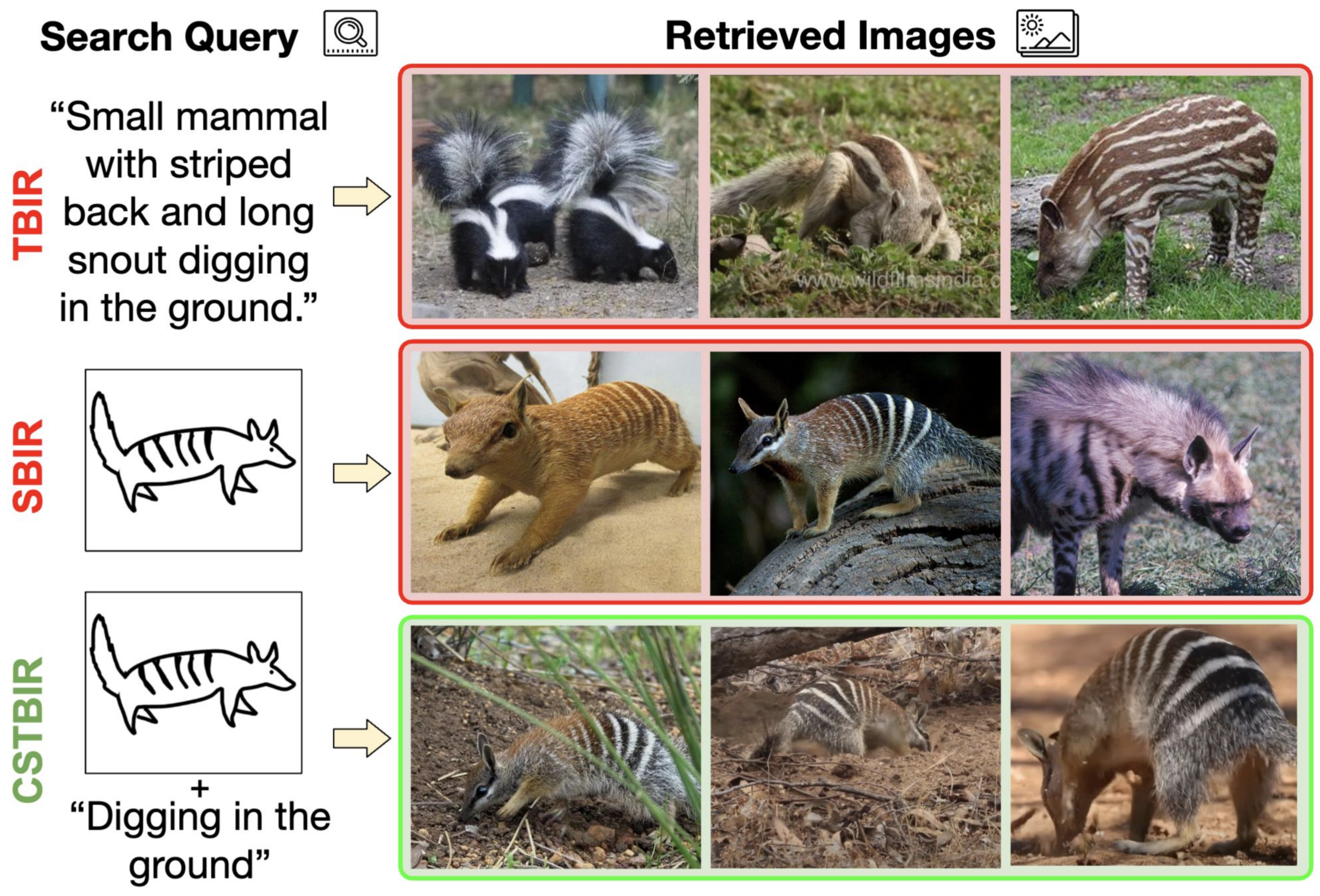
Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions

Prajwal Gatti[1], Kshitij Parikh[1], Dhriti Prasanna Paul[1],
Manish Gupta[2], Anand Mishra[1]

[1]Indian Institute of Technology Jodhpur, [2]Microsoft

Searching for objects in scenes with sketch+text queries

The CSTBIR Task



Search Query

Retrieved Images

TBIR

“Small mammal with striped back and long snout digging in the ground.”

SBIR

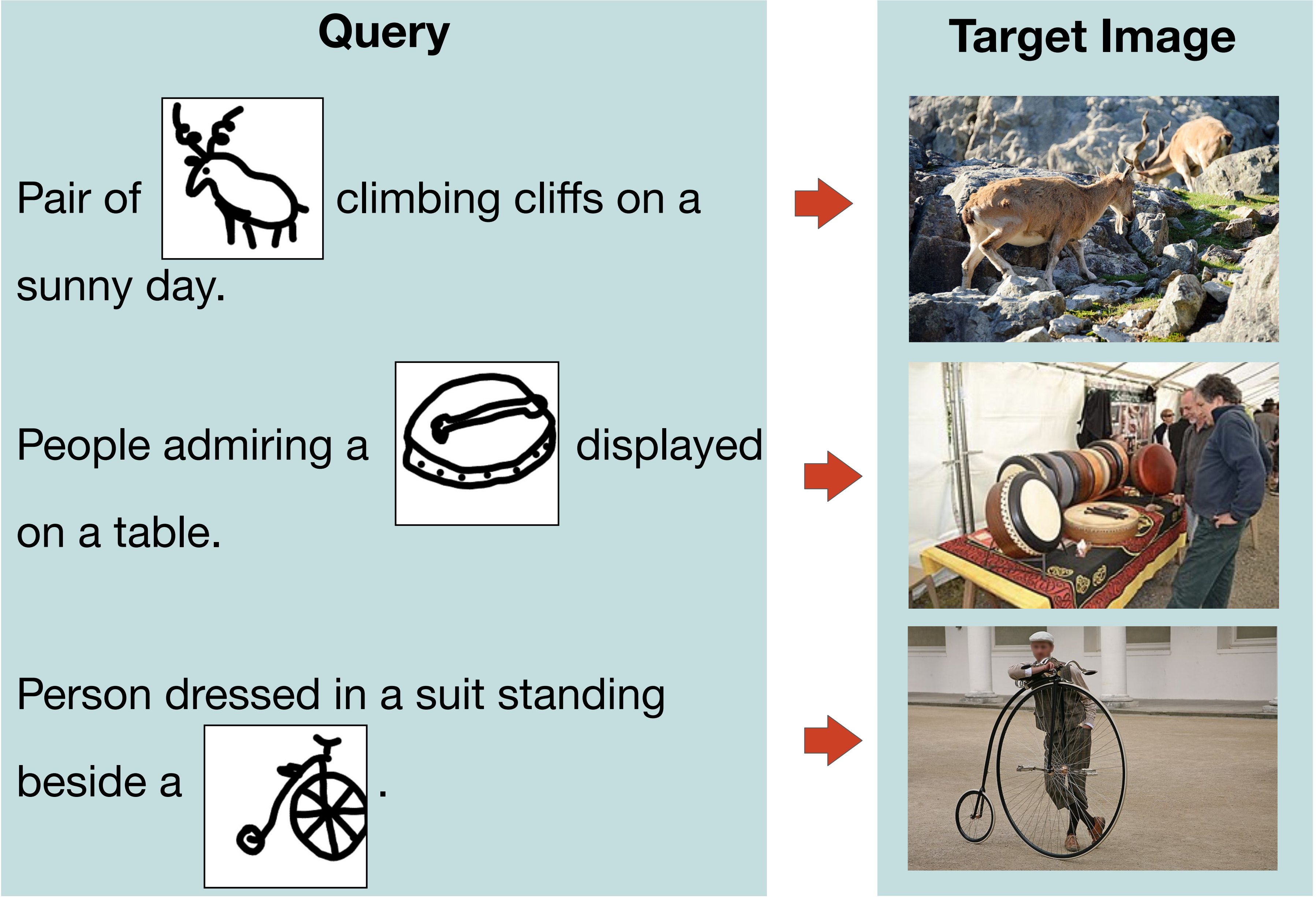
CSTBIR

+ “Digging in the ground”

Given: a hand-drawn sketch **S**, a complementary text **T** and a database **D** of **N** natural scene images with multiple objects

Rank: **N** images according to relevance to composite ⟨**S**, **T**⟩ query.

CSTBIR Dataset



Query

Target Image

Pair of climbing cliffs on a sunny day.

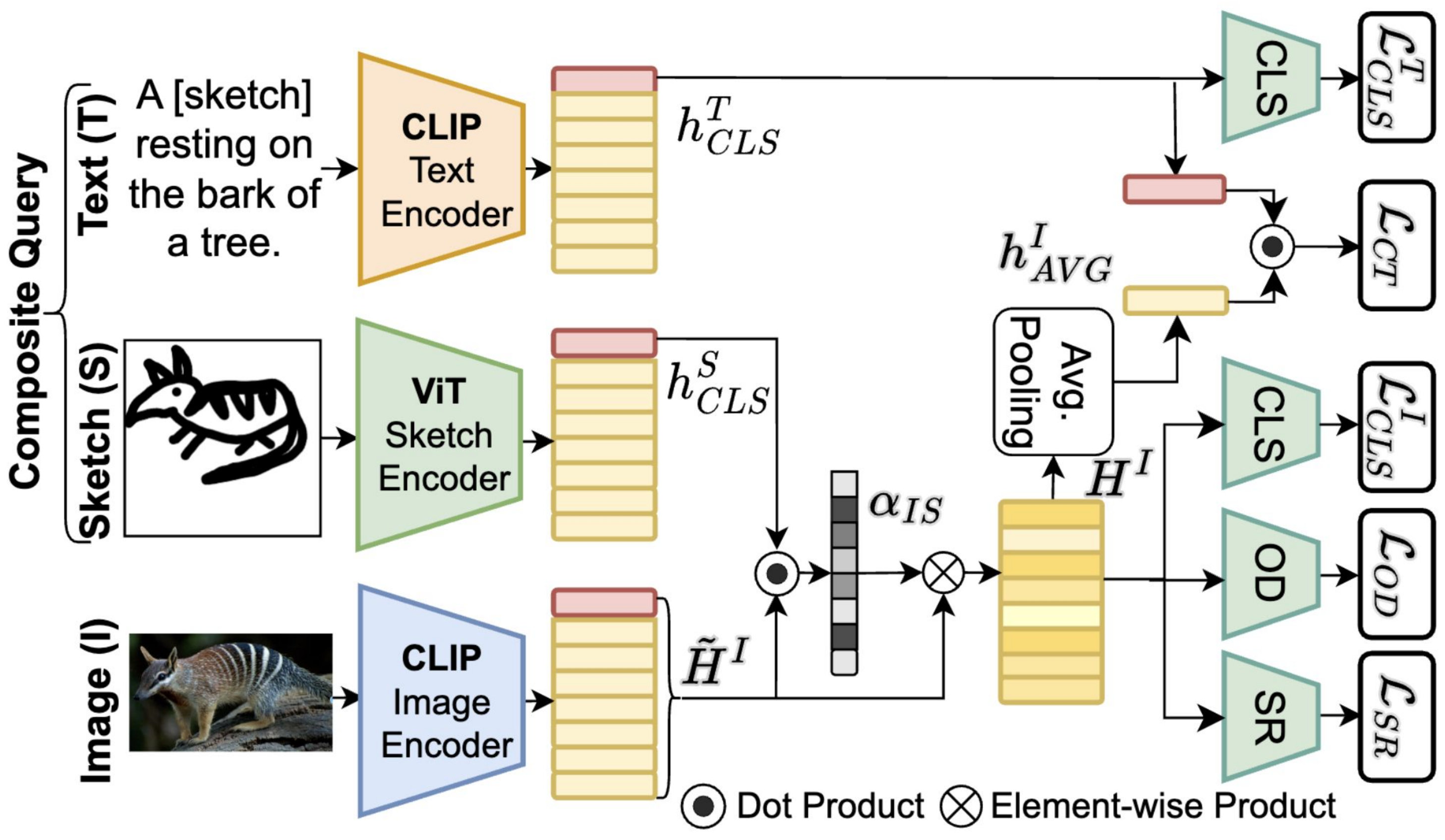
People admiring a displayed on a table.

Person dressed in a suit standing beside a .

(objects: markhor, bodhran, penny-farthing)

- Natural images and text descriptions from Visual Genome and sketches from Quick, Draw!
- Train (~**1.89M** queries, ~97K images, 258 object classes)
- Validation (~**5K** images, ~97K queries)
- Test-1K: 1K queries, 1K images
- Test-5K: 4K queries, 5K images
- Open-Category set: 750 queries, 70 objects, 1K images.

STNet: Sketch+Text Network



Composite Query

Text (T)

A [sketch] resting on the bark of a tree.

Sketch (S)

Image (I)

CLIP Text Encoder

ViT Sketch Encoder

CLIP Image Encoder

⊙ Dot Product ⊗ Element-wise Product

Training objectives: (i) Contrastive Training, (ii) Object Classification, (iii) Sketch-Guided Object Localization, and (iv) Sketch Reconstruction

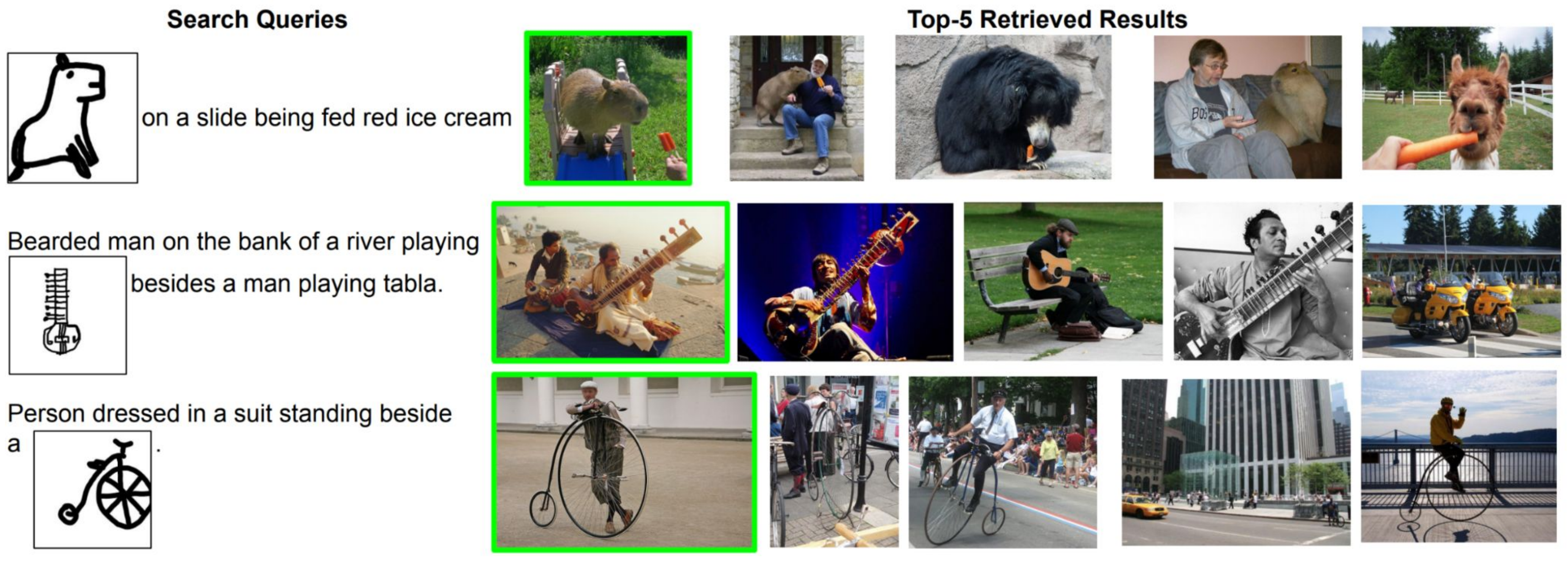
Results

	Method	R@10 ↑		R@20 ↑		R@50 ↑		R@100 ↑		MdR ↓	
		T1K	T5K	T1K	T5K	T1K	T5K	T1K	T5K	T1K	T5K
Sketch	Doodle2Search	14.3	3.6	24.5	6.7	36.2	14.5	45.7	24.4	129.0	573.5
	DeepSBIR	5.2	1.6	8.8	3.0	18.9	5.7	27.4	9.5	258.5	1288.0
	ViT-Siamese	20.4	5.2	34.2	9.9	51.0	22.2	62.6	34.9	48.0	233.0
Text	VisualBERT	23.3	7.6	35.9	15.4	40.8	27.8	54.0	40.2	46.0	246.0
	ViLT	28.1	10.5	42.7	16.5	60.2	30.1	74.3	43.8	30.0	163.0
	CLIP	50.6	24.2	63.1	33.7	78.8	49.1	86.7	62.5	10.0	52.0
Sketch+Text	TIRG	31.9	10.4	44.2	17.3	62.8	31.6	73.2	45.4	27.5	128.0
	Taskformer	22.4	9.3	35.6	14.8	42.3	27.6	53.8	38.3	48.0	204.0
	Two-stage	67.0	34.8	77.4	46.9	88.6	64.7	**93.7**	**76.2**	5.0	24.0
	Two-stage (desc)	60.1	30.5	73.7	41.7	85.5	59.6	91.6	72.0	7.0	32.0
	STNet (Ours)	**73.7**	**38.7**	**80.6**	**50.0**	**89.4**	**64.6**	93.5	74.5	**3.0**	**20.5**

Table 1: Performance comparison on the CSTBIR test-1K/5K.

M	Query	Objective	R@10	R@20	R@50	R@100	MdR
1	S	$\mathcal{L}_{CT}$	20.2	33.7	50.9	62.9	50.5
2	T	$\mathcal{L}_{CT}$	50.6	63.1	78.8	86.7	10.0
3	T+S	$\mathcal{L}_{CT}$	68.4	77.2	85.6	89.8	5.0
4	T+S	$\mathcal{L}_{CT} + \mathcal{L}_{OD} + \mathcal{L}_{SR}$	69.4	80.4	85.6	90.4	5.0
5	T+S	$\mathcal{L}_{CT} + \mathcal{L}_{CLS} + \mathcal{L}_{SR}$	70.4	79.6	86.2	91.1	5.0
6	T+S	$\mathcal{L}_{CT} + \mathcal{L}_{CLS} + \mathcal{L}_{OD}$	71.2	79.0	87.0	93.0	4.0
7	T+S	(6) + $\mathcal{L}_{SR}$	**73.7**	**80.6**	**89.4**	**93.5**	**3.0**

Table 2: Ablation study for STNet model



Search Queries

Top-5 Retrieved Results

on a slide being fed red ice cream

Bearded man on the bank of a river playing besides a man playing tabla.

Person dressed in a suit standing beside a .

Summary

CSTBIR: Composite Sketch+Text Based Image Retrieval Task

New Dataset: containing ~2M queries and ~108K natural images

STNet: Pre-trained multimodal transformer based method with task-specific training objectives.

AAAI 2024

This work is supported by the Science and Engineering Research Board (SERB) Grant. DST No: SRG/2021/001948.

Paper and resources available here: https://vl2g.github.io/projects/cstbir/