Going Beyond Nouns With Vision & Language Models Using Synthetic Data

[1] RICE [3] MIT [4] GT

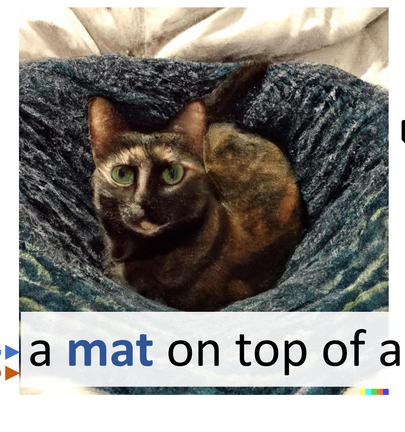
Paola Cascante-Bonilla[1,2], Khaled Shehada[2,3], James Seale Smith[2,4], Sivan Doveh[6,7], Donghyun Kim[2,7], Rameswar Panda[2,7], Gül Varol[5], Aude Oliva[2,3], Vicente Ordonez[1], Rogerio Feris[2,7], Leonid Karlinsky[2,7]

[2] MIT-IBM Watson AI Lab

[5] École des Ponts ParisTech

[6] Weizmann Institute of Science

Large-scale vision & language models struggle with **compositional reasoning**

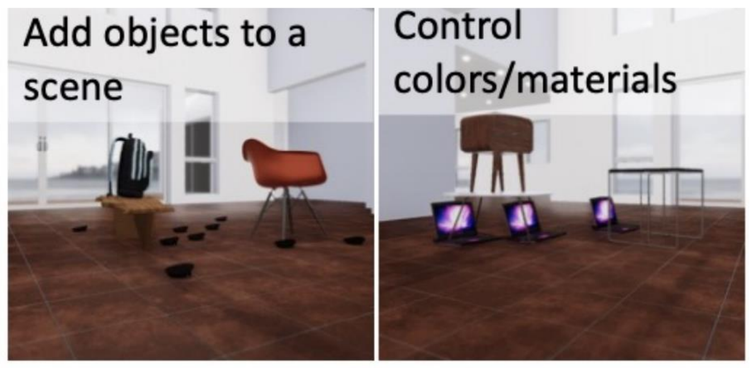


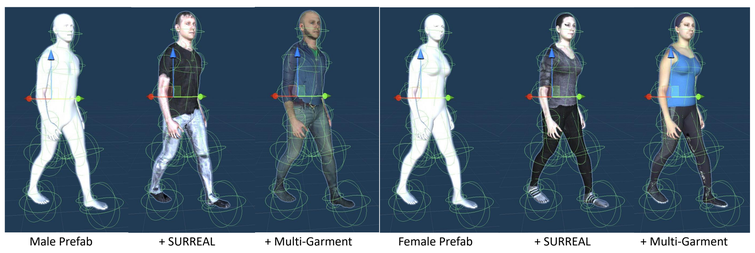
Synthetic data is cheap to produce and comes automatically labeled; it sidesteps many ethical and privacy issues that come with natural images.

ThreeDWorld (TDW) physics-based realistic simulation platform

Unity

- sample scene: 39 choices
- sample cameras (4-12)
- sample objects (0-8): 2304 choices
- sample attributes: 585 × 551 choices
- sample digital humans: 3 × 514 choices

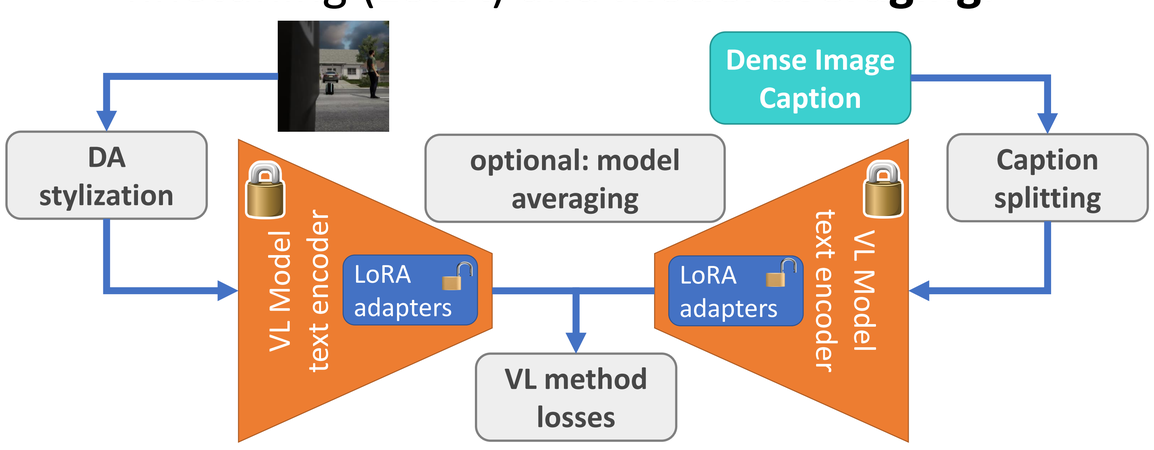




We propose **SyViC**, a million-scale synthetic dataset, and data generation codebase. By adding **digital humans** and enabling **physics,** we aim to cover transitive and intransitive **human actions**.

This scene contains a shirt, a stool, a fire hydrant, and two humans. They are in a room with blue floor and white walls. The first human is to the right of the fire hydrant. The second human wears a green football jersey and blue jeans pants (...)

We finetune a large-scale pre-trained VL model using our **SyViC** dataset. Our proposed methodology includes **domain adaptation** and **avoiding forgetting** through parameter-efficient finetuning (**LoRA**) and **model averaging**.



Results

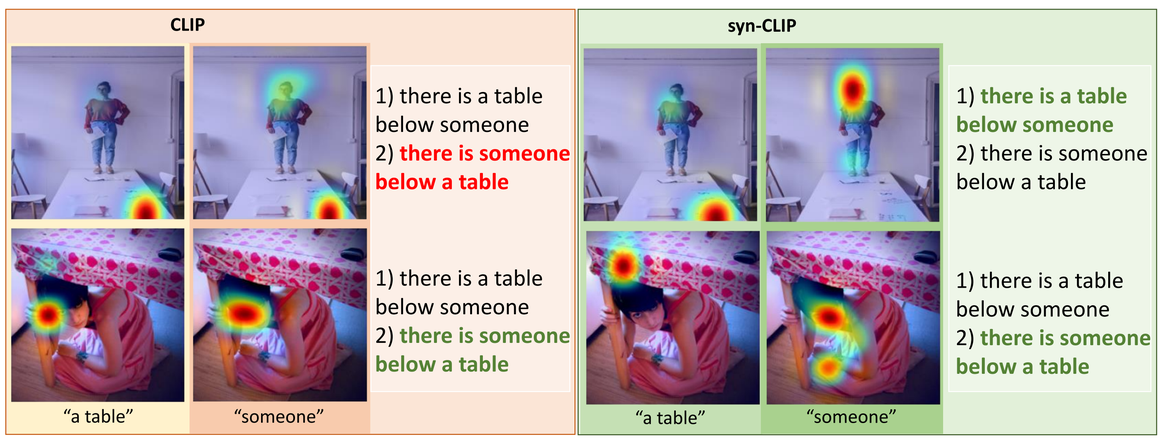
We evaluate our finetuned model on three benchmarks: *VL-Checklist*, *ARO*, and *Winoground*. The compositional reasoning evaluation includes understanding the meaning of the sentence after changing the **word order**, **attributes,** and **relations** of humans/objects.

	VL Checklist			Zero-Short
	Relation	Attribute	**Average**	(21 tasks)
CLIP	63.57	67.51	65.54	56.07
CyCLIP	61.15	66.96	64.06	55.99
syn-CLIP	69.39 (+5.82)	70.37 (+2.86)	69.88 (+4.34)	55.27 (-0.8)
syn-CyCLIP	65.73 (+4.58)	68.06 (+1.1)	66.89 (+2.83)	55.40 (-0.6)

	ARO				
	VG-Rel.	VG-Att.	Flickr30k	COCO	**Average**
CLIP	58.84	63.19	47.20	59.46	57.17
CyCLIP	59.12	65.41	20.82	29.54	43.72
syn-CLIP	71.40 (+12.56)	66.94 (+3.75)	59.06 (+11.86)	70.96 (+11.5)	67.09 (+9.9)
syn-CyCLIP	69.02 (+9.9)	63.65 (-1.76)	49.17 (+28.35)	59.36 (+29.82)	60.30 (+16.58)

	Winoground			Winoground†		
	Text	Image	**Group**	Text	Image	**Group**
CLIP	31.25	10.50	8.00	31.58	10.53	8.19
syn-CLIP	30.00	11.50	9.50 (+1.50)	29.82	12.28	9.94 (+1.75)

Grad-CAM on a Winoground sample before and after finetuning CLIP using our **SyViC** dataset.



Project Page

https://synthetic-vic.github.io/

[7] IBM Research