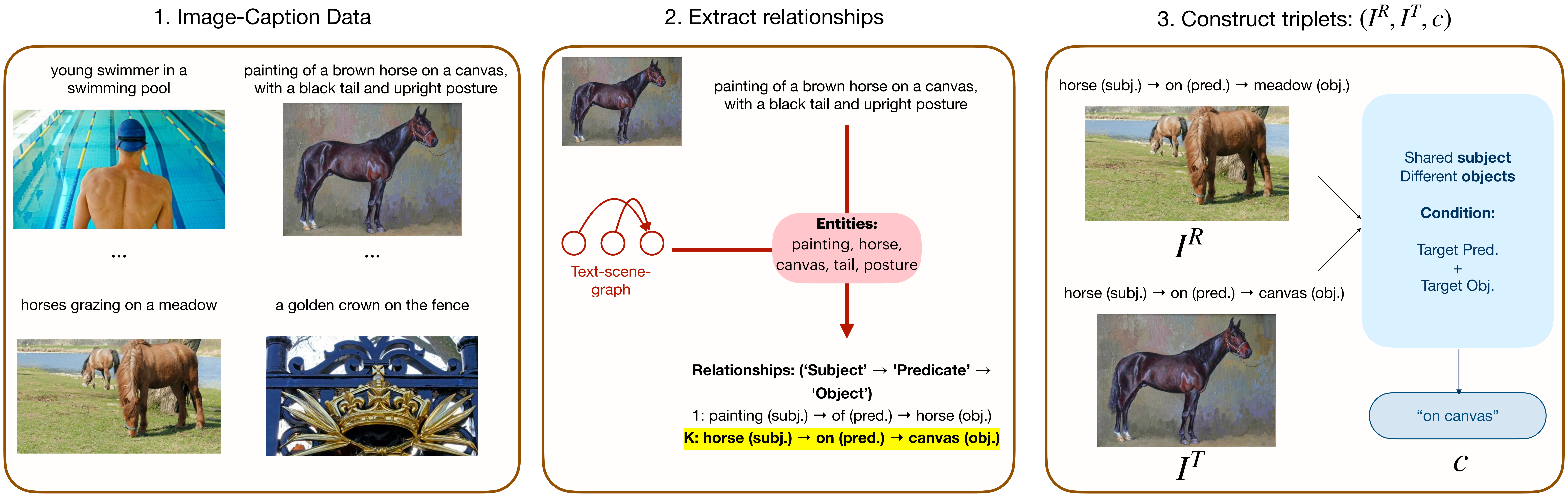
GeneCIS: Change an Attribute

- **Inputs:** (i) Reference Image; (ii) Conditioning Text; (iii) Gallery of Target Images
- **Outputs:** Best matching gallery image (**one correct answer**)
- **Distractors** in gallery prevent shortcut solutions



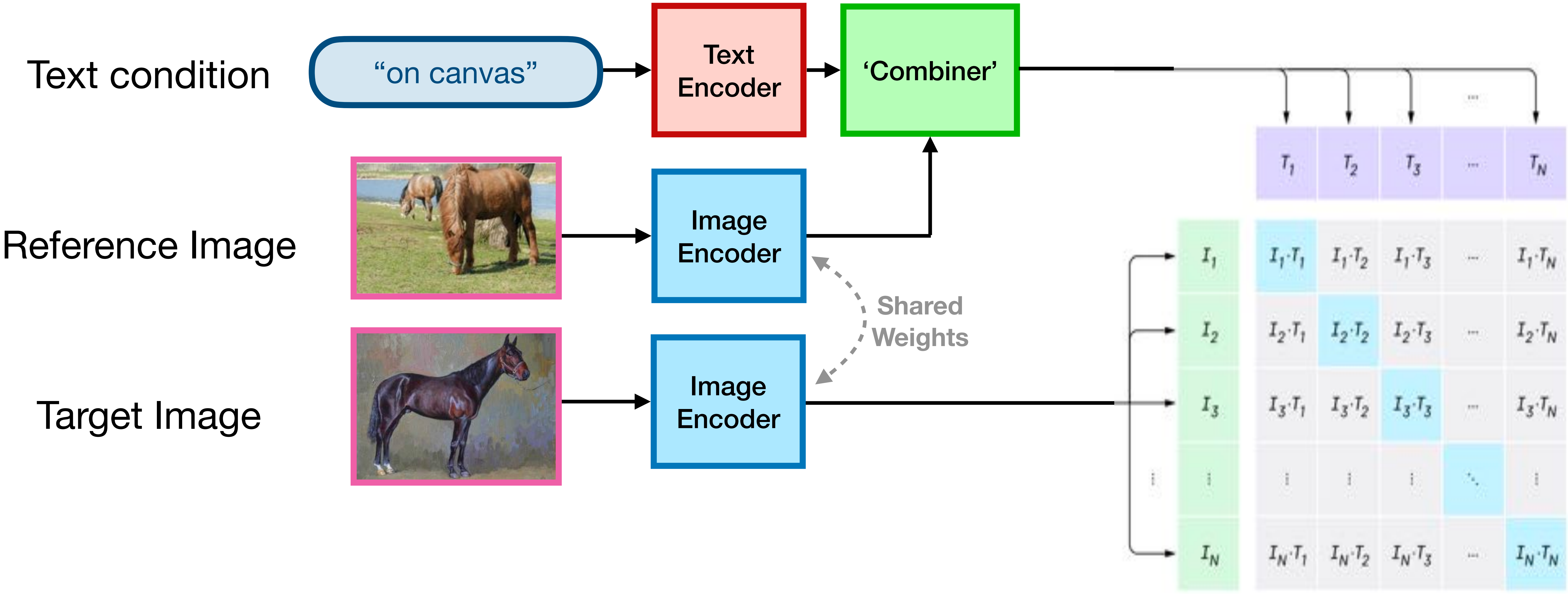
Method

- **Key challenge:** Open-set of similarity conditions.
  - Impossible to get exhaustively annotated training data
- **Solution:** Mine training data from image-caption datasets (Conceptual Captions 3M, CC3M)
  - Collect millions triplets of **(Reference Image, Target Image, Condition)**



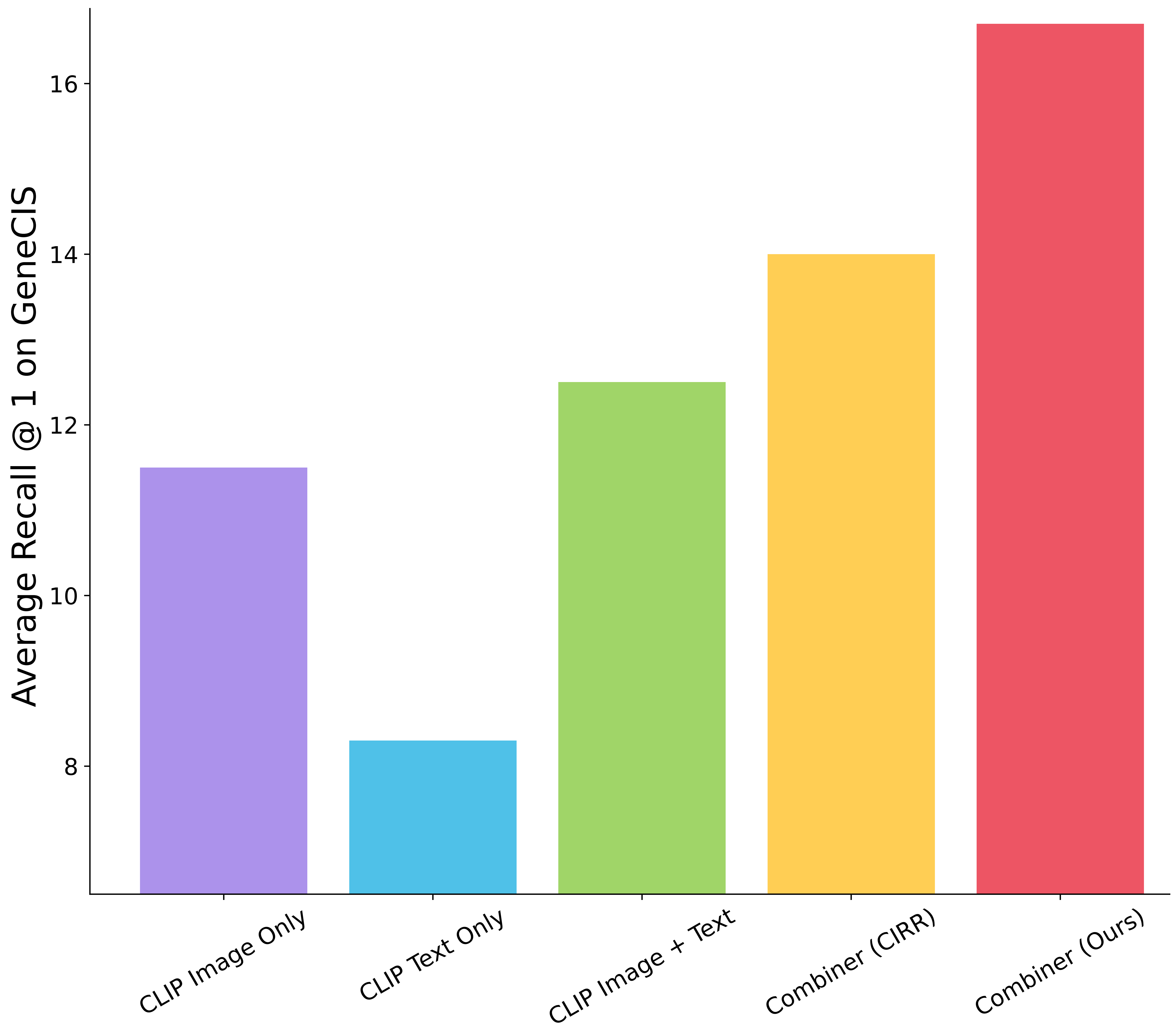
Method

- We now have millions of training triplets (we mine 1.6M triplets)
- Embed images and text with CLIP-initialized encoders
- Condition **reference image** features on **text condition** with ‘Combiner’ module [1]
- Train contrastively



[1] Conditioned and Composed Image Retrieval Combining and Partially Fine-Tuning CLIP-Based Features, Baldrati et al.,CVPRW 2022
https://openaiassets.blob.core.windows.net/$web/clip/draft/20210104b/overview-a.svg

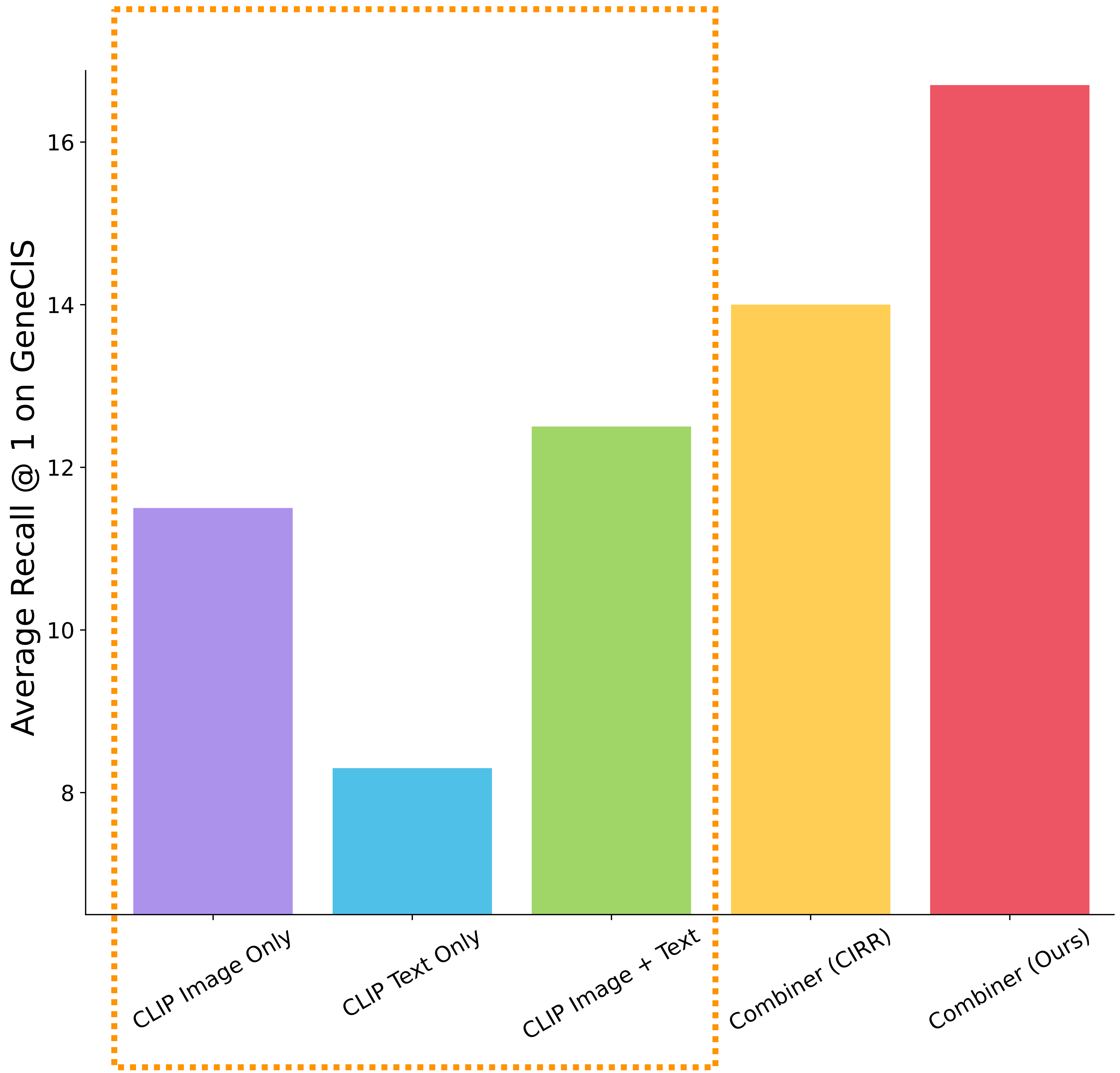
Results



Average Recall @ 1 on GeneCIS
16
14
12
10
8
CLIP Image Only
CLIP Text Only
CLIP Image + Text
Combiner (CIRR)
Combiner (Ours)

Results

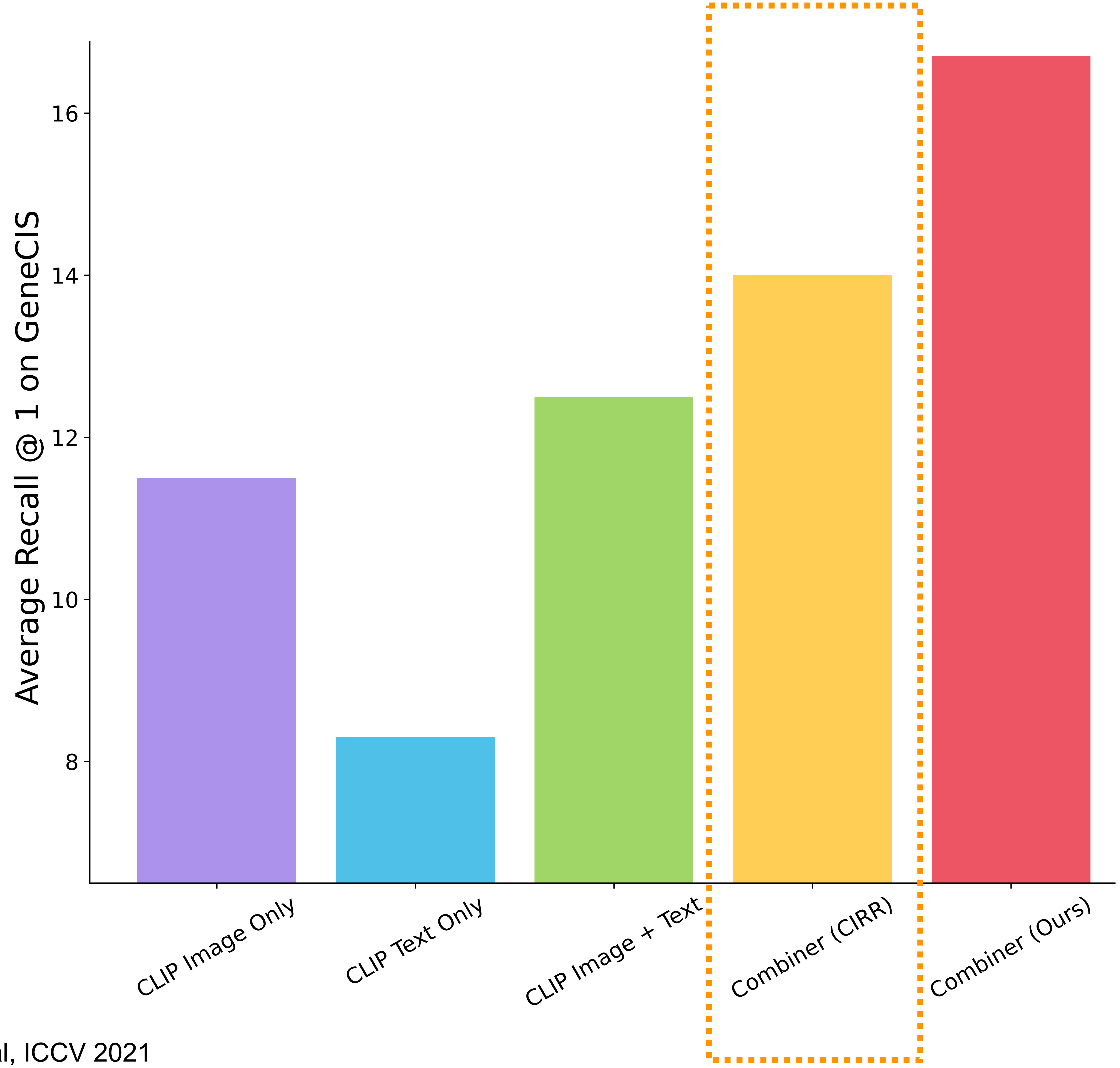
CLIP-Only Baselines



Results

Training with manual supervision from CIRR [1]

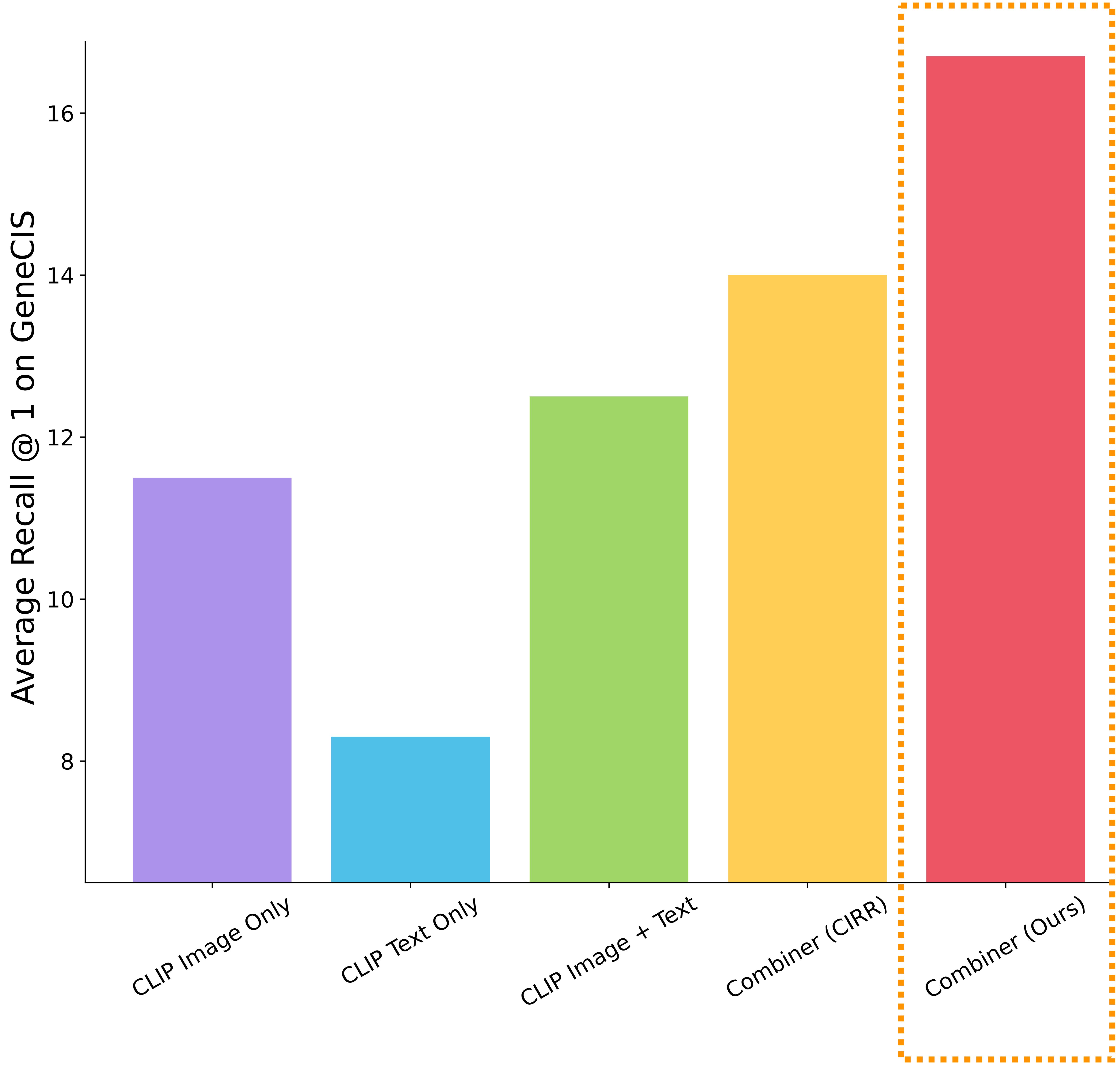
Dataset of 30k triplets



[1] Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models, Liu et. al, ICCV 2021

Results

Ours trained on 1.6M **automatically curated** triplets



Further Analysis

- **Zero-shot** evaluation of our model outperforms many **supervised** baselines on similar benchmarks
- Our model gets state-of-the-art on MIT-States, despite zero-shot evaluation

CIRR

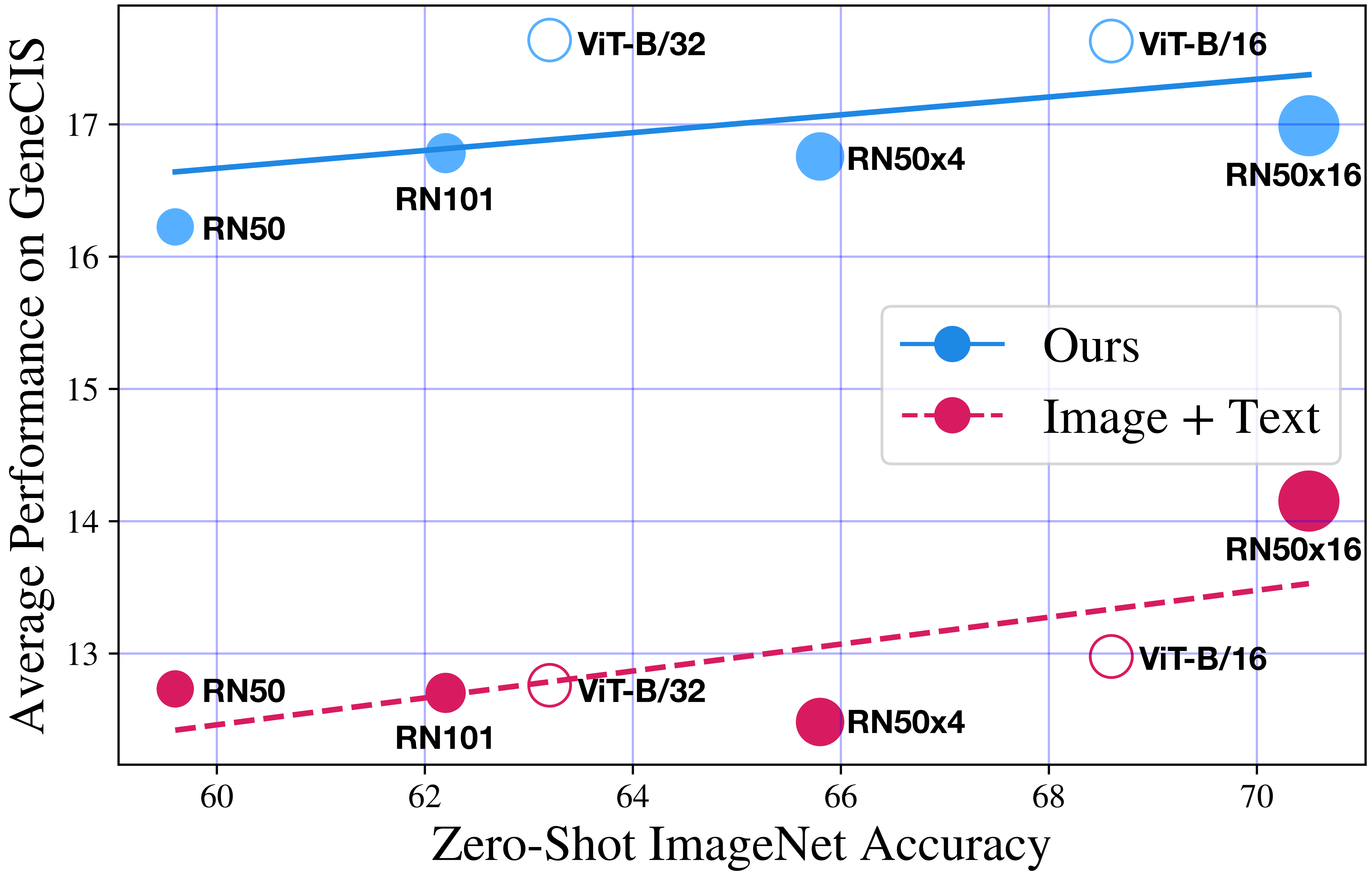
	Zero-shot	Recall @ 1	Recall @ 5	Recall @ 10
ARTEMIS [17]	✗	17.0	46.1	61.3
CIRPLANT [45]	✗	19.6	52.6	68.4
Combiner (CIRR, [4])	✗	38.5	70.0	81.9
Combiner (CIRR, improved)	✗	40.9	73.4	84.8
Image Only	✓	7.5	23.9	34.7
Text Only	✓	20.7	43.9	56.1
Image + Text	✓	21.8	50.9	63.7
Combiner (CC3M, Ours)	✓	**27.3**	**57.0**	**71.1**

MIT-States

	Zero-shot	Recall @ 1	Recall @ 5	Recall @ 10
TIRG [74]	✗	12.2	31.9	43.1
ComposeAE [2]	✗	13.9	35.3	47.9
LBF [28]	✗	14.7	35.3	46.6
HCL [79]	✗	15.2	36.0	46.7
MAN [20]	✗	15.6	36.7	47.7
Image Only	✓	3.7	14.0	22.9
Text Only	✓	11.2	21.7	11.2
Image + Text	✓	12.8	31.4	42.5
Combiner (CC3M, Ours)	✓	**15.6**	**37.5**	**49.2**

Further Analysis

- GeneCIS performance is only **weakly correlated with ImageNet accuracy** of backbone
    - In contrast to common vision tasks like detection and segmentation
- GeneCIS probes an **important** but **orthogonal** visual capability to most benchmarks



Thank you for listening

Email:
sagar@robots.ox.ac.uk

Project page (+QR Link):
https://sgvaze.github.io/genecis/