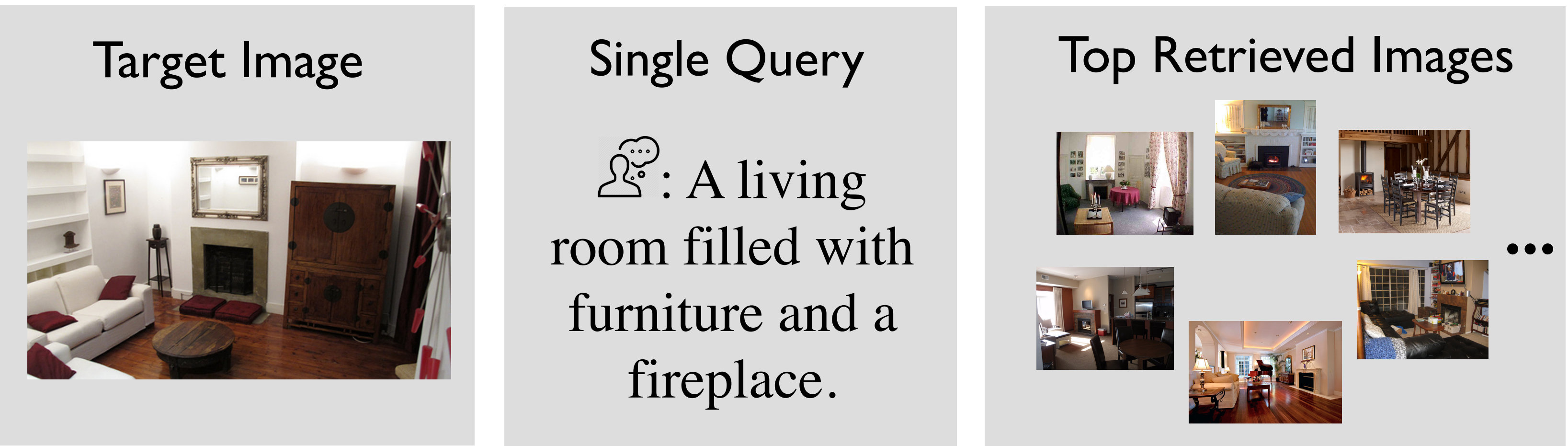
Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries

Fuwen Tan[1], Paola Cascante-Bonilla[1], Xiaoxiao Guo[2], Hui Wu[2] Song Feng[2],Vicente Ordonez[1]

[1]University of Virginia, [2]IBM Research AI

Retrieving an Image using a Single Query is HARD

Especially when the scene is complex

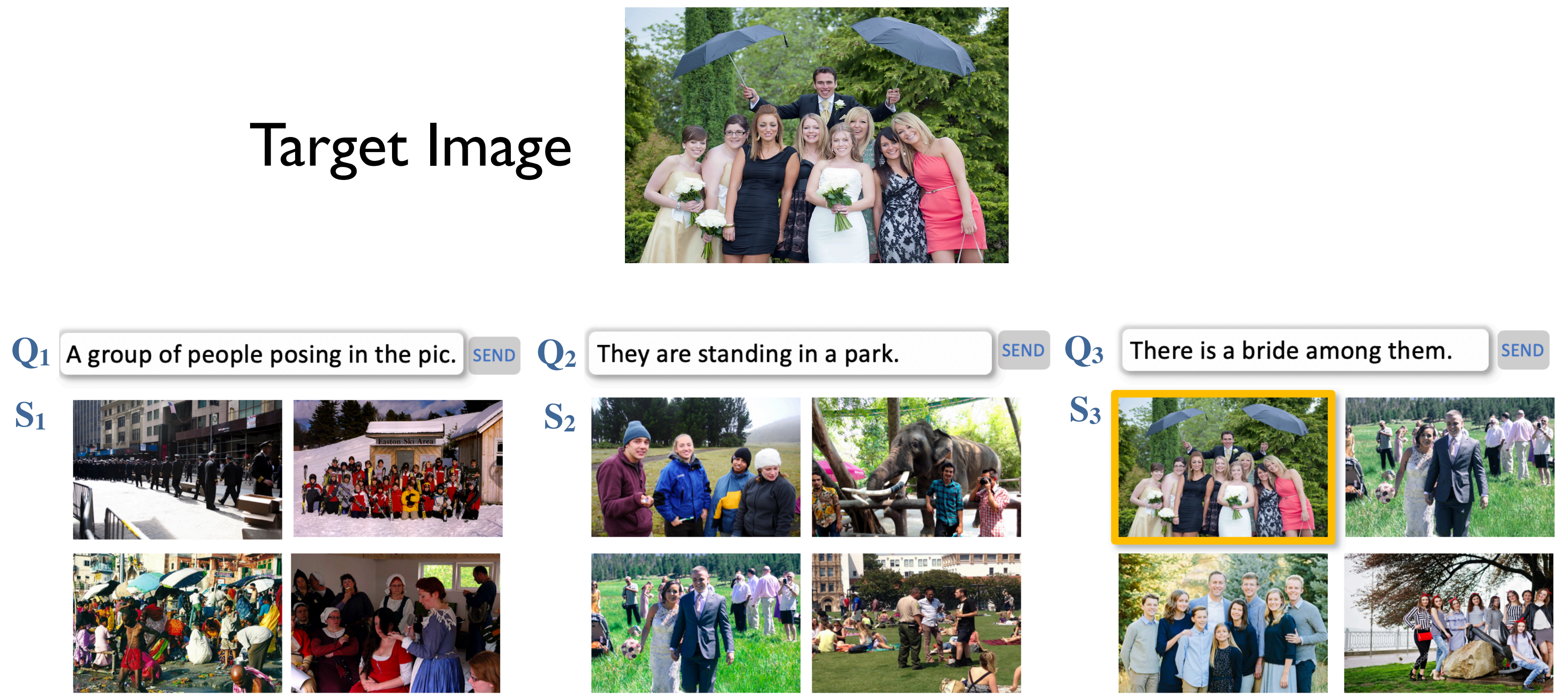

Target Image

Single Query

: A living room filled with furniture and a fireplace.

Top Retrieved Images

Goal: Retrieving an Image by Multiple Round Queries


Target Image

Q1 A group of people posing in the pic. SEND

Q2 They are standing in a park. SEND

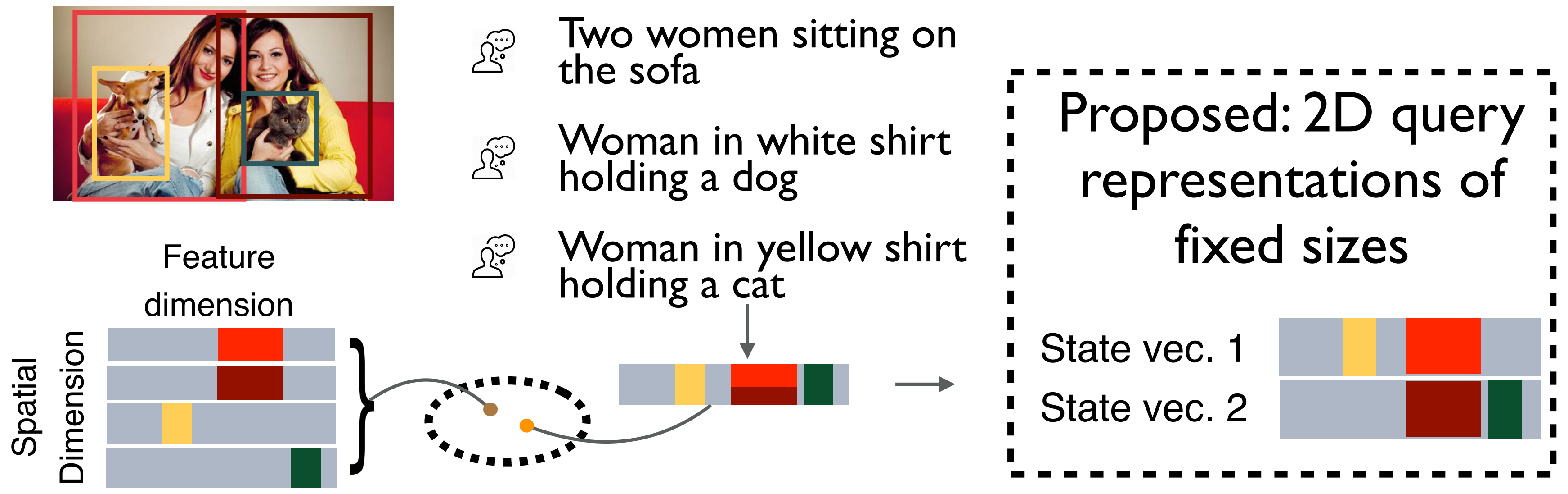
Q3 There is a bride among them. SEND

Contributions

- Drill-down, an **interactive image search** approach with multiple round queries which leverages region captions as a form of **weak supervision** during training;
- A **compact representation**, outperforming competing baseline methods by a significant margin;
- Experiments on a **large-scale natural image dataset**:Visual Genome, demonstrating superior performance of our model on both simulated and real user queries.

Observation

1D query representations can NOT distinguish entities sharing the same feature space

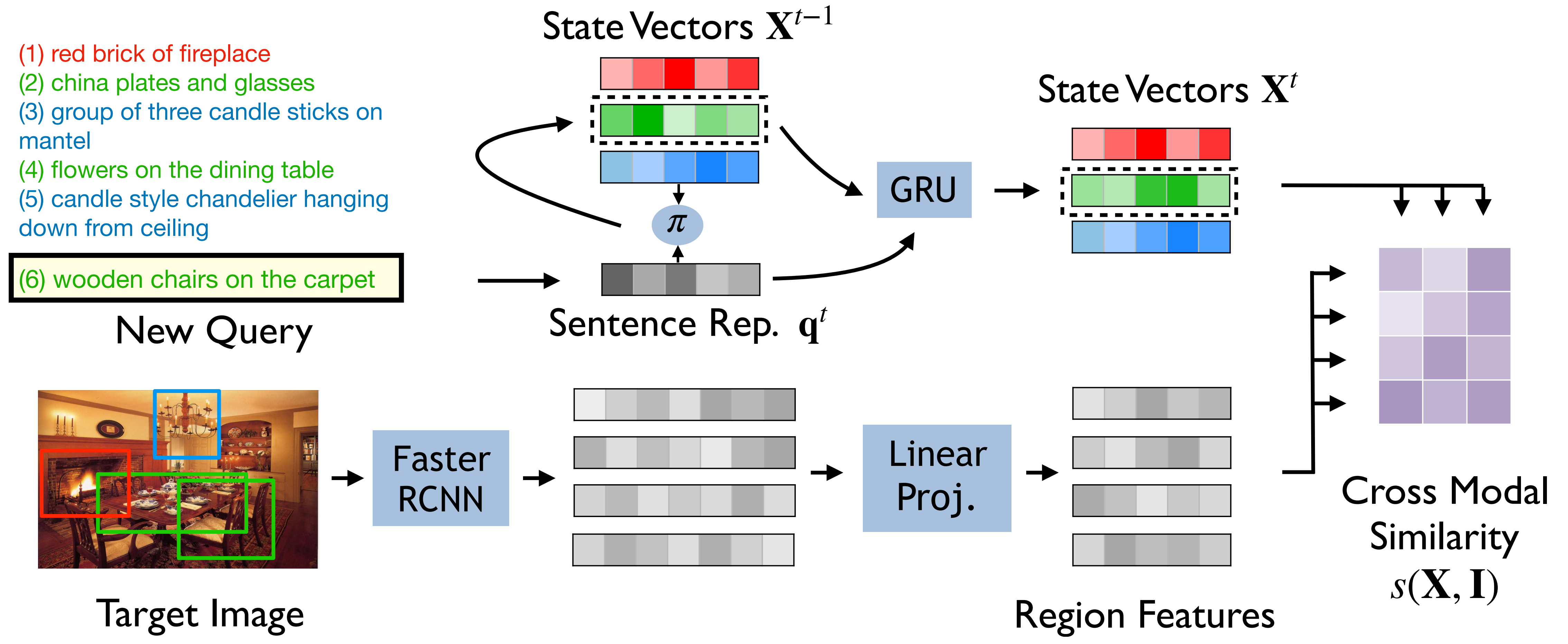

Two women sitting on the sofa

Woman in white shirt holding a dog

Woman in yellow shirt holding a cat

Proposed: 2D query representations of fixed sizes

Model


(1) red brick of fireplace

(2) china plates and glasses

(3) group of three candle sticks on mantel

(4) flowers on the dining table

(5) candle style chandelier hanging down from ceiling

(6) wooden chairs on the carpet

New Query
State Vectors $\mathbf{X}^{t-1}$
π
Sentence Rep. $\mathbf{q}^t$
GRU
State Vectors $\mathbf{X}^t$
Target Image
Faster RCNN
Linear Proj.
Region Features
Cross Modal Similarity $s(\mathbf{X}, \mathbf{I})$

Region Captions as Weak Supervision

Training

Q1:A cat on the left

Q2:A dog on the right

Q3:A small dog

VISUALGENOME

	Training	Validation	Testing
Samples	92105	5000	9896

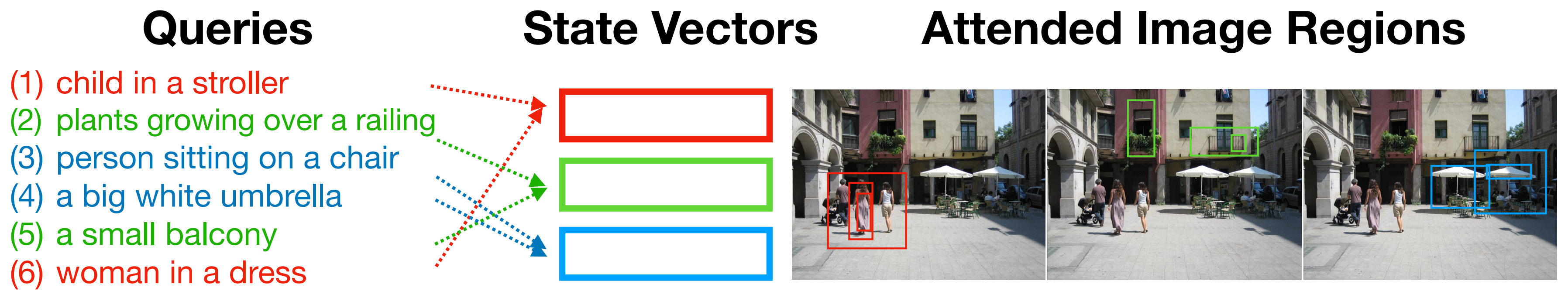
Pros:

- Free, no extra annotation
- "Abstract" of real queries,
  - More invariant signals, e.g. image content
  - Fewer irrelevant signals, e.g. speaking style

Cons:

- Domain shift

Interpretable & Compact Representation


Queries

(1) child in a stroller

(2) plants growing over a railing

(3) person sitting on a chair

(4) a big white umbrella

(5) a small balcony

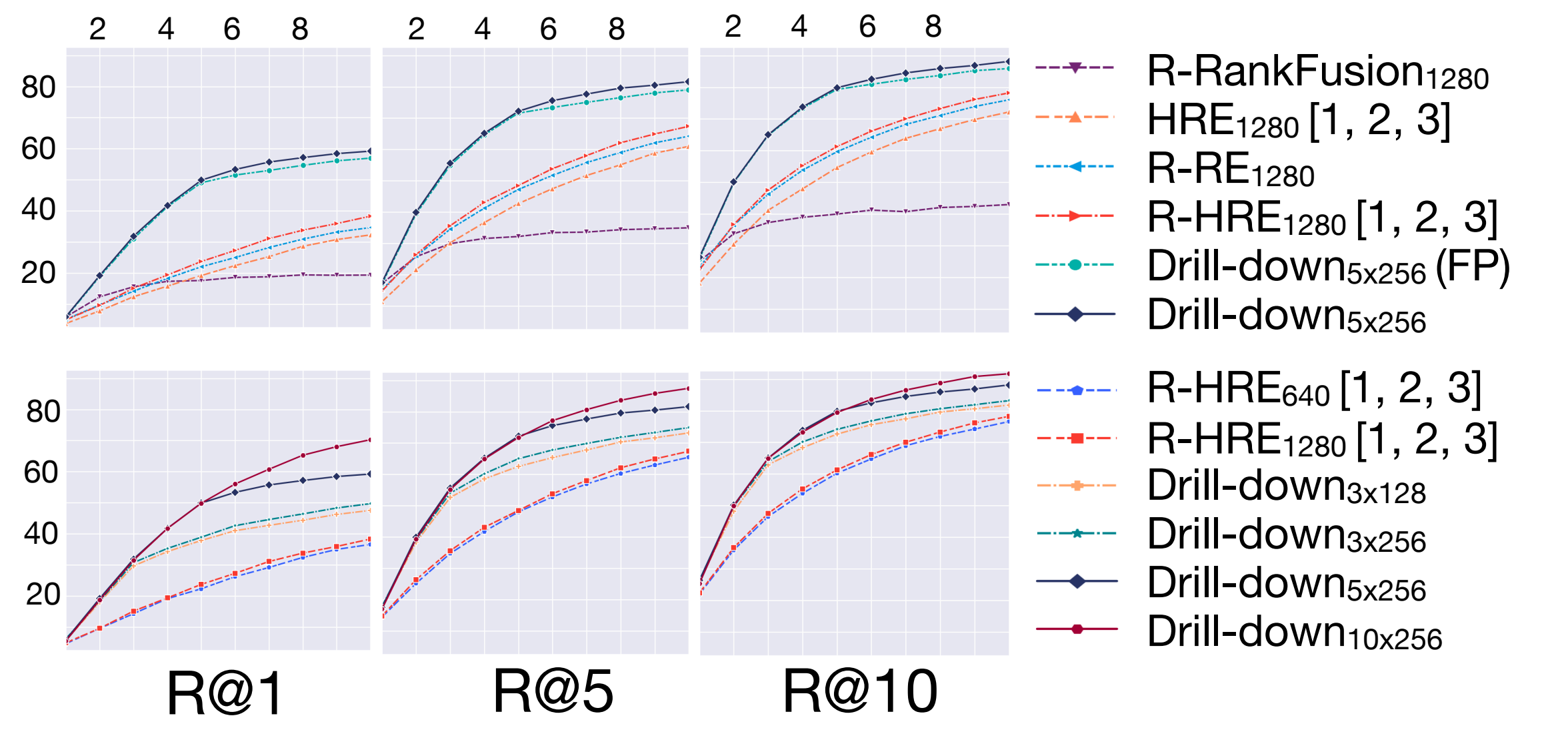
(6) woman in a dress

State Vectors

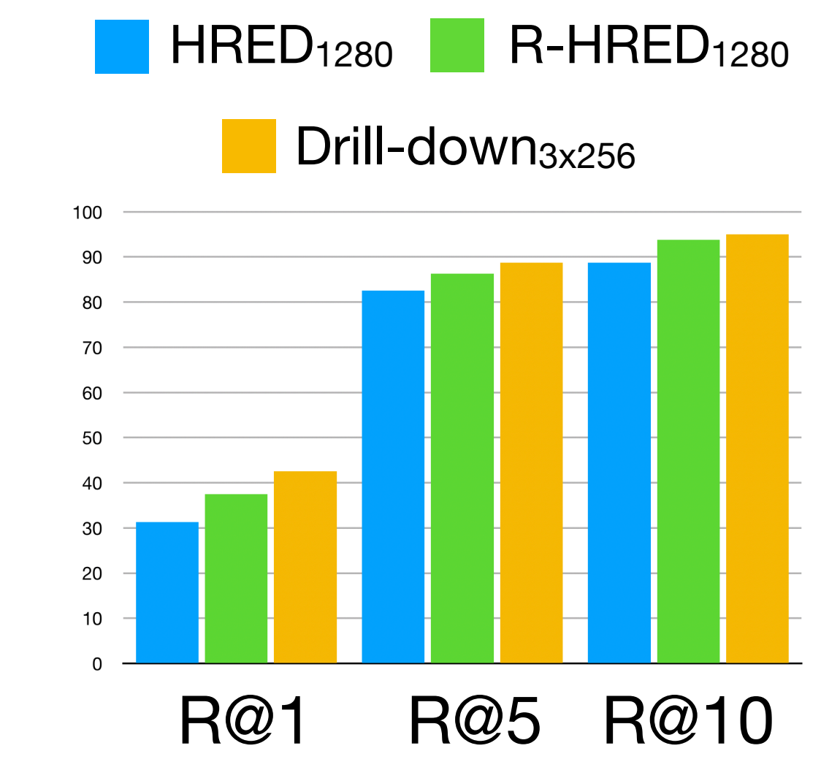
Attended Image Regions

Methods	R-RE$_{1280}$/HRE$_{1280}$ [1, 2, 3]	R-HRE$_{640/1280}$ [1, 2, 3]	Drill-down$_{3\times128}$ / $_{3\times256}$ / $_{5\times256}$
# Query Rep.	1280	640 / 1280	384 / 768 / 1280
# Image Rep.	1280 / 36 × 1280	36×640 / 36 × 1280	36 × 128 / 36 × 256 / 36 × 256
# Parameters	22820k	9866k / 22820k	4861k / 5830k / 5830k

Evaluations on Simulated/Real Scenarios



Evaluations on Region Captions



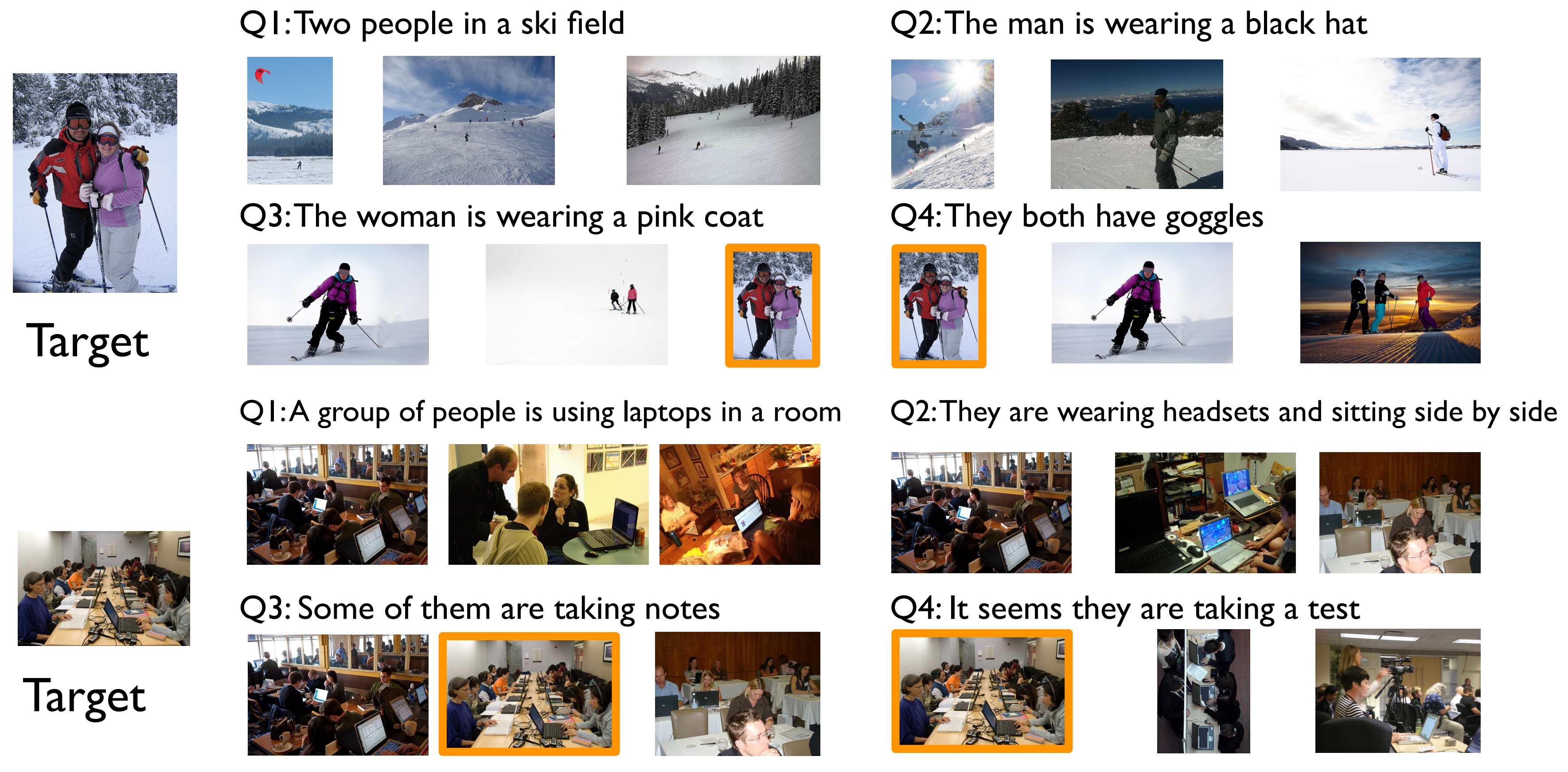
Human Evaluations

[1] *Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models*. Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, Joelle Pineau. AAAI 2016

[2] *Knowledge-aware multimodal dialogue systems*. Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. ACM MM 2018

[3] *Dialog-based interactive image retrieval*. Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. NeurIPS 2018

Multiple Round Retrieval Examples


Q1:Two people in a ski field

Q2:The man is wearing a black hat

Q3:The woman is wearing a pink coat

Q4:They both have goggles

Target

Q1:A group of people is using laptops in a room

Q2:They are wearing headsets and sitting side by side

Q3: Some of them are taking notes

Q4: It seems they are taking a test

Target