Composing Object Relations and Attributes for Image-Text Matching

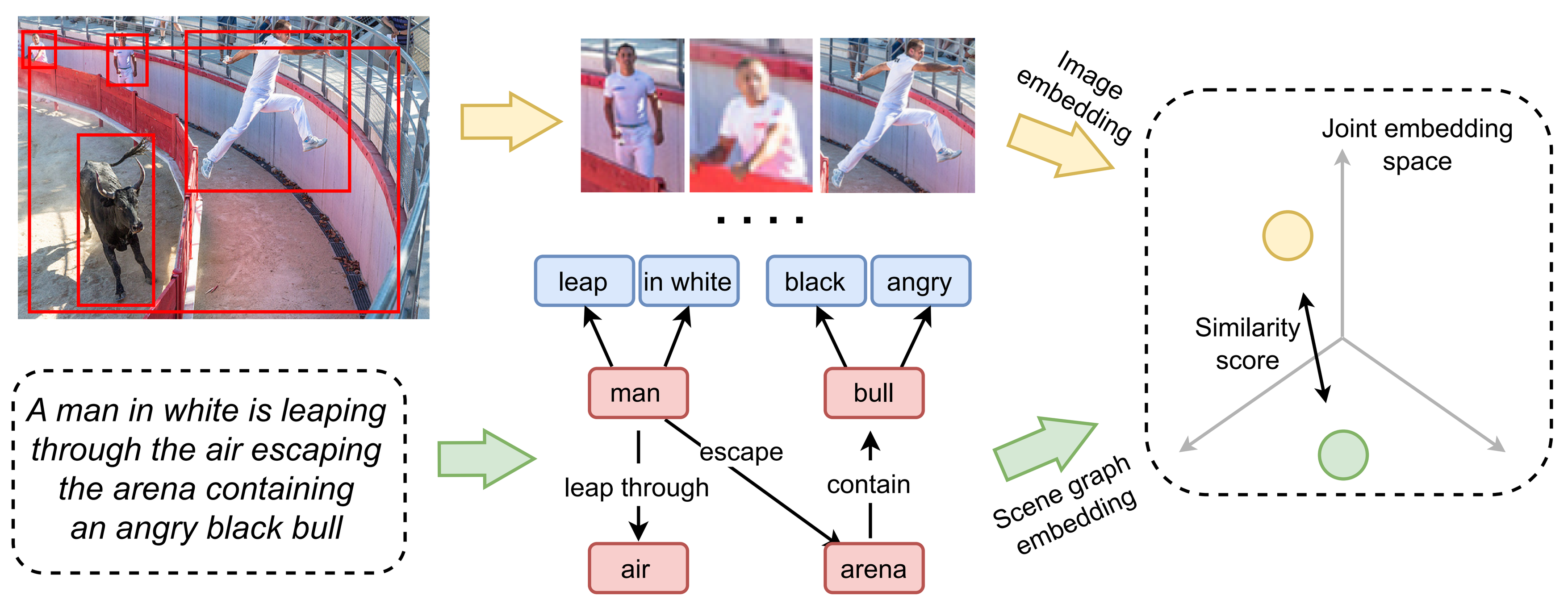
Khoi Pham[1], Chuong Huynh[1], Ser-Nam Lim[2], Abhinav Shrivastava[1]

[1]University of Maryland, College Park [2]University of Central Florida

GitHub

CVPR JUNE 17-21, 2024 SEATTLE, WA

Introduction

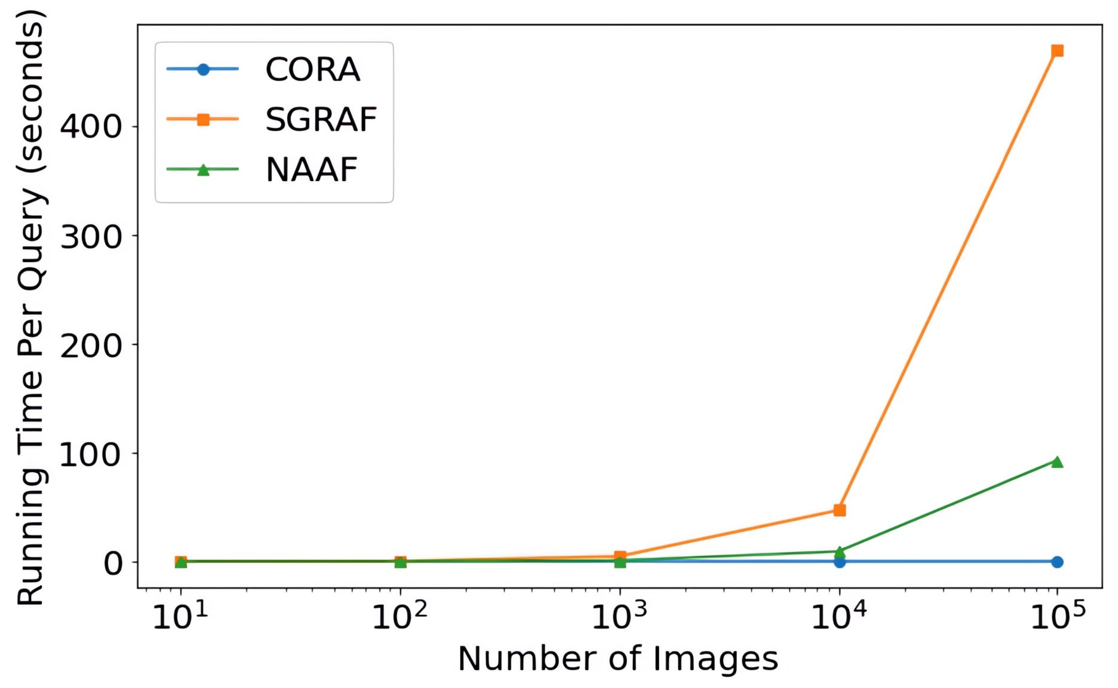


Motivations

- Cross-attention networks are powerful but more computationally expensive than dual encoders.
- Text encoder (GRU, LSTM, BERT) needs to learn semantic parsing: 1) Which tokens are objects/attributes/relations? 2) How to bind attribute/relation with the correct object?
- Prior work only align images with texts holistically. Can we design objectives to align image with individual object entity?

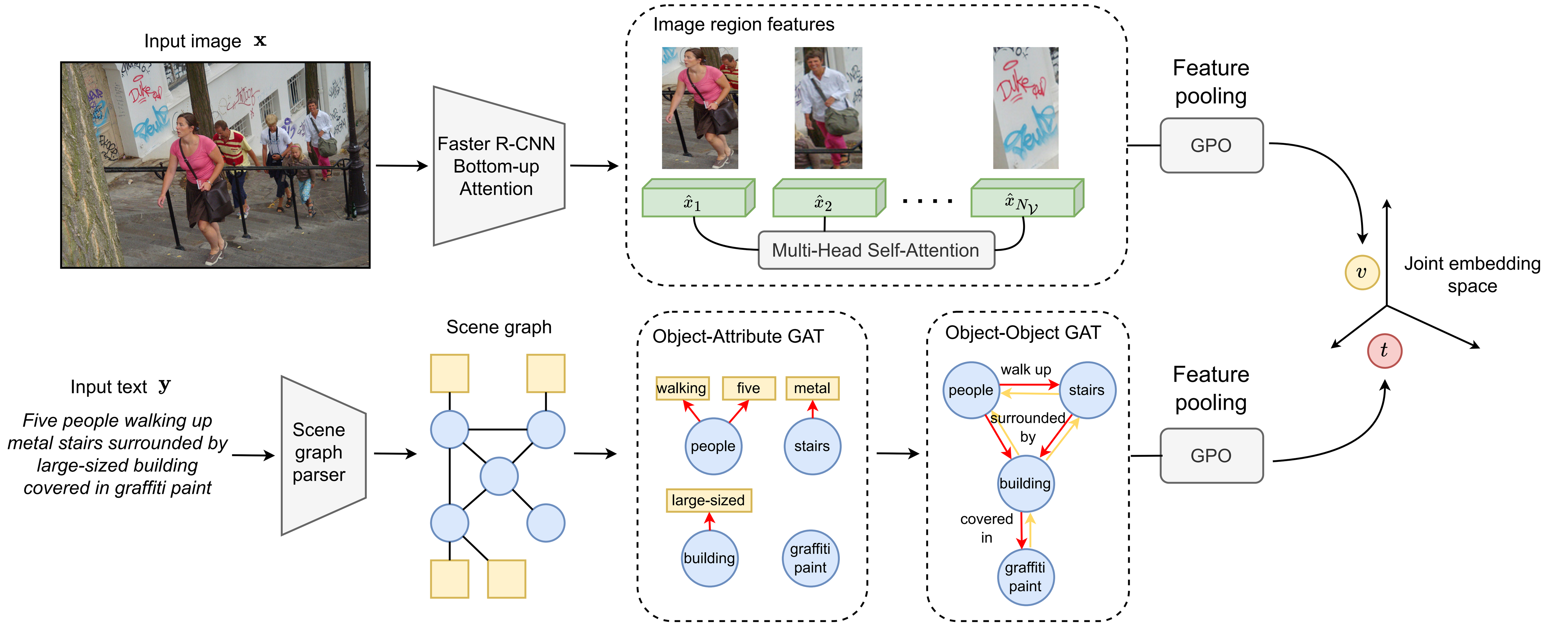
Proposal – CORA:

- A scene graph-based dual encoder for image-text matching.
- Is trained with objectives to make global image-text and local image-object entity alignment.

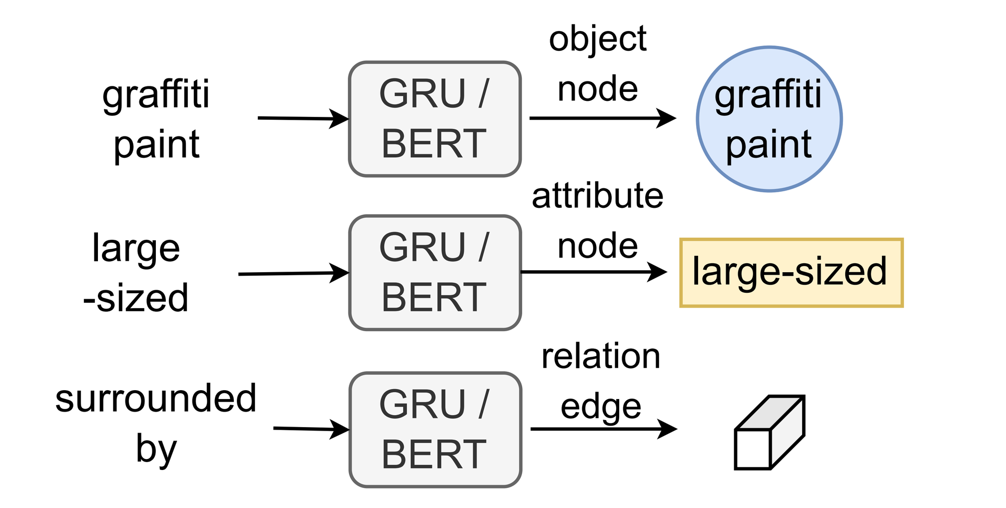


- CORA is superior to recent SOTA cross-attention image-text matching methods in terms of retrieval speed and accuracy.

CORA Framework

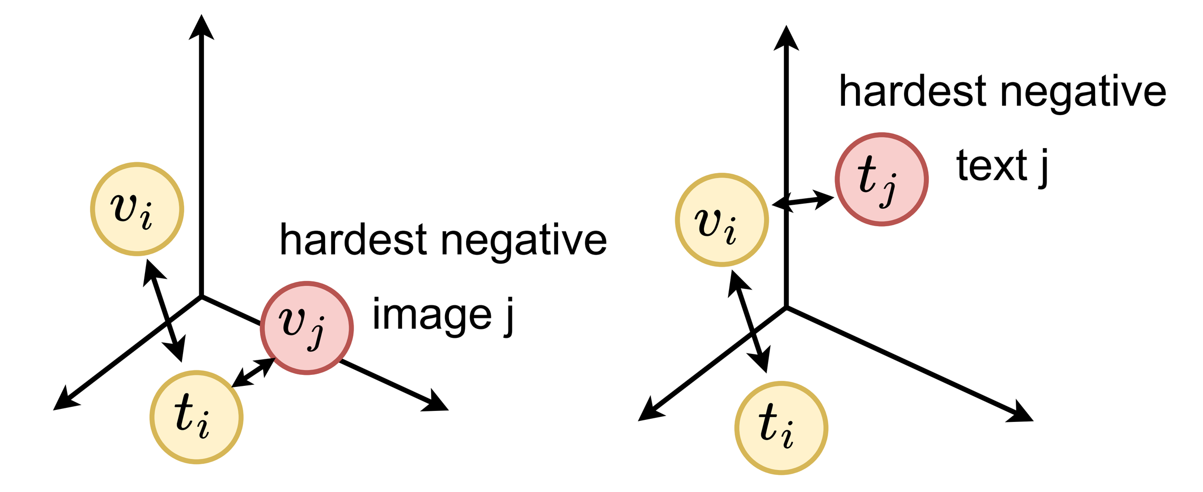


Semantic Concept Encoding



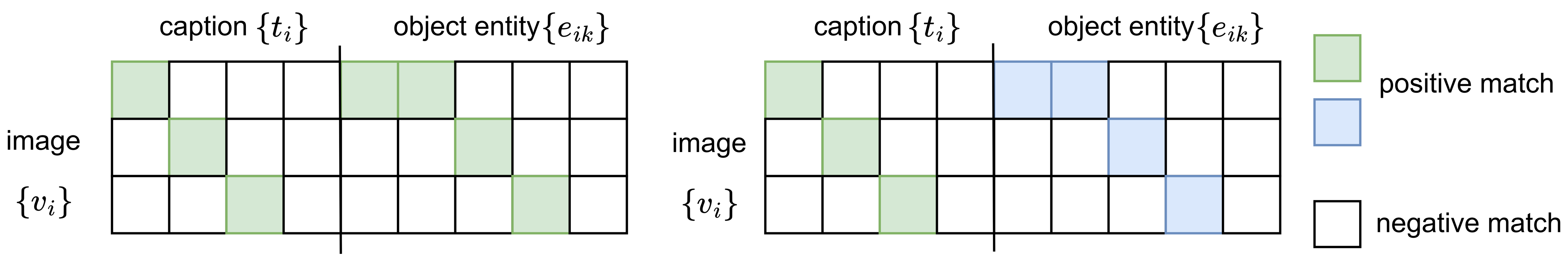
Training. Given pairwise similarity score in a batch:

1) Triplet hardest loss



2) Contrastive loss: green >> white

3) Triplet specificity loss: green >> blue



Experiments

Dataset	Method	Venue	CA	Image → Text R@1	Image → Text R@5	Image → Text R@10	Text → Image R@1	Text → Image R@5	Text → Image R@10	RSUM
Faster R-CNN + Bi-GRU										
Flickr30K	CHAN	CVPR'23	✓	79.7	94.5	97.3	60.2	85.3	90.7	507.7
	NAAF†	CVPR'23	✓	81.9	96.1	98.3	61.0	85.3	90.6	513.2
	SDE†	CVPR'23		80.9	94.7	97.6	59.4	85.6	91.1	509.3
	HREM†	CVPR'23		81.4	96.5	98.5	60.9	85.6	91.3	514.2
	Ours†			82.3	96.1	98.0	63.0	87.4	92.8	**519.6**
MS-COCO	CHAN	CVPR'23	✓	60.2	85.9	92.4	41.7	71.5	81.7	433.4
	NAAF†	CVPR'23	✓	58.9	85.2	92.0	42.5	70.9	81.4	430.9
	SDE†	CVPR'23		60.4	86.2	92.4	42.6	73.1	83.1	437.8
	HREM†	CVPR'23		60.6	86.4	92.5	41.3	71.9	82.4	435.1
	Ours†			63.0	86.8	92.7	44.2	73.9	84.0	**444.6**
Faster R-CNN + BERT										
Flickr30K	MV-VSE†	IJCAI'22		82.1	95.8	97.9	63.1	86.7	92.3	517.5
	CHAN	CVPR'23	✓	80.6	96.1	97.8	63.9	87.5	92.6	518.5
	HREM†	CVPR'23		84.0	96.1	98.6	64.4	88.0	93.1	**524.2**
	Ours†			83.4	95.9	98.6	64.1	88.1	93.1	<u>523.3</u>
MS-COCO	MV-VSE†	IJCAI'22		59.1	86.3	92.5	42.5	72.8	83.1	436.3
	CHAN	CVPR'23	✓	59.8	87.2	93.3	44.9	74.5	84.2	443.9
	HREM†	CVPR'23		64.0	88.5	93.7	45.4	75.1	84.3	**451.0**
	Ours†			64.3	87.5	93.6	45.4	74.7	84.6	<u>450.1</u>

- New state-of-the-art results on Flickr30K while being competitive on MS-COCO. Superior performance compared to even cross-attention methods.

Image-to-text retrieval

1. A woman dressed in black with a tattoo on her right arm is taking a picture...
2. A woman with long hair in black clothing is taking a photograph.
3. A person with tattoos is looking at a photo on a digital camera , or cellphone.
4. A tattooed woman taking a picture with a digital camera.
5. Somebody took a photo of a girl with long black hair taking a photo.

Image-to-entity retrieval:

digital camera, camera lens, woman wearing black, gun range, mobile phone, photographer, black blouse, black backpack, black purse, black leather pumps, black leather bag, dark haired woman

A large white dog sits on a bench with people next to a path .