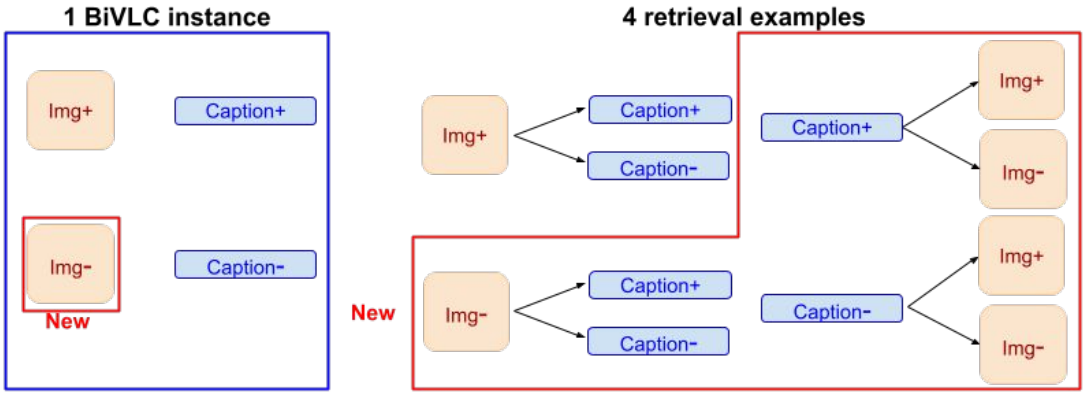
BiVLC: Extending Vision-Language Compositionality Evaluation with Text-to-Image Retrieval

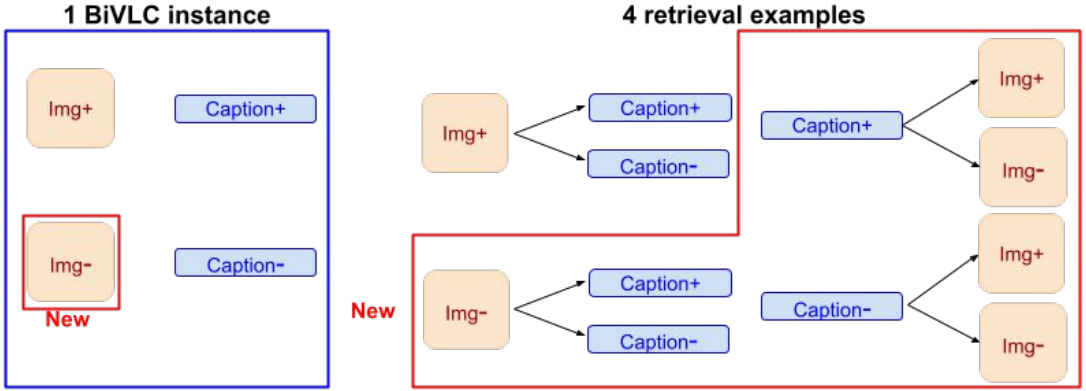
Imanol Miranda, Ander Salaberria, Eneko Agirre, Gorka Azkune
HiTZ Center – Ixa, University of the Basque Country (UPV/EHU)
{imanol.miranda, ander.salaberria, e.agirre, gorka.azcune}@ehu.eus

Universidad del País Vasco Euskal Herriko Unibertsitatea
HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology
NEURAL INFORMATION PROCESSING SYSTEMS

Adding negative images opens the door to many more retrieval examples

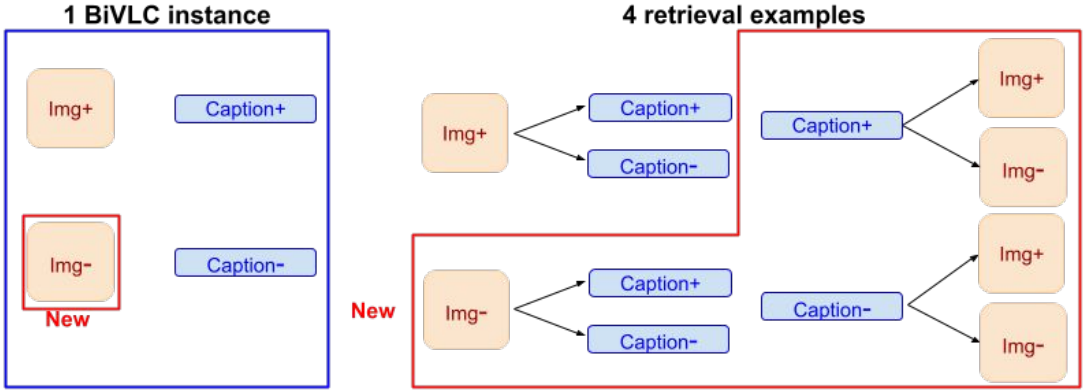


Adding negative images opens the door to many more retrieval examples



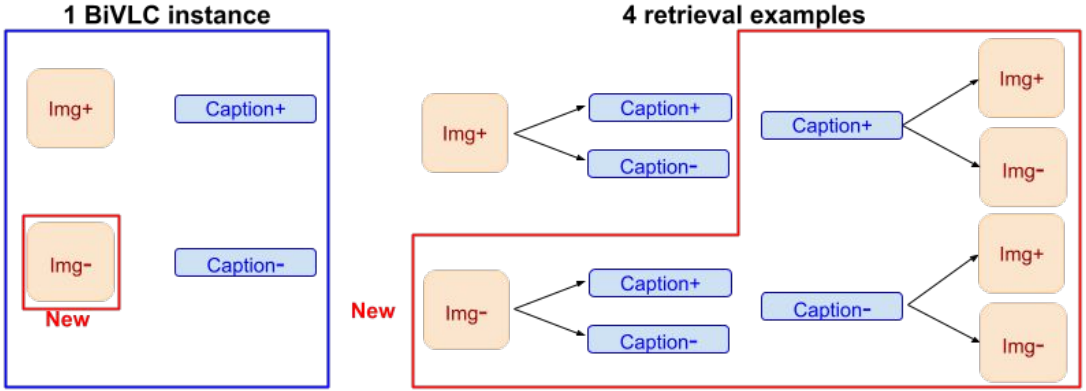
Dataset	I2T	T2I	REPLACE			SWAP			ADD		Total
			OBJ	ATT	REL	OBJ	ATT	REL	OBJ	ATT	
Winoground	✓	✓				668		1,036			1,600†
SUGARCREPE	✓		1,652	788	1,406	246	666		2,062	692	7,512
BIVLC (ours)	✓	✓	4,800	1,748	1,848	324	1,112		1,596	304	11,732

Adding negative images opens the door to many more retrieval examples



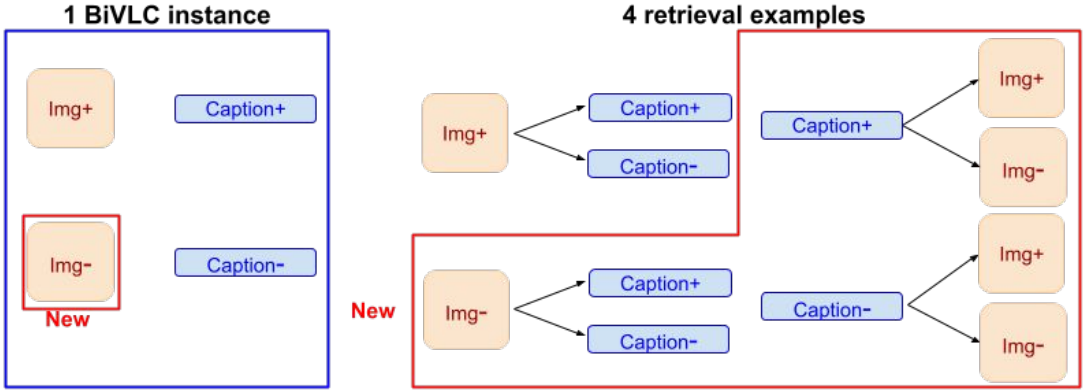
Dataset	I2T	T2I	REPLACE			SWAP			ADD		Total
			OBJ	ATT	REL	OBJ	ATT	REL	OBJ	ATT	
Winoground	✓	✓				668		1,036			1,600†
SUGARCREPE	✓		1,652	788	1,406	246	666		2,062	692	7,512
BIVLC (ours)	✓	✓	4,800	1,748	1,848	324	1,112		1,596	304	11,732

Adding negative images opens the door to many more retrieval examples



Dataset	I2T	T2I	REPLACE			SWAP			ADD		Total
			OBJ	ATT	REL	OBJ	ATT	REL	OBJ	ATT	
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

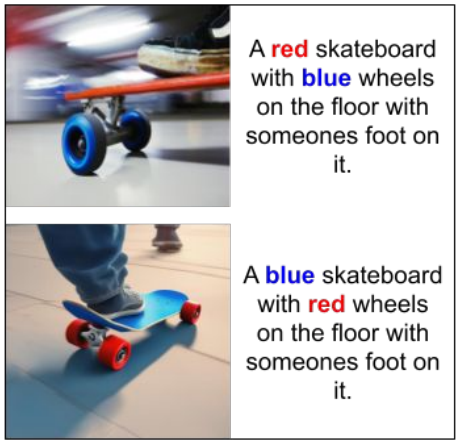
Adding negative images opens the door to many more retrieval examples



Dataset	I2T	T2I	REPLACE			SWAP			ADD		Total
			OBJ	ATT	REL	OBJ	ATT	REL	OBJ	ATT	
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

BiVLC: Bidirectional Vision-Language Compositionality dataset

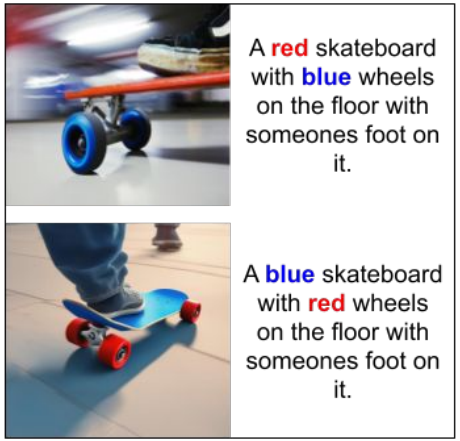
BiVLC is a **Bi**directional **V**ision-**L**anguage **C**ompositionality dataset with almost 3k instances formed by 2 images and 2 captions.



BiVLC: Bidirectional Vision-Language Compositionality dataset

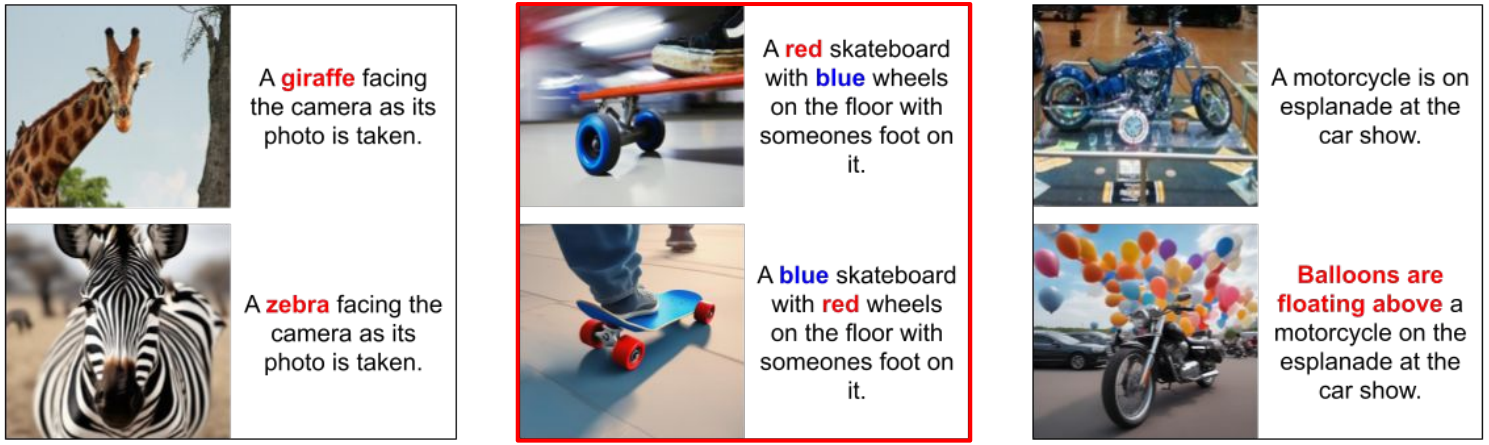
BiVLC is a **Bi**directional **V**ision-**L**anguage **C**ompositionality dataset with almost 3k instances formed by 2 images and 2 captions.

Replace



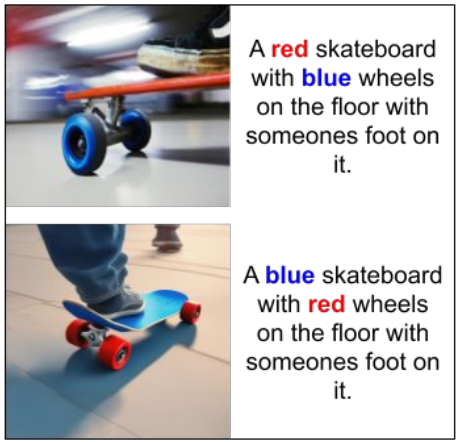
BiVLC: Bidirectional Vision-Language Compositionality dataset

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



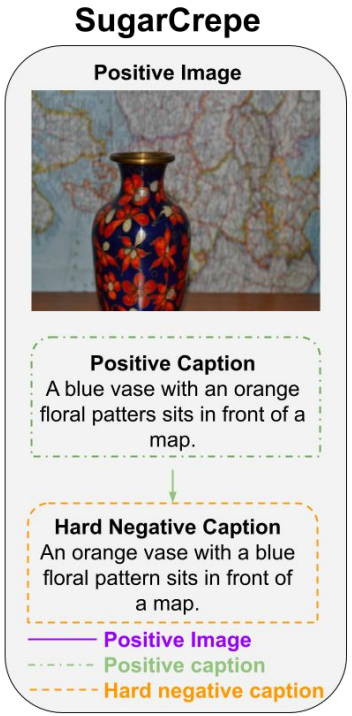
BiVLC: Bidirectional Vision-Language Compositionality dataset

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

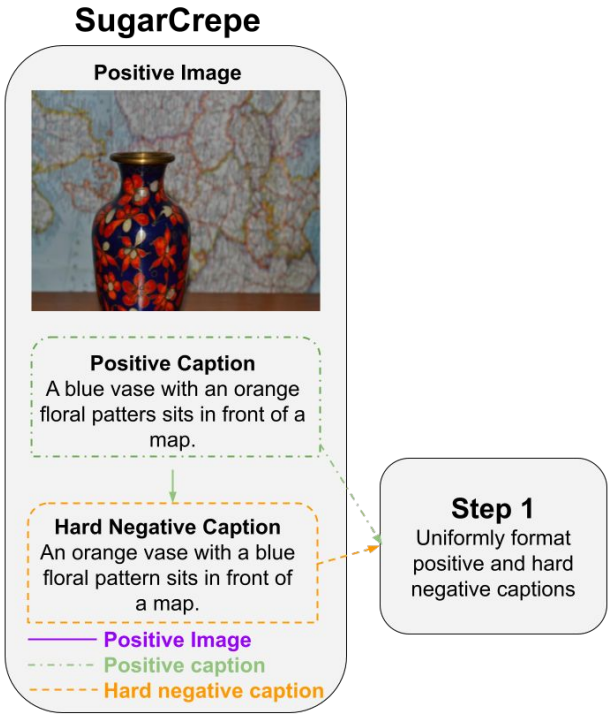


ADD

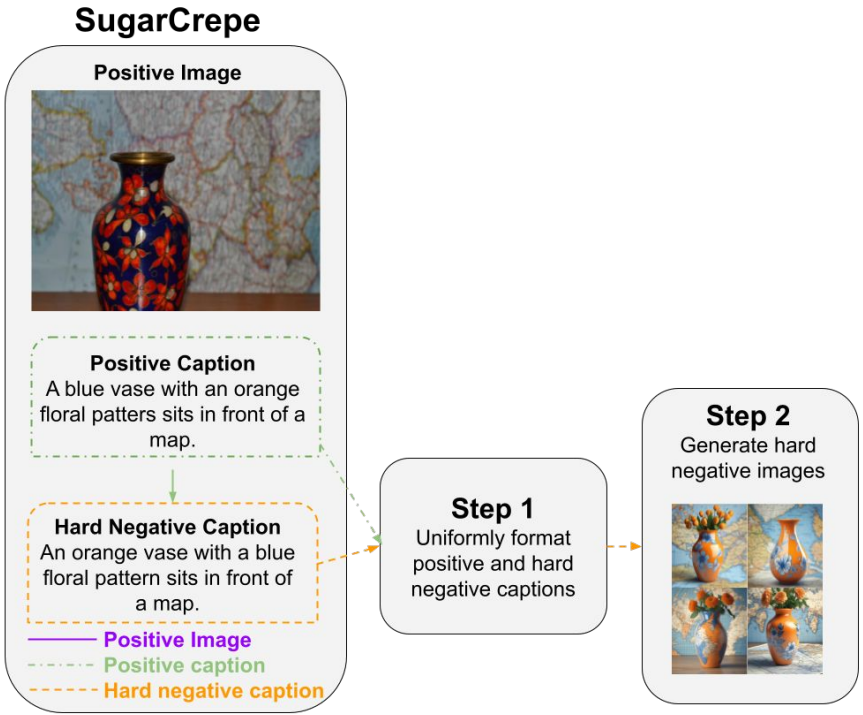
Semi-automatic dataset construction



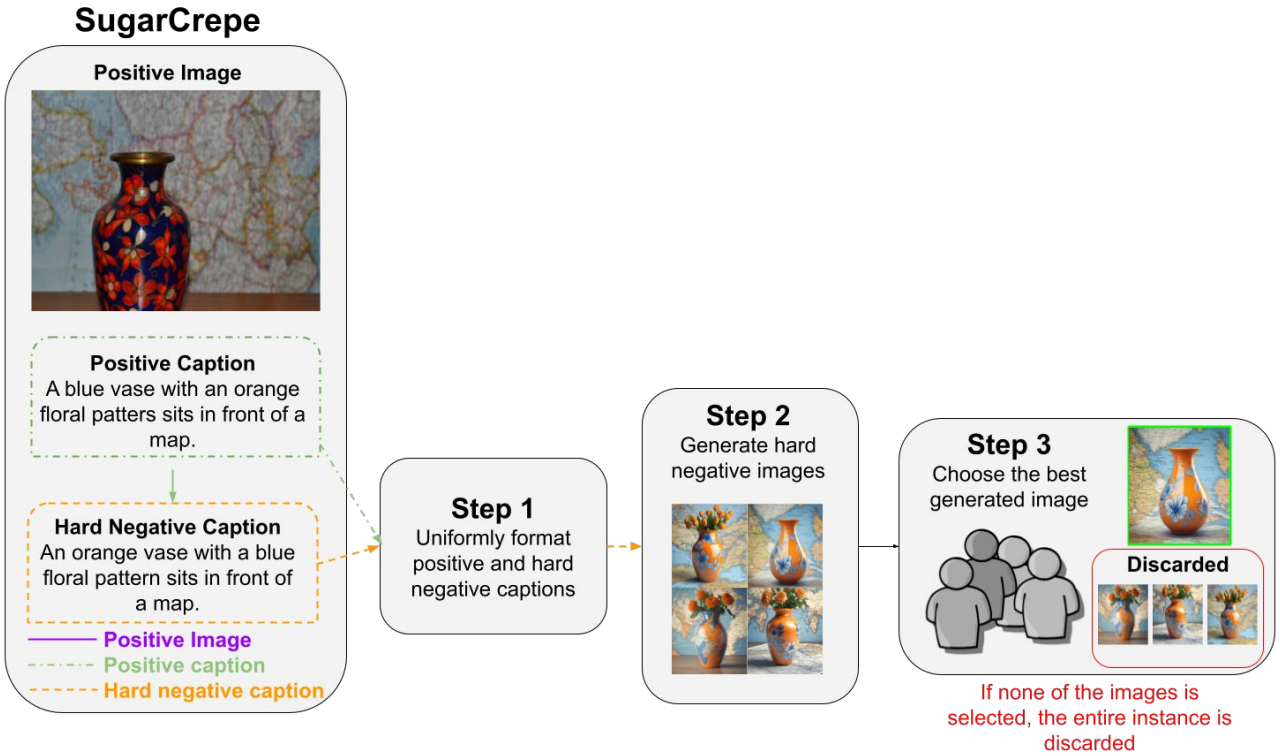
Semi-automatic dataset construction



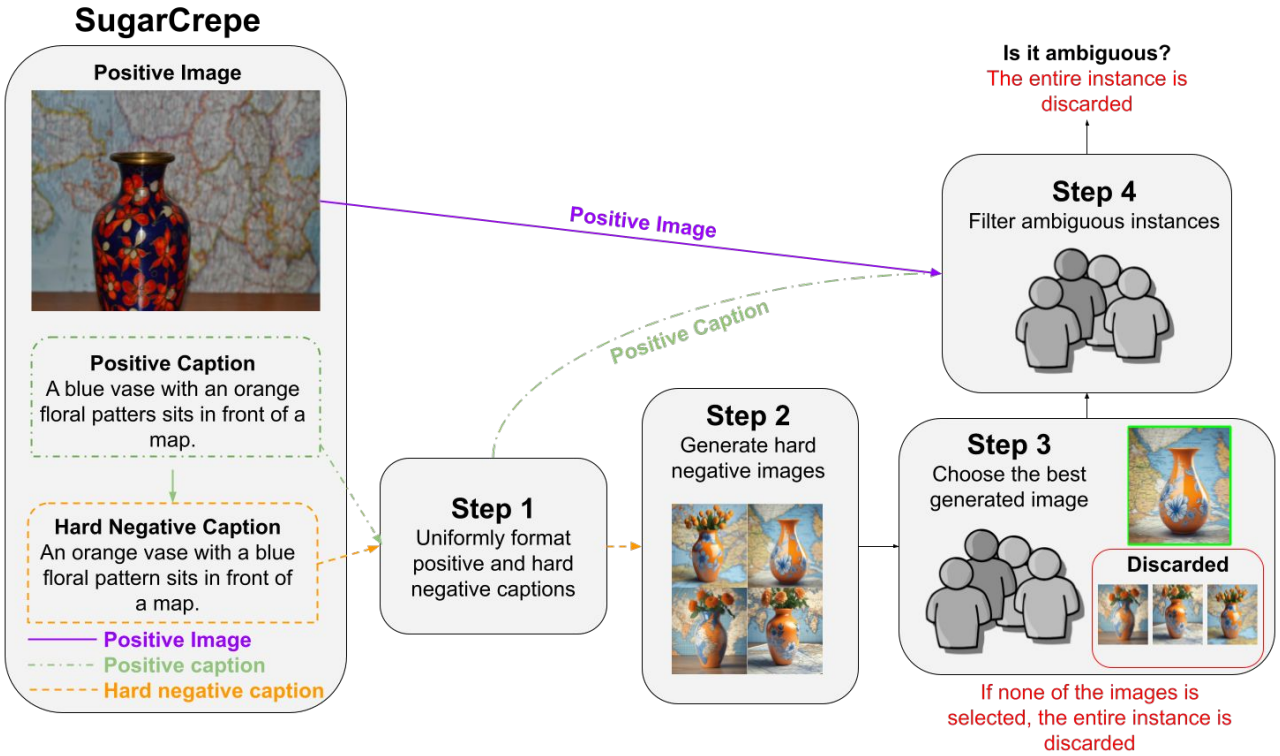
Semi-automatic dataset construction



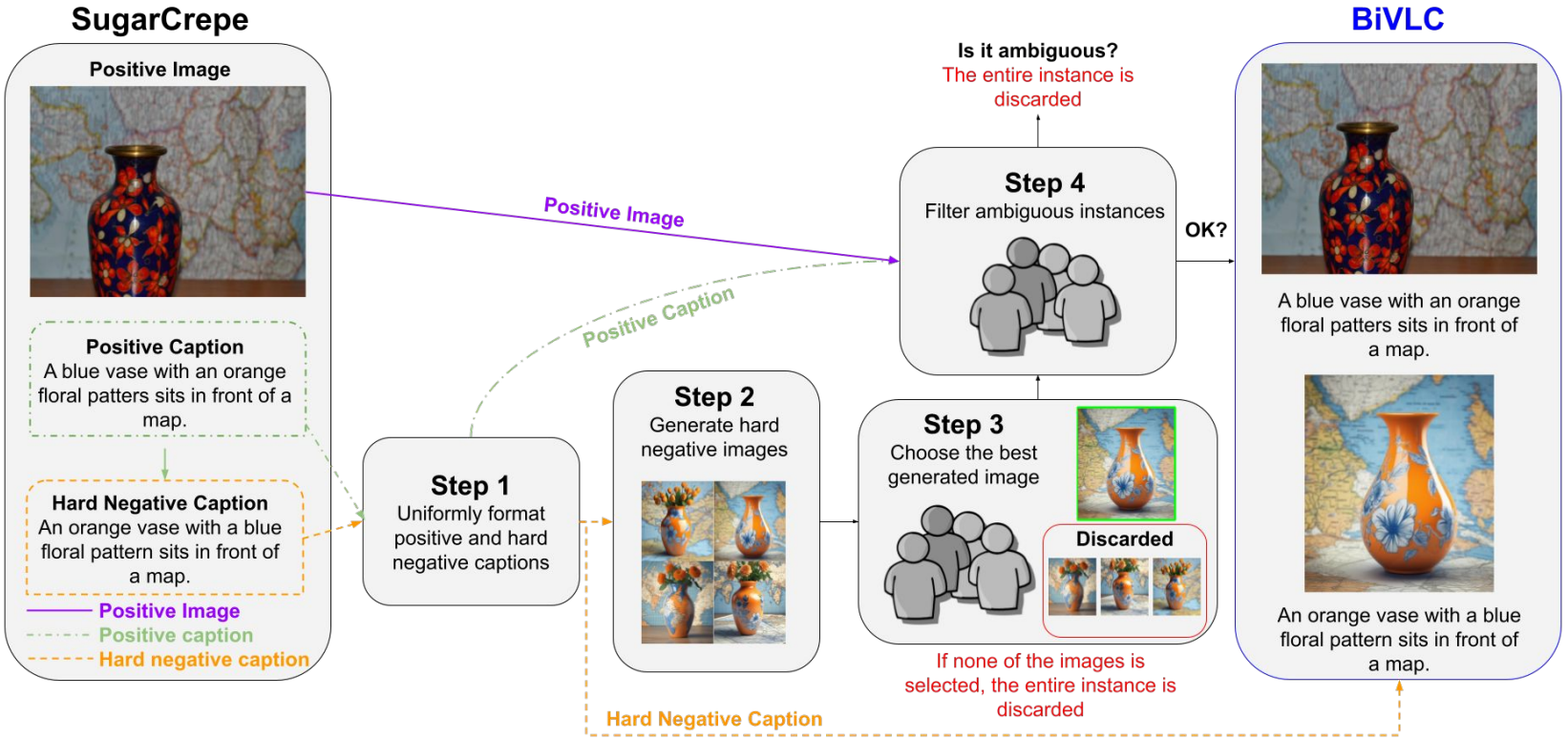
Semi-automatic dataset construction



Semi-automatic dataset construction



Semi-automatic dataset construction



SOTA models results

	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	90.46	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**

SOTA models results

	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	90.46	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**

SOTA models results

	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	90.46	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**

SOTA models results

	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	90.46	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**

SOTA models results

	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	90.46	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**

SOTA models results

	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	90.46	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**

SOTA models results

	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	90.46	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**

SOTA models results

	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	90.46	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**

Findings

	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	90.46	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.

Findings

	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	90.46	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 2: The gap to humans is bigger in BiVLC than in SugarCrepe

Findings

	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	90.46	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 3: SugarCrepe and BiVLC performance are not correlated

Exploring training strategies

We **propose two new models based on the two main strategies in the literature** to improve the VLC skills of a multimodal model:

1. **CLIP$_{\text{TROHN-TEXT}}$** using hard negative texts for training**.**

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

Exploring training strategies

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**

Exploring training strategies

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**

Exploring training strategies

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**

VQAScore-XXL 93.72

Exploring training strategies

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**

Exploring training strategies

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**

SugarCrepe and BiVLC performance are not correlated

Exploring training strategies

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	82.75	63.89	60.96
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
$\text{CLIP}_{\text{TROHN-Img}}$	89.40	**88.54**	**71.84**	**69.25**

Exploring training strategies

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**

TROHN-Text has 10 times more hard negative texts

Exploring training strategies

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**

Exploring training strategies

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**

VQAScore-XL Group score 70.20

3B vs 151M parameters for CLIP

Further analysis

- Why does training with hard negative images help?
- Which category is the most difficult?
- Why is $\text{CLIP}_{\text{TROHN-IMG}}$ still far from humans?
- Are our models just distinguishing between synthetic and natural?

Thank you!

- Project page: https://imirandam.github.io/BiVLC_project_page
- Github: https://github.com/IMirandaM/BiVLC
- Dataset: https://huggingface.co/datasets/imirandam/BiVLC
- Contact
  - by email {imanol.miranda, ander.salaberria, e.agirre, gorka.azcune}@ehu.eus
  - X @I_MirandaM @AnderSala @eagirre @gazkune