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

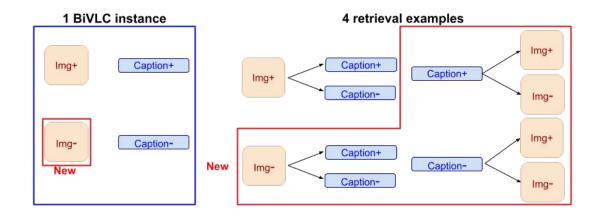
Imanol Miranda, Ander Salaberria, Eneko Agirre, Gorka Azkune HiTZ Center – Ixa, University of the Basque Country (UPV/EHU) {imanol.miranda, ander.salaberria, e.agirre, gorka.azcune}@ehu.eus

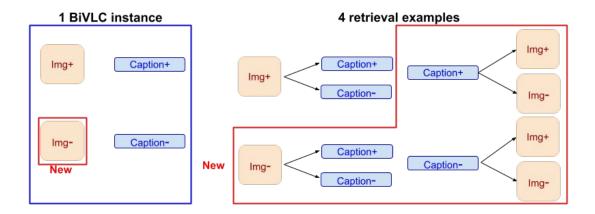


Universidad Euskal Herriko del País Vasco Unibertsitatea HiTZ

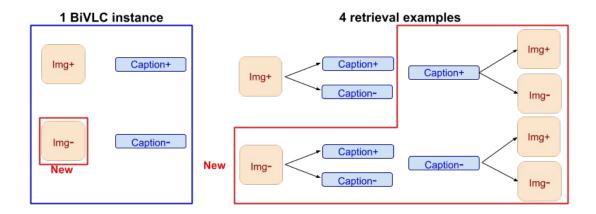
Hizkuntza Teknologiako Zentroa Basque Center for Language Technology



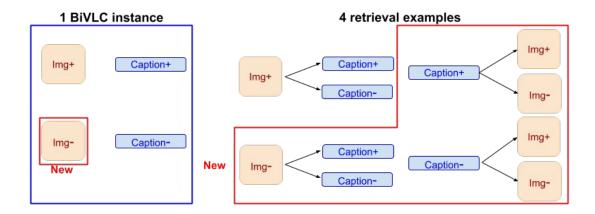




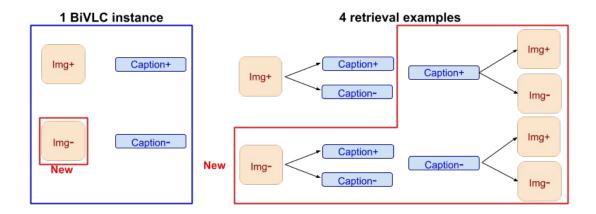
Detect	ют	I2T T2I		REPLACE		SWAP			ADD		Total
Dataset	141	1 41	OBJ	ATT	REL	Obj	ATT	REL	OBJ	ATT	Total
Winoground	\checkmark	\checkmark				668		1,036			1,600†
SUGARCREPE	\checkmark		1,652	788	1,406	246	666		2,062	692	7,512
BIVLC (ours)	\checkmark	\checkmark	4,800	1,748	1,848	324	1,112		1,596	304	11,732



Dataset	I2T	T2I	F	REPLAC	E		SWAP		AD	D	Total
Dataset	141	1 41	OBJ	ATT	REL	Obj	ATT	REL	OBJ	ATT	Total
Winoground	\checkmark	\checkmark				668		1,036			1,600†
SUGARCREPE	~		1,652	788	1,406	246	666		2,062	692	7,512
BIVLC (ours)	\checkmark	\checkmark	4,800	1,748	1,848	324	1,112		1,596	304	11,732



Dataset	I2T T2I		REPLACE		SWAP			ADD		Total	
Dataset	141	1 41	OBJ	ATT	REL	Obj	ATT	REL	OBJ	ATT	Total
Winoground	\checkmark	\checkmark				668		1,036			1,600†
SUGARCREPE	\checkmark		1,652	788	1,406	246	666		2,062	692	7,512
BIVLC (ours)	\checkmark	\checkmark	4,800	1,748	1,848	324	1,112		1,596	304	11,732



Dataset	I2T	T2I	F	REPLAC	E		SWAP		AD	D	Total
Dataset	141	1 41	OBJ	ATT	REL	Obj	ATT	REL	OBJ	ATT	Total
Winoground	\checkmark	\checkmark				668		1,036			1,600†
SUGARCREPE	\checkmark		1,652	788	1,406	246	666	a di kana	2,062	692	7,512
BIVLC (ours)	\checkmark	\checkmark	4,800	1,748	1,848	324	1,112		1,596	304	11,732

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



A **giraffe** facing the camera as its photo is taken.



A zebra facing the

camera as its photo is taken.



A red skateboard with blue wheels on the floor with someones foot on it.



A **blue** skateboard with **red** wheels on the floor with someones foot on it.



A motorcycle is on esplanade at the car show.



Balloons are floating above a motorcycle on the esplanade at the car show.

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



A **giraffe** facing the camera as its photo is taken.



Replace

A **zebra** facing the camera as its photo is taken.



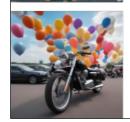
A red skateboard with blue wheels on the floor with someones foot on it.



A **blue** skateboard with **red** wheels on the floor with someones foot on it.



A motorcycle is on esplanade at the car show.



Balloons are floating above a motorcycle on the esplanade at the car show.

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



A **giraffe** facing the camera as its photo is taken.



A **zebra** facing the

camera as its photo is taken.



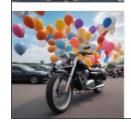
A red skateboard with blue wheels on the floor with someones foot on it.



A **blue** skateboard with **red** wheels on the floor with someones foot on it.



A motorcycle is on esplanade at the car show.



Balloons are floating above a motorcycle on the esplanade at the car show.

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



A **giraffe** facing the camera as its photo is taken.



A **zebra** facing the camera as its photo is taken.



A red skateboard with blue wheels on the floor with someones foot on it.



A **blue** skateboard with **red** wheels on the floor with someones foot on it.



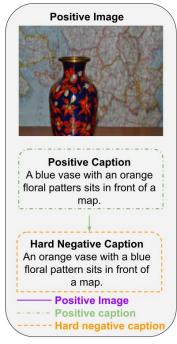
A motorcycle is on esplanade at the car show.



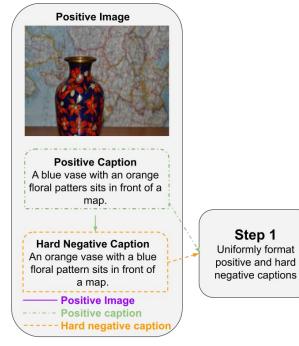
Balloons are floating above a motorcycle on the esplanade at the car show.

ADD

SugarCrepe



SugarCrepe



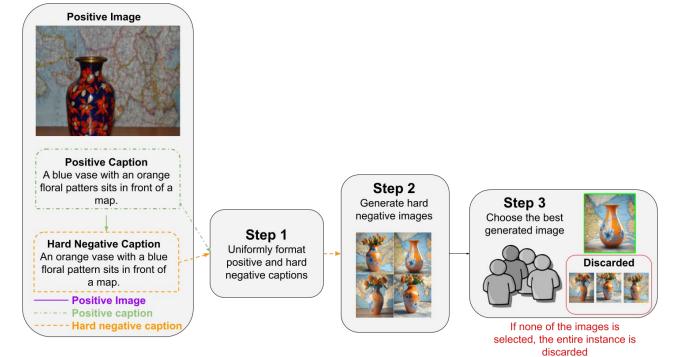
SugarCrepe

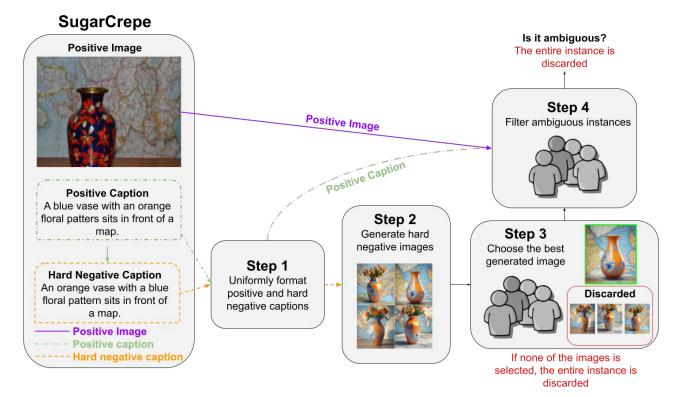


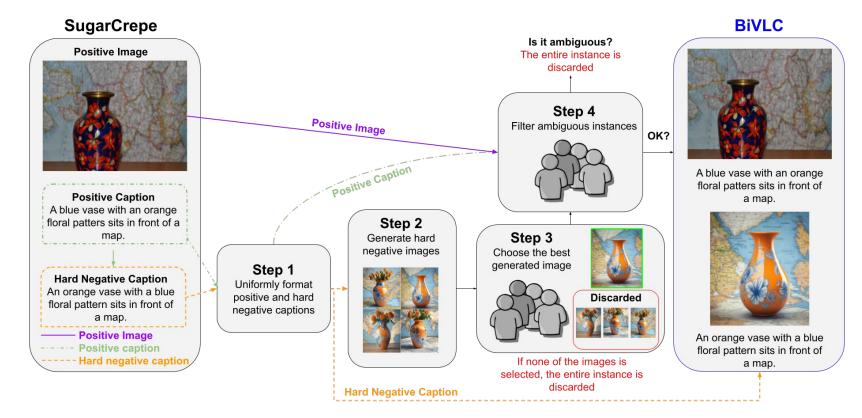
Step 1 Uniformly format positive and hard negative captions Step 2 Generate hard negative images



SugarCrepe







	Madal	D	Gua - D Coppo		BIVLO	2
	Model	Params	SUGARCREPE	I2T	T2I	Group
	Human	N/A	98.93	90.40	93.00	86.80
	Random	N/A	50.00	25.00	25.00	16.67
	CLIP		76.56	75.83	52.40	49.06
Contractive	CLIP _{COCO}	15134	84.66	82.75	63.89	60.96
Contrastive	NEGCLIP	151M	85.64	80.74	61.95	58.75
	GNM		81.83	81.32	60.86	57.96
	Open CapPa	676M	90.46	57.72	56.19	41.97
Generative	VQASCore-XL	3B	90.85	81.96	76.61	70.20
	VQASCore-XXL	11B	93.72	86.16	81.93	76.47

	Nr. 11	D	Guai n Canan	BIVLC			
	Model	Params	SUGARCREPE	I2T	T2I	Group	
	Human	N/A	98.93	90.40	93.00	86.80	
	Random	N/A	50.00	25.00	25.00	16.67	
	CLIP		76.56	75.83	52.40	49.06	
Contractivo	CLIP _{COCO}	15114	84.66	82.75	63.89	60.96	
Contrastive	NEGCLIP	151M	85.64	80.74	61.95	58.75	
	GNM		81.83	81.32	60.86	57.96	
	Open CapPa	676M	90.46	57.72	56.19	41.97	
Generative	VQASCore-XL	3B	90.85	81.96	76.61	70.20	
	VQASCore-XXL	11B	93.72	86.16	81.93	76.47	

	M. 11	D	Guai n Canan		BIVLO	2
	Model	Params	SUGARCREPE	I2T	T2I	Group
	Human	N/A	98.93	90.40	93.00	86.80
	Random	N/A	50.00	25.00	25.00	16.67
	CLIP		76.56	75.83	52.40	49.06
Contraction	CLIP _{COCO}	15114	84.66	82.75	63.89	60.96
Contrastive	NEGCLIP	151M	85.64	80.74	61.95	58.75
	GNM		81.83	81.32	60.86	57.96
	Open CapPa	676M	90.46	57.72	56.19	41.97
Generative	VQASCore-XL	3B	90.85	81.96	76.61	70.20
	VQASCore-XXL	11B	93.72	86.16	81.93	76.47

	Madal	D	Cura D C D D D D		BIVLO	2
	Model	Params	SUGARCREPE	I2T	T2I	Group
	Human	N/A	98.93	90.40	93.00	86.80
	Random	N/A	50.00	25.00	25.00	16.67
	CLIP		76.56	75.83	52.40	49.06
Contrastive	CLIP _{COCO}	151M	84.66	82.75	63.89	60.96
Contrastive	NEGCLIP	13111	85.64	80.74	61.95	58.75
	GNM		81.83	81.32	60.86	57.96
	Open CapPa	676M	90.46	57.72	56.19	41.97
Generative	VQASCore-XL	3B	90.85	81.96	76.61	70.20
	VQASCore-XXL	11B	93.72	86.16	81.93	76.47

	Madal	D	Cuch D CD DDD		BIVLO	2
	Model	Params	SUGARCREPE	I2T	T2I	Group
	Human	N/A	98.93	90.40	93.00	86.80
	Random	N/A	50.00	25.00	25.00	16.67
	CLIP		76.56	75.83	52.40	49.06
Contractivo	CLIP _{COCO}	15114	84.66	82.75	63.89	60.96
Contrastive	NEGCLIP	151M	85.64	80.74	61.95	58.75
	GNM		81.83	81.32	60.86	57.96
	Open CapPa	676M	90.46	57.72	56.19	41.97
Generative	VQASCore-XL	3B	90.85	81.96	76.61	70.20
	VQASCore-XXL	11B	93.72	86.16	81.93	76.47

	M 11	D	Gua in Conno		BIVLO	2
	Model	Params	SUGARCREPE	I2T	T2I	Group
	Human	N/A	98.93	90.40	93.00	86.80
	Random	N/A	50.00	25.00	25.00	16.67
	CLIP		76.56	75.83	52.40	49.06
Contraction	CLIP _{COCO}	151M	84.66	82.75	63.89	60.96
Contrastive	NEGCLIP		85.64	80.74	61.95	58.75
	GNM		81.83	81.32	60.86	57.96
	Open CapPa	676M	90.46	57.72	56.19	41.97
Generative	VQASCore-XL	3B	90.85	81.96	76.61	70.20
	VQASCore-XXL	11B	93.72	86.16	81.93	76.47

	Nr. 11	D	Guai n Canan	BIVLC			
	Model	Params	SUGARCREPE	I2T	T2I	Group	
	Human	N/A	98.93	90.40	93.00	86.80	
	Random	N/A	50.00	25.00	25.00	16.67	
	CLIP		76.56	75.83	52.40	49.06	
Contractive	CLIP _{COCO}	15114	84.66	82.75	63.89	60.96	
Contrastive	NEGCLIP	151M	85.64	80.74	61.95	58.75	
	GNM		81.83	81.32	60.86	57.96	
	Open CapPa	676M	90.46	57.72	56.19	41.97	
Generative	VQASCore-XL	3B	90.85	81.96	76.61	70.20	
	VQASCore-XXL	11B	93.72	86.16	81.93	76.47	

	M 11	D	Guai a Casaa	BIVLC			
	Model	Params	SUGARCREPE	I2T	T2I	Group	
	Human	N/A	98.93	90.40	93.00	86.80	
	Random	N/A	50.00	25.00	25.00	16.67	
	CLIP		76.56	75.83	52.40	49.06	
Contraction	CLIP _{COCO}	151M	84.66	82.75	63.89	60.96	
Contrastive	NEGCLIP		85.64	80.74	61.95	58.75	
	GNM		81.83	81.32	60.86	57.96	
	Open CapPa	676M	90.46	57.72	56.19	41.97	
Generative	VQASCore-XL	3B	90.85	81.96	76.61	70.20	
	VQASCore-XXL	11B	93.72	86.16	81.93	76.47	

Findings

	Madal	el Params SUGARCREPE		BIVLC		
	Model			I2T	T2I	Group
	Human	N/A	98.93	90.40	93.00	86.80
	Random	N/A	50.00	25.00	25.00	16.67
	CLIP		76.56	75.83	52.40	49.06
Contractivo	CLIP _{COCO}	151M	84.66	82.75	63.89	60.96
Contrastive	NEGCLIP	151M	85.64	80.74	61.95	58.75
	GNM		81.83	81.32	60.86	57.96
	Open CapPa	676M	90.46	57.72	56.19	41.97
Generative	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

	Madal	D	Cuch D CD DDD		BIVLO	2
	Model	Params	SUGARCREPE	I2T	T2I	Group
	Human	N/A	98.93	90.40	93.00	86.80
	Random	N/A	50.00	25.00	25.00	16.67
	CLIP		76.56	75.83	52.40	49.06
Contractive	CLIP _{COCO}	15114	84.66	82.75	63.89	60.96
Contrastive	NEGCLIP	151M	85.64	80.74	61.95	58.75
	GNM		81.83	81.32	60.86	57.96
	Open CapPa	676M	90.46	57.72	56.19	41.97
Generative	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

	Nr. 11	D. Guain Coros		BIVLC		
	Model	Params	SUGARCREPE	I2T	T2I	Group
	Human	N/A	98.93	90.40	93.00	86.80
	Random	N/A	50.00	25.00	25.00	16.67
	CLIP		76.56	75.83	52.40	49.06
Contractive	CLIP _{COCO}	15114	84.66	82.75	63.89	60.96
Contrastive	NEGCLIP	151M	85.64	80.74	61.95	58.75
	GNM		81.83	81.32	60.86	57.96
	Open CapPa	676M	90.46	57.72	56.19	41.97
Generative	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

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**_{TROHN-TEXT} using hard negative texts for training.
- 2. **CLIP**_{TROHN-IMG} using both, hard negative texts and images.

Madal	SugarCorpe	BIVLC			
Model	SUGARCREPE	I2T	T2I	Group	
Random	50.00	25.00	25.00	16.67	
CLIP	76.56	75.83	52.40	49.06	
CLIP _{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	
CLIP _{TROHN-TEXT}	93.40	78.18	62.19	57.48	
CLIP _{TROHN-IMG}	89.40	88.54	71.84	69.25	

Model	SUGARCREPE	BIVLC			
wiouei	SUGARCREPE	I2T	T2I	Group	
Random	50.00	25.00	25.00	16.67	
CLIP	76.56	75.83	52.40	49.06	
CLIP _{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	
CLIP _{TROHN-TEXT}	93.40	78.18	62.19	57.48	
CLIP _{TROHN-IMG}	89.40	88.54	71.84	69.25	

Madal	Sug a D Control	BIVLC			
Model	SUGARCREPE	I2T	T2I	Group	
Random	50.00	25.00	25.00	16.67	
CLIP	76.56	75.83	52.40	49.06	
CLIP _{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	
CLIP _{TROHN-TEXT}	93.40	78.18	62.19	57.48	
CLIP _{TROHN-IMG}	89.40	88.54	71.84	69.25	

VQAScore-XXL 93.72

Madal	Sug + D CD DDD	BIVLC			
Model	SUGARCREPE	I2T	T2I	Group	
Random	50.00	25.00	25.00	16.67	
CLIP	76.56	75.83	52.40	49.06	
CLIP _{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	
CLIP _{TROHN-TEXT}	93.40	78.18	62.19	57.48	
CLIP _{TROHN-IMG}	89.40	88.54	71.84	69.25	

Model	SUGARCREPE	I2T	BIVLO T2I	Group
Random	50.00	25.00	25.00	16.67
CLIP	76.56	75.83	52.40	49.06
CLIP _{COCO}	84.66	<u>82.75</u>	63.89	<u>60.96</u>
NEGCLIP	85.64	80.74	61.95	58.75
GNM	81.83	81.32	60.86	57.96
CLIP _{TROHN-TEXT}	93.40	78.18	62.19	57.48
CLIP _{TROHN-IMG}	89.40	88.54	71.84	69.25

SugarCrepe and BiVLC performance are not correlated

Madal	SugarCorpe	BIVLC			
Model	SUGARCREPE	I2T	T2I	Group	
Random	50.00	25.00	25.00	16.67	
CLIP	76.56	75.83	52.40	49.06	
CLIP _{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	
CLIP _{TROHN-TEXT}	93.40	78.18	62.19	57.48	
CLIP _{TROHN-IMG}	89.40	88.54	71.84	69.25	

Model	SUGARCREPE	I2T	BIVLO T2I	Group
Random	50.00	25.00	25.00	16.67
CLIP	76.56	75.83	52.40	49.06
CLIP _{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
CLIP _{TROHN-TEXT}	93.40	78.18	62.19	57.48
CLIP _{TROHN-IMG}	89.40	88.54	71.84	69.25

TROHN-Text has 10 times more hard

negative texts

Madal	SugarCorpe	BIVLC			
Model	SUGARCREPE	I2T	T2I	Group	
Random	50.00	25.00	25.00	16.67	
CLIP	76.56	75.83	52.40	49.06	
CLIP _{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	
CLIP _{TROHN-TEXT}	93.40	78.18	62.19	57.48	
CLIP _{TROHN-IMG}	89.40	88.54	71.84	69.25	

Madal	Cuc + D CD DDD	BIVLC		
Model	SUGARCREPE	I2T	T2I	Group
Random	50.00	25.00	25.00	16.67
CLIP	76.56	75.83	52.40	49.06
CLIP _{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
CLIP _{TROHN-TEXT}	93.40	78.18	62.19	57.48
CLIP _{TROHN-IMG}	89.40	88.54	71.84	69.25
	V	QAScore-XL Group score 70		
	3	B vs 151	A parame	ters for C

Further analysis

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

Thank you!

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