BiVLC: Extending Vision-Language Compositionality Evaluation with Text-to-Image Retrieval

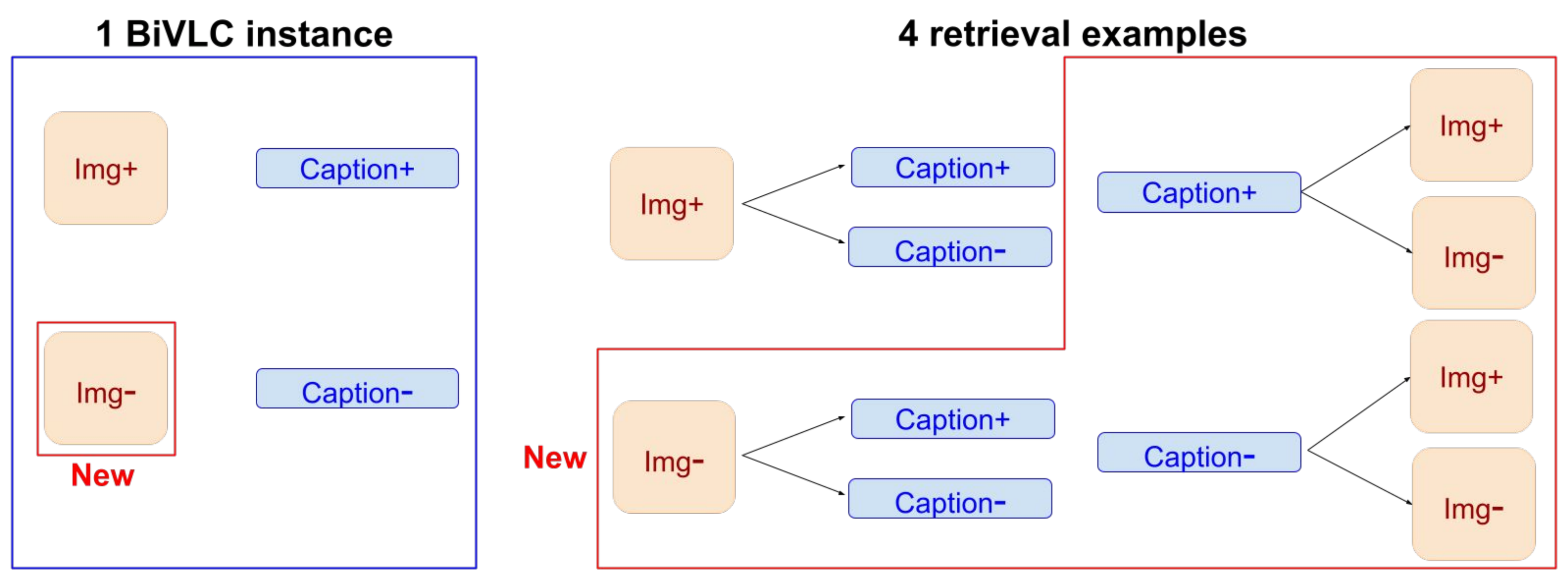
HiTZ Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology
Universidad del País Vasco / Euskal Herriko Unibertsitatea

Imanol Miranda, Ander Salaberria, Eneko Agirre, Gorka Azkune
HiTZ Center – Ixa, University of the Basque Country (UPV/EHU)

NEURAL INFORMATION PROCESSING SYSTEMS

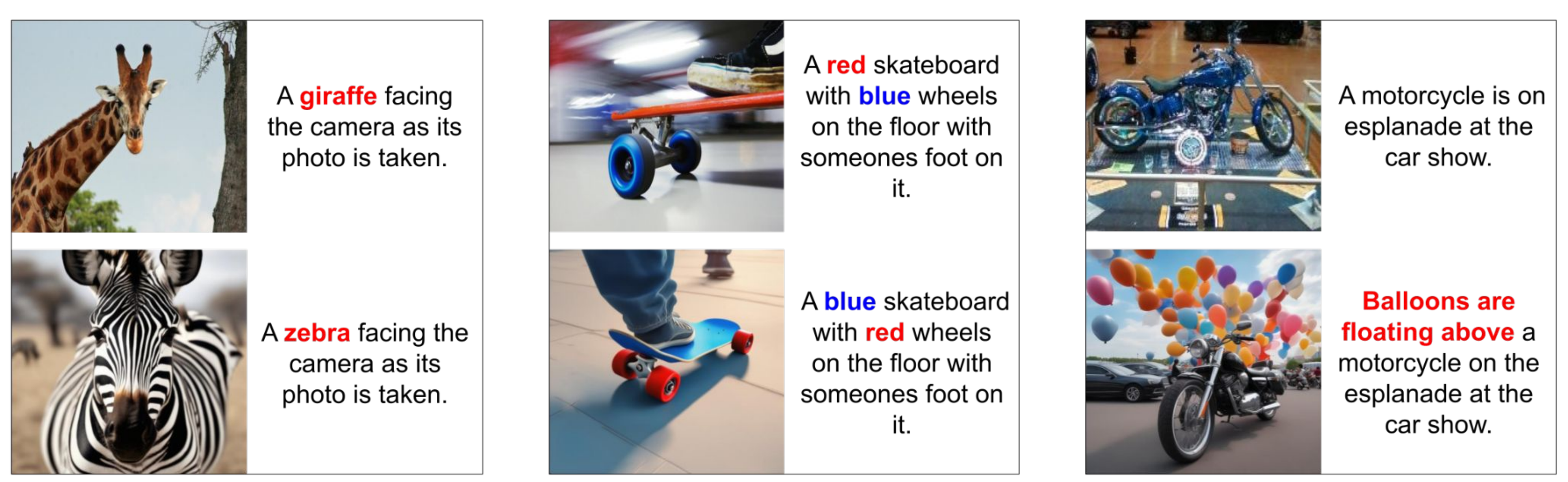
Motivation

Previous datasets focused mainly on image-to-text retrieval. Why don't we include text-to-image retrieval also?



What is BiVLC?

BiVLC is a **Bi**directional **V**ision-**L**anguage **C**ompositionality dataset with almost 3k instances formed by 2 images and 2 captions.

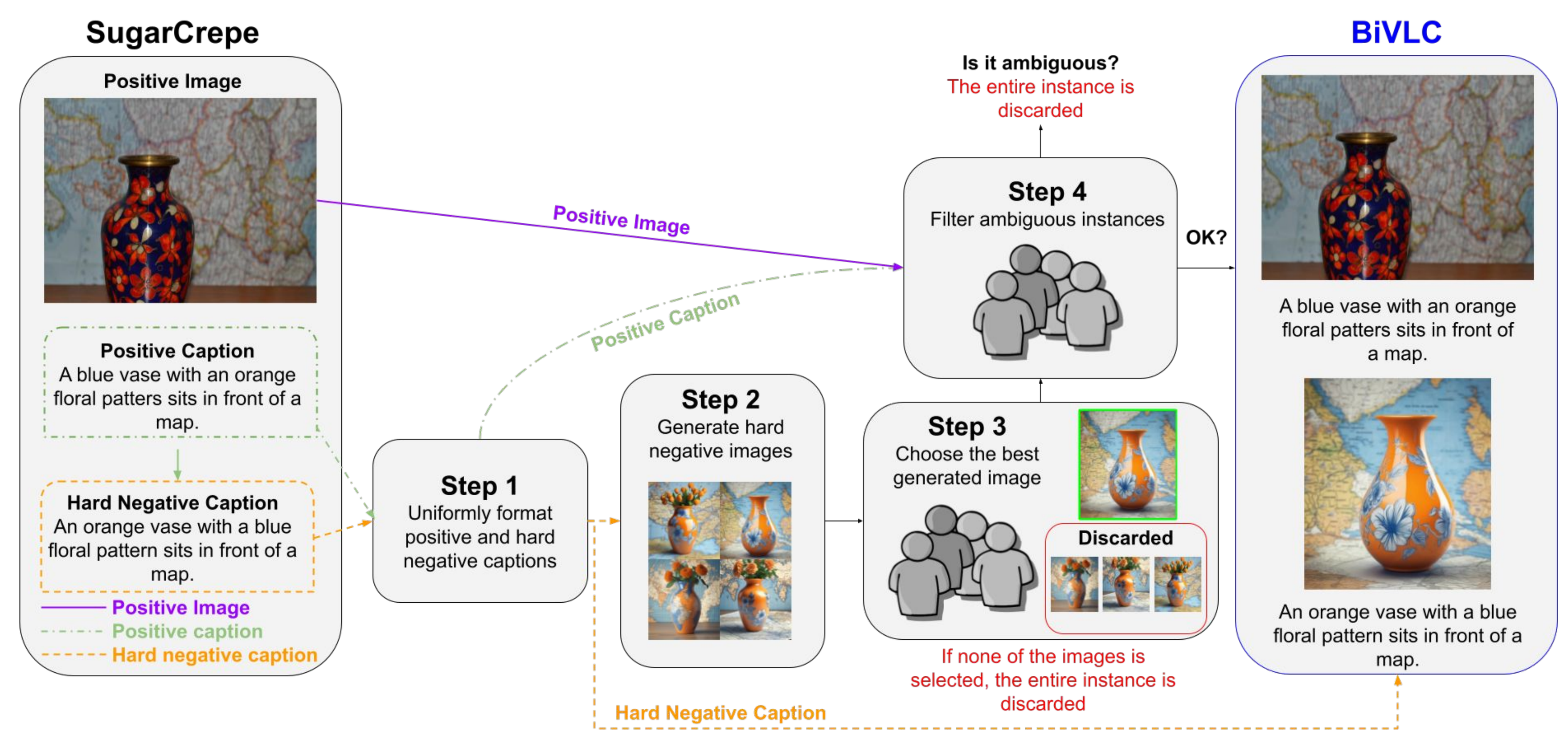


Replace | **Swap** | **Add**

Dataset	I2T	T2I	Replace Obj	Replace Att	Replace Rel	Swap Obj	Swap Att	Swap Rel	Add Obj	Add Att	Total
Winoground	✓	✓				668		1,036			1,600†
SugarCrepe	✓		1,652	788	1,406	246	666		2,062	692	7,512
BiVLC (ours)	✓	✓	4,800	1,748	1,848	324	1,112		1,596	304	11,732

How is BiVLC constructed?

We propose a semi-automatic dataset construction method:



Findings with BiVLC

We evaluated SOTA models in SugarCrepe and BiVLC divided into Contrastive and Generative.

	Model	Params	SugarCrepe	BiVLC I2T	BiVLC T2I	BiVLC Group
	Human	N/A	98.93	90.40	93.00	86.80
	Random	N/A	50.00	25.00	25.00	16.67
Contrastive	CLIP	151M	76.56	75.83	52.40	49.06
	$\text{CLIP}_{\text{COCO}}$		84.66	**82.75**	**63.89**	**60.96**
	NegCLIP		**85.64**	80.74	61.95	58.75
	GNM		81.83	81.32	60.86	57.96
Generative	Open CapPa	676M	90.59	57.72	56.19	41.97
	VQAScore-XL	3B	90.85	81.96	76.61	70.20
	VQAScore-XXL	11B	**93.72**	**86.16**	**81.93**	**76.47**

Finding 1: Current models underperform on text-to-image retrieval.
Finding 2: The gap to humans is bigger in BiVLC than in SugarCrepe.
Finding 3: SugarCrepe and BiVLC performance are not correlated.

Exploring training strategies

We propose two new models:

1. **$\text{CLIP}_{\text{TROHN-TEXT}}$** using hard negative **texts.**
2. **$\text{CLIP}_{\text{TROHN-IMG}}$** using hard negative **texts and images.**

Model	SugarCrepe	BiVLC I2T	BiVLC T2I	BiVLC Group
Random	50.00	25.00	25.00	16.67
CLIP	76.56	75.83	52.40	49.06
$\text{CLIP}_{\text{COCO}}$	84.66	<u>82.75</u>	<u>63.89</u>	<u>60.96</u>
NegCLIP	85.64	80.74	61.95	58.75
GNM	81.83	81.32	60.86	57.96
$\text{CLIP}_{\text{TROHN-Text}}$	**93.40**	78.18	62.19	57.48
$\text{CLIP}_{\text{TROHN-Img}}$	<u>89.40</u>	**88.54**	**71.84**	**69.25**

Finding 4: Training with hard negative images can boost the performance of multimodal contrastive models.

Are our models cheating?

We develop two new systems which are trained to detect synthetic and natural images and captions: **CLIP_{Det}**, based on original pretrained CLIP encoders and **$\text{CLIP}_{\text{TROHN-IMG/Det}}$**, our $\text{CLIP}_{\text{TROHN-IMG}}$ model encoders.

Model	Text detection acc	Img detection acc	I2T	T2I	Group
Random	50.00	50.00	25.00	25.00	16.67
CLIP_{Det}	57.00	100.00	66.69	19.64	19.64
$\text{CLIP}_{\text{TROHN-Img/Det}}$	61.34	100.00	75.04	26.42	26.42

Finding 5: Distinguishing between natural and synthetic inputs is not enough to perform well in BiVLC.
Finding 6: I2T is more sensitive to natural vs synthetic.

Highlights

- The largest I2T and T2I compositionality dataset.
- A new semi-automatic dataset construction method.
- BiVLC offers a more complete view of compositionality skills.
- Multimodal models lag behind humans by a large margin.

Contact

- **by email** {imanol.miranda, ander.salaberria, e.agirre, gorka.azcune}@ehu.eus
- **X** @I_MirandaM @AnderSala @eagirre @gazkune

Project page
https://imirandam.github.io/BiVLC_project_page