



Unlocking Compositional Reasoning ~~through~~ Vision-Language Models *in*

Paola Cascante-Bonilla

Postdoctoral Associate / University of Maryland, College Park

Assistant Professor / Stony Brook University (SUNY)

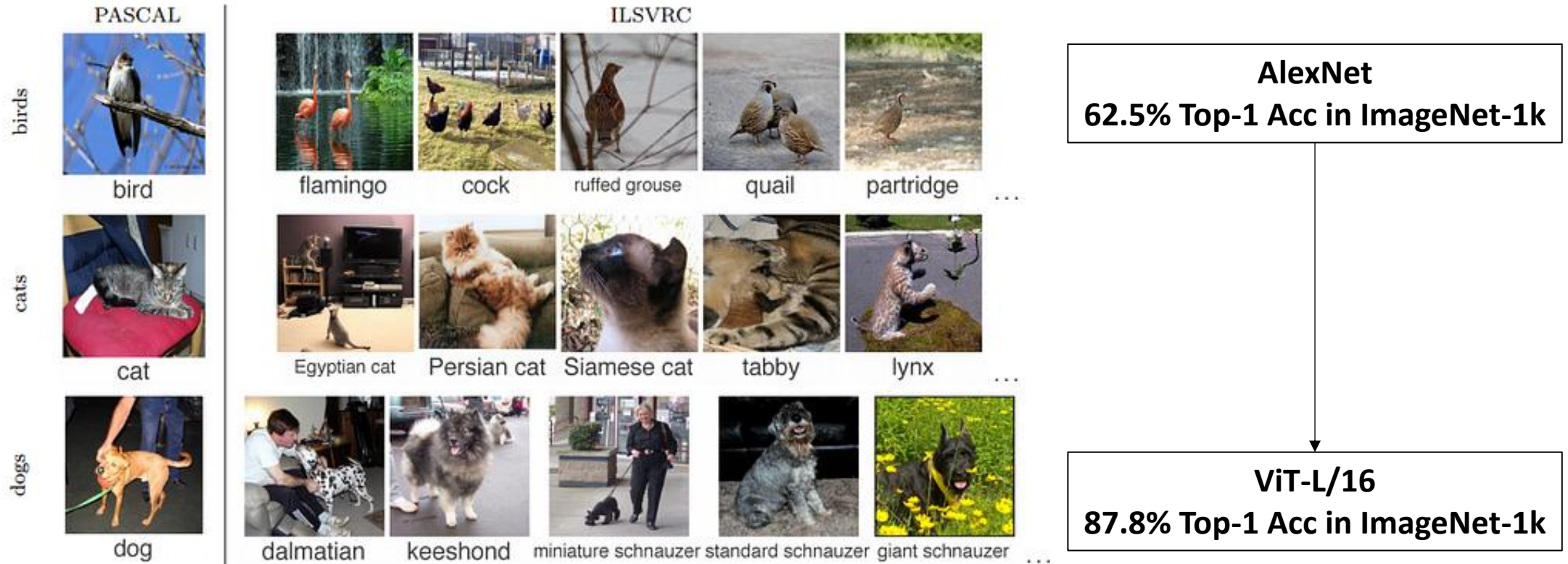


Classes: 1,000
Training Data: 1,381,167

ImageNet Large Scale Visual Recognition Challenge.

<http://image-net.org/about-overview>

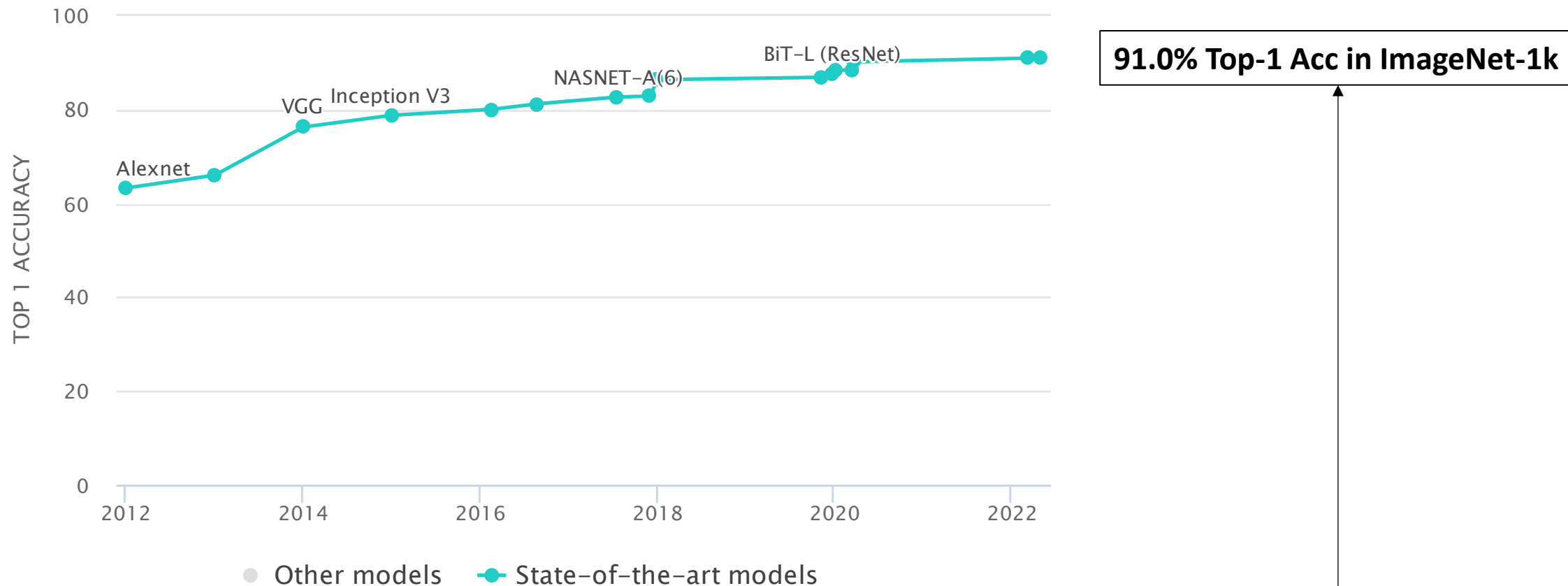
Vision Models are Good Classifiers



(NeurIPS 2012) ImageNet Classification with Deep Convolutional Neural Networks. Krizhevsky et al.

(ICLR 2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al. 3

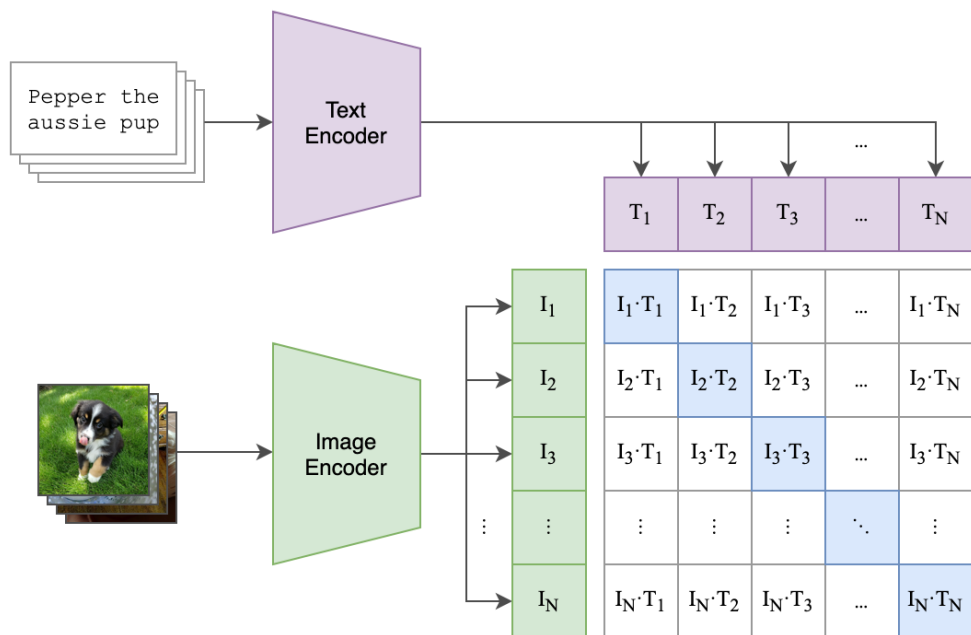
Vision-Language Models are Better Classifiers



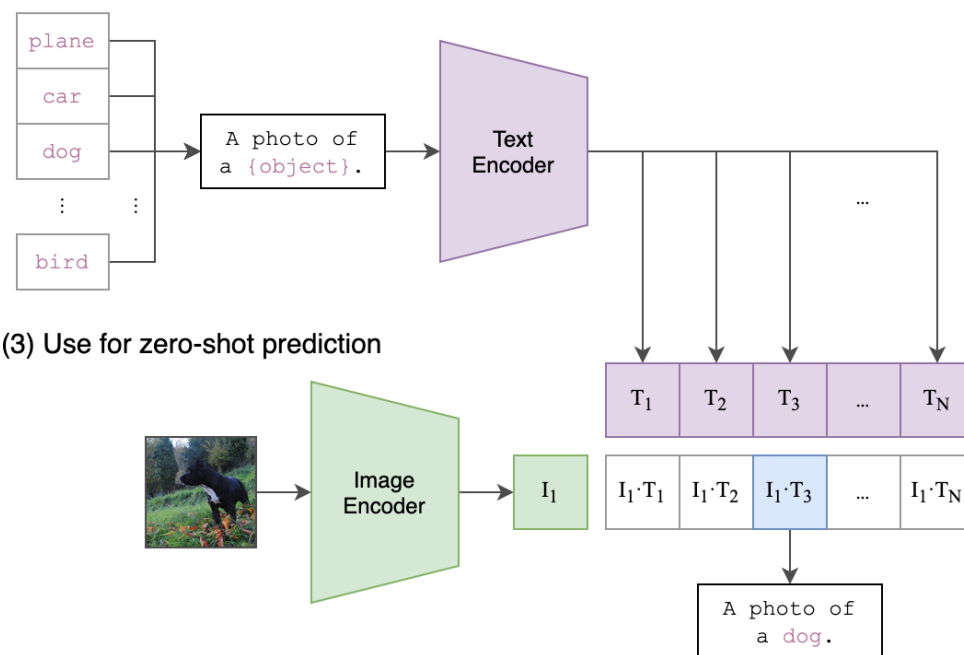
(TMLR 2022) CoCa: Contrastive Captioners are Image-Text Foundation Models. Yu, et al. ←

Vision-Language Models are Better Classifiers

(1) Contrastive pre-training



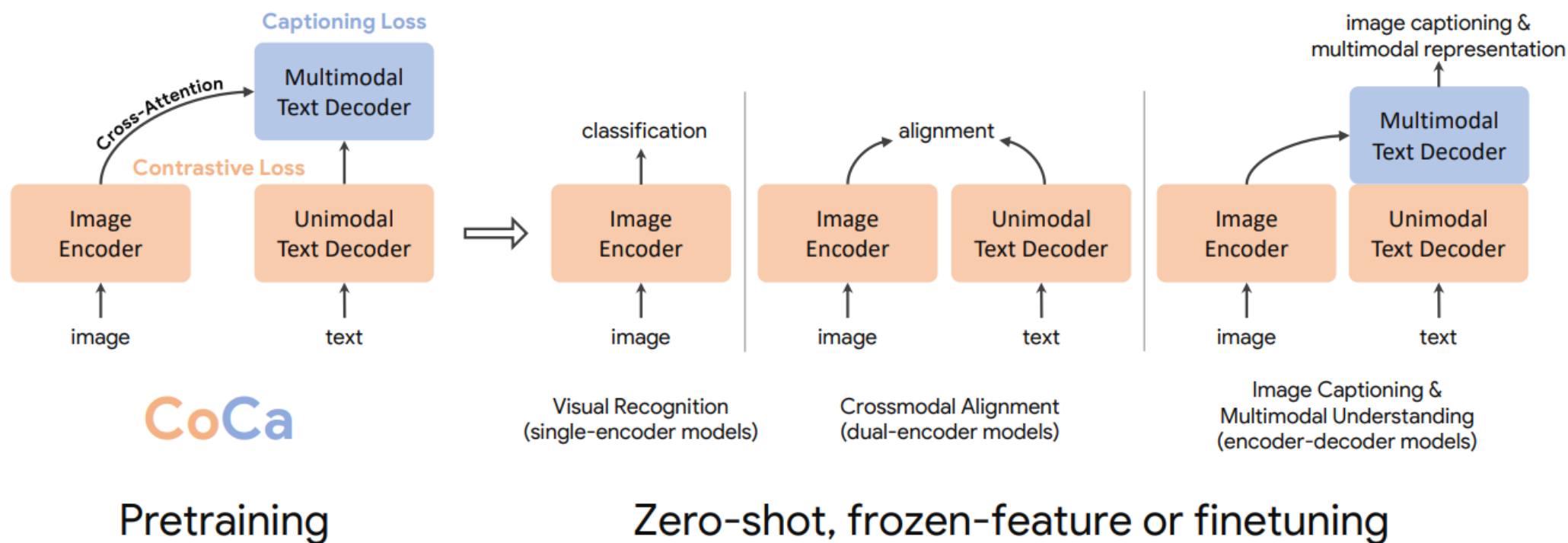
(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

(ICML 2021) Learning Transferable Visual Models From Natural Language Supervision. Radford, Alec, et al.

Vision-Language Models are Better Classifiers



(TMLR 2022) CoCa: Contrastive Captioners are Image-Text Foundation Models. Yu, et al.

Are Vision-Language Models Better Classifiers?





The image features a **large brown dog** and a small cat **sitting together on a couch**. The dog is positioned on the left side of the couch, while the cat is on the right side. They appear to be enjoying each other's company, **with the dog's paw resting on the cat's paw**. In the background, **there is a person sitting on the couch**, possibly observing the interaction between the dog and the cat.



A red cube, on top of a yellow cube,
to the left of a green cube

Can we teach models to distinguish attributes, states, and relations?

The principle of compositionality:

The meaning of a complex expression is a function of the meaning of its constituents and the way they are combined. – Zoltán Gendler Szabó, The case of compositionality.

What is compositionality?

The cat is on the mat



vs

The mat is on the cat



What is compositionality?

The **cat** is on the **mat**

vs

The **mat** is on the **cat**



The promise of synthetic data

*“Synthetic data is computer-generated information for testing and training AI models that have become indispensable in our **data-driven era**.*

*It’s **cheap to produce**, comes **automatically labeled**, and sidesteps many of the logistical, ethical, and privacy issues that come with training deep learning models on real-world examples.*

The research firm Gartner estimates that, by 2030, synthetic data will overtake actual data in training AI models.”

Meta May Turn to Synthetic Data to Feed its AI Growth

Nat Rubio-Licht

October 5, 2023



Deep Dives

PATENT DROP

Google's Smart S
to Global Data M



FINANCE

Berkshire Hath
Massive Mounta

<https://www.thedailyupside.com/technology/artificial-intelligence/meta-may-turn-to-synthetic-data-to-feed-its-ai-growth/>

*Exploring 3D Simulation Platforms to generate realistic **synthetic data** is a promising direction; it may provide similar insights of real data without exposing sensitive information.*

SimVQA: Exploring Simulated Environments for Visual Question Answering

ThreeDWorld (TDW)
physics-based realistic
simulation platform



(CVPR 2022) SimVQA: Exploring Simulated Environments for Visual Question Answering. Cascante-Bonilla et al.



Q/ Is there a table in the room? A/ yes



Q/ How many chairs are in the picture? A/ 2



Q/ What color is the fire hydrant? A/ yellow



Q/ Is there a teddy bear on top of the table? A/ yes

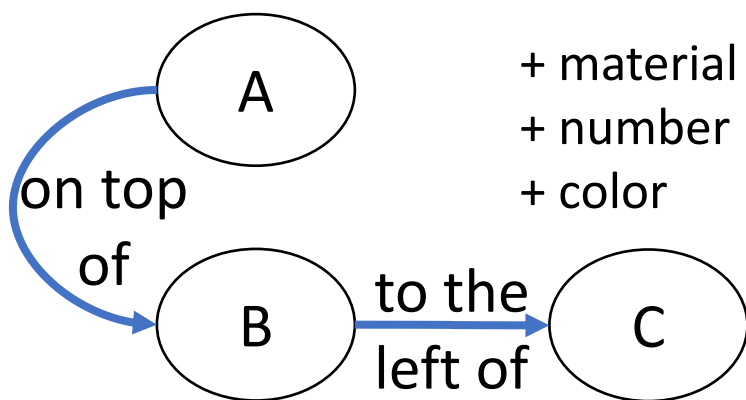


Q/ How many lamps are in the room? A/ 3

Synthetic Data Generation



From a Scene Graph:



We can generate question/answer pairs:

Q/ What is **[position]** the **[object_b]**?

A/ a **[object_a]**

Q/ What **is on top of** the **table**?

A/ a **brown backpack**

Q/ What **is on top of** the **table**? A/ a **red lamp**

Synthetic to Real-world data

Learning to count on real data using only synthetic samples.

Training data			R-VQA Accuracy		
Real	Synthetic				
R-VQA _{NC}	H-VQA _C	W-VQA _C	Numeric	Others	Overall
✓			6.08	68.94	60.69
✓	✓		15.99 _{+9.91}	68.97	62.02 _{+1.33}
✓		✓	21.18 _{+15.1}	68.91	62.65 _{+1.96}
✓	✓	✓	24.96 _{+18.8}	68.91	63.14 _{+2.45}

Synthetic Data Problem: Domain Shift

A gap between real and synthetic images.

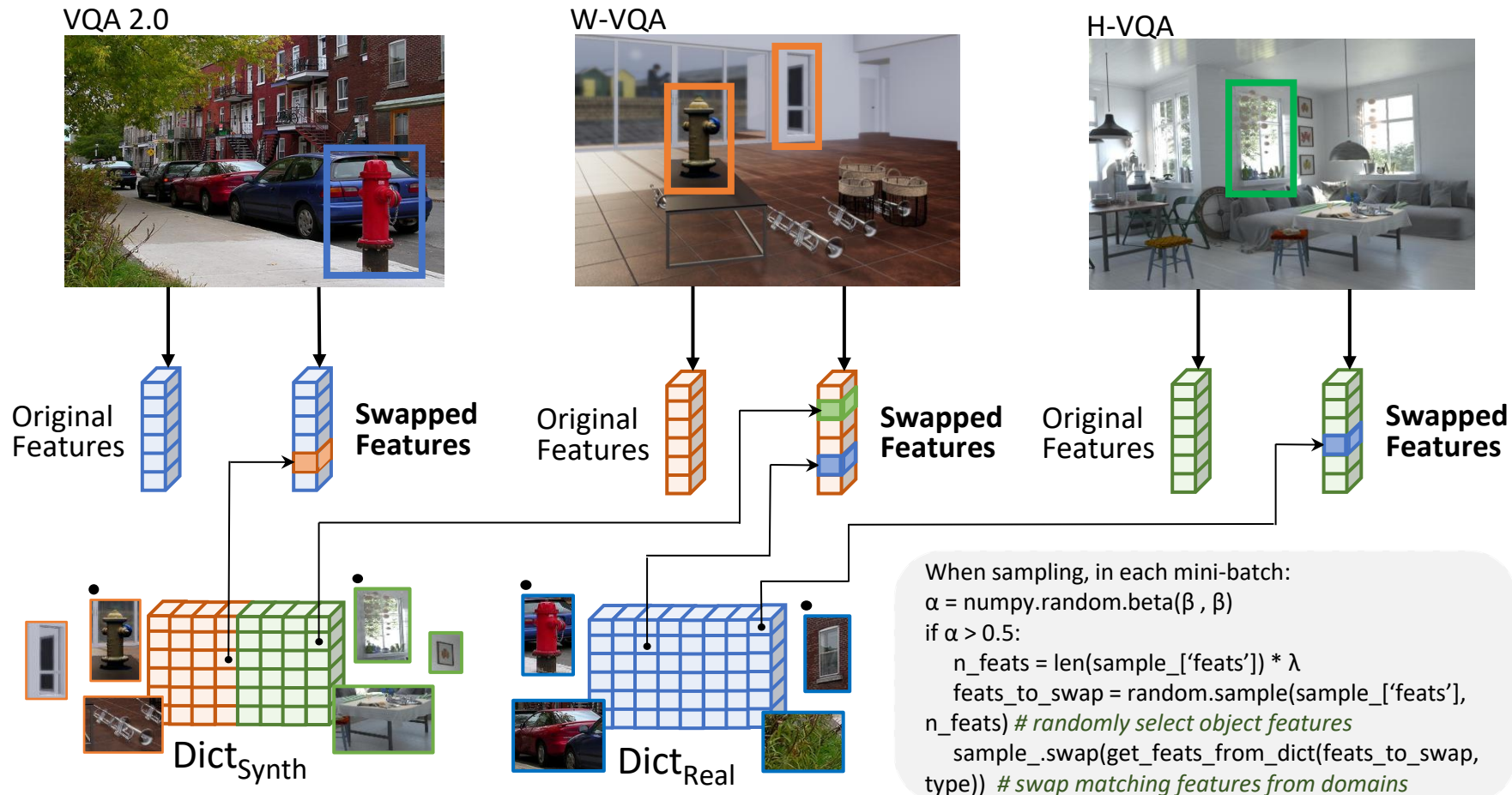


VS



Solution: Feature Alignment

Proposed Method: Feature Swapping (F-SWAP)



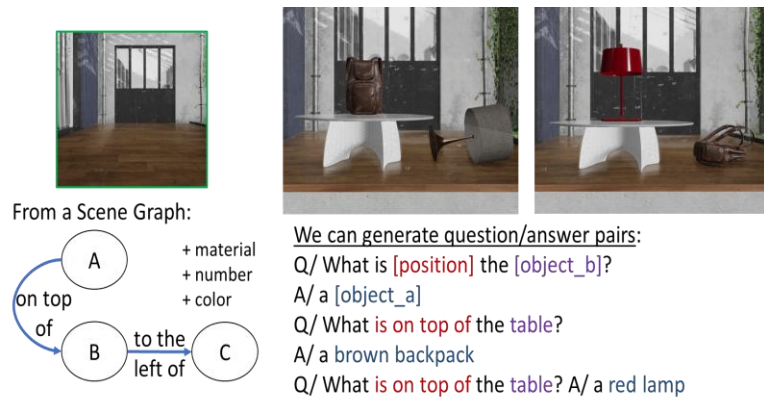
Synthetic to Real-world data

Learning to count under different low-regime settings for VQA v2.

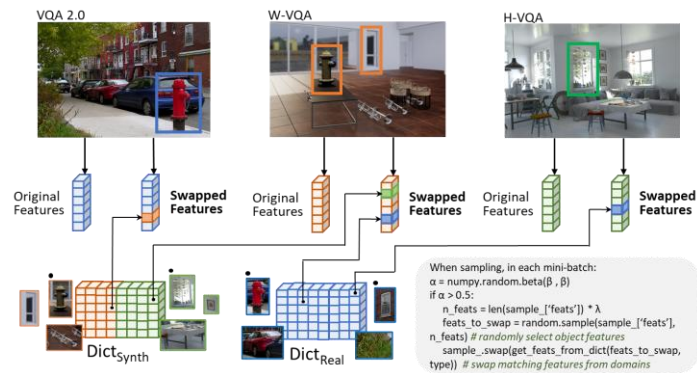
Data	Method	+0% R-VQA _C		
		<i>Numeric</i>	<i>Others</i>	<i>Overall</i>
H-VQA _C	Simple Augmentation	15.99	68.97	62.02
H-VQA _C	Feature Swapping (F-SWAP)	23.38 _{+7.39}	69.07 _{+0.10}	63.07 _{+1.05}
W-VQA _C	Simple Augmentation	21.18	68.91	62.65
W-VQA _C	Feature Swapping (F-SWAP)	26.84 _{+5.66}	68.89 _{-0.02}	63.67 _{+1.02}

Contributions:

Generation Pipeline



Transferable Knowledge



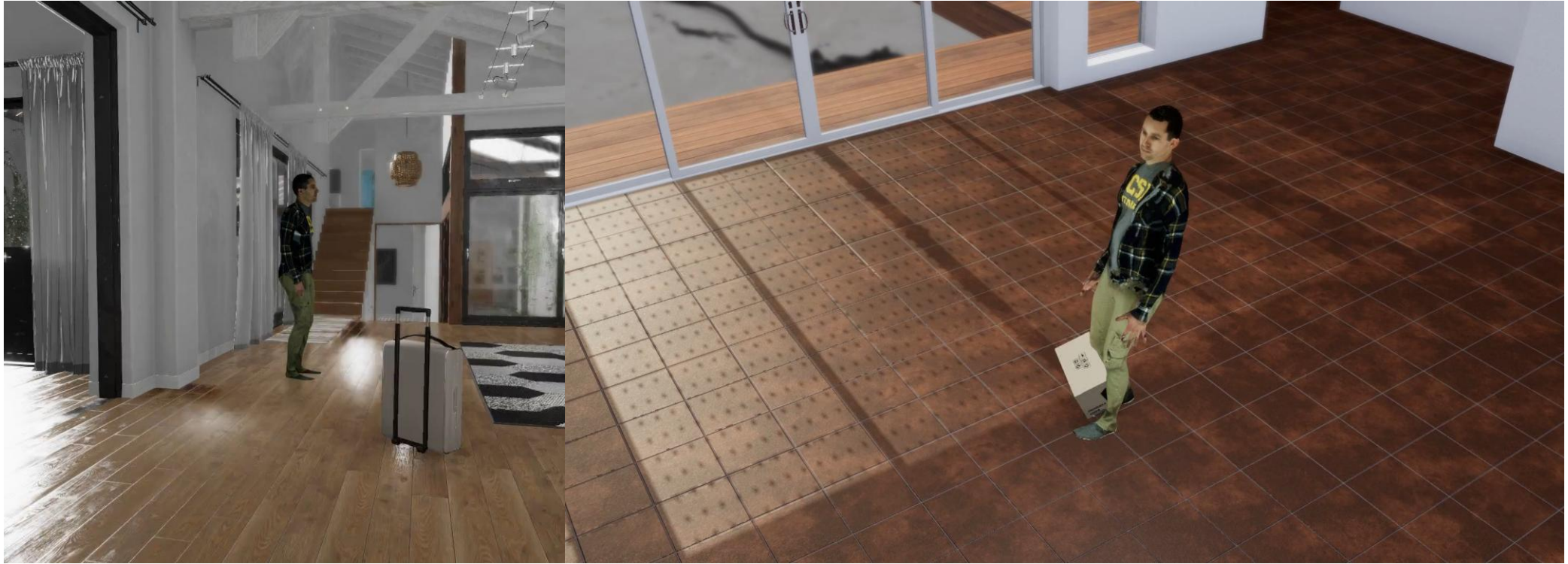
Avoid Privacy Concerns



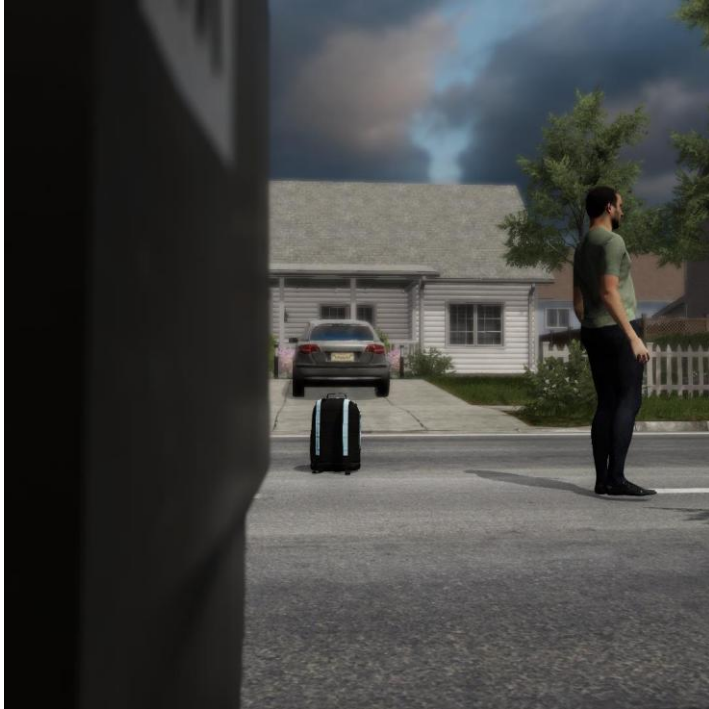
Synthetic images vs real images



Learning from *better* Synthetic Data



Learning from *better* Synthetic Data



This scene contains a backpack and one human. The human is to the right of the backpack. The backpack is behind the human. The backpack is to the left of the human. The human is in front of the backpack. The human side walks. The human is male. The human wears a green t-shirt and black pants. The human has short light-brown hair and a beard (...)



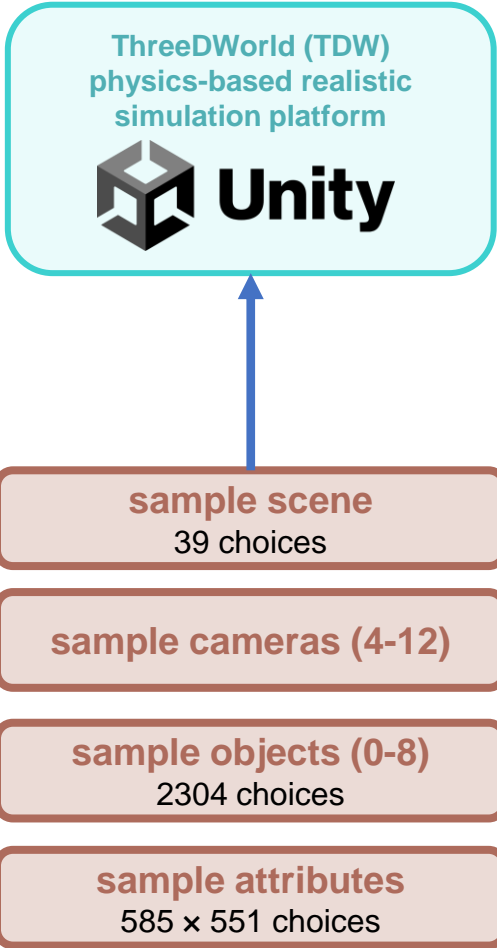
This scene contains a gift wrapping and two humans. They are in a street with grey floor, green plants on the side, and houses around. The first human is to the right of the gift wrapping. The first human is walking. The first human wears a red shirt and solid black pants. The first human has brown hair. The second human stands straight. The second human has brown hair (...)

Going Beyond Nouns ... Using Synthetic Data



(ICCV 2023) Going Beyond Nouns With Vision & Language Models Using Synthetic Data. International Conference on Computer Vision. Cascante-Bonilla et al.

Going Beyond Nouns...



An image of a *chair*
next to a *backpack*



An image of a *brown* table *on top*
of a *white* table

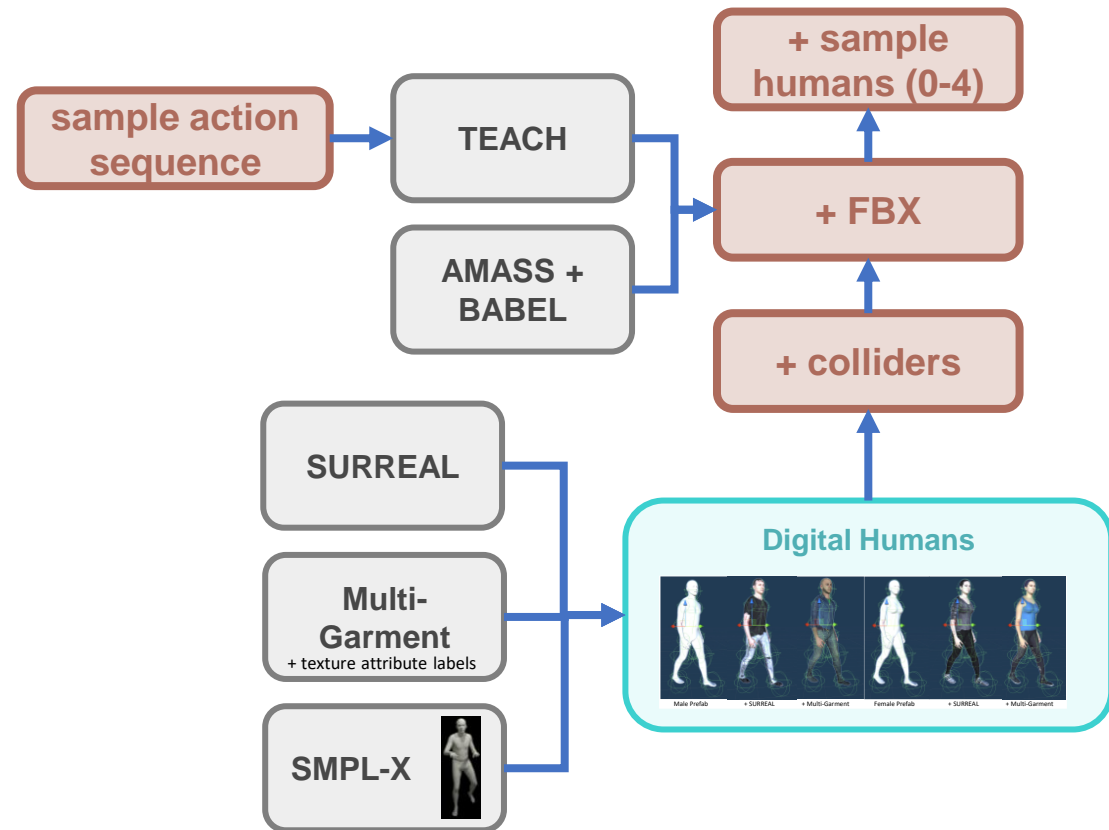
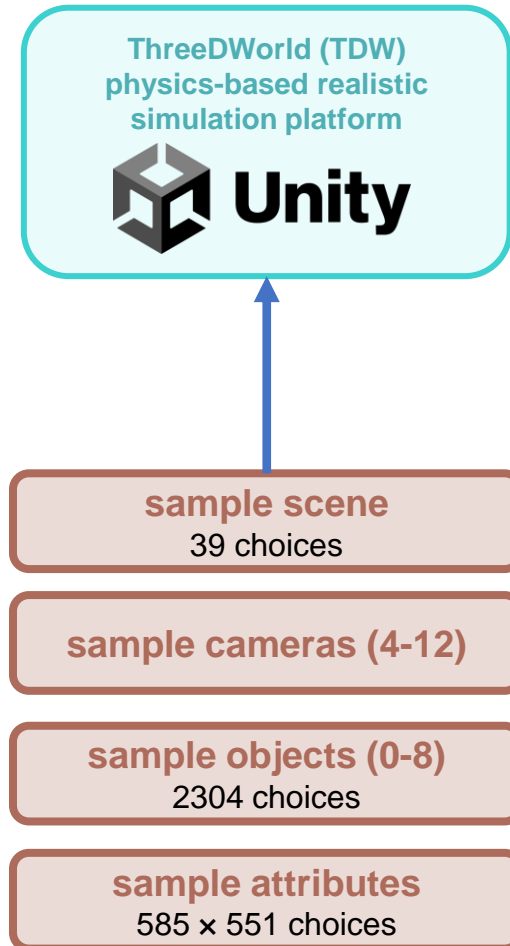


A person *standing in*
front of a *window*



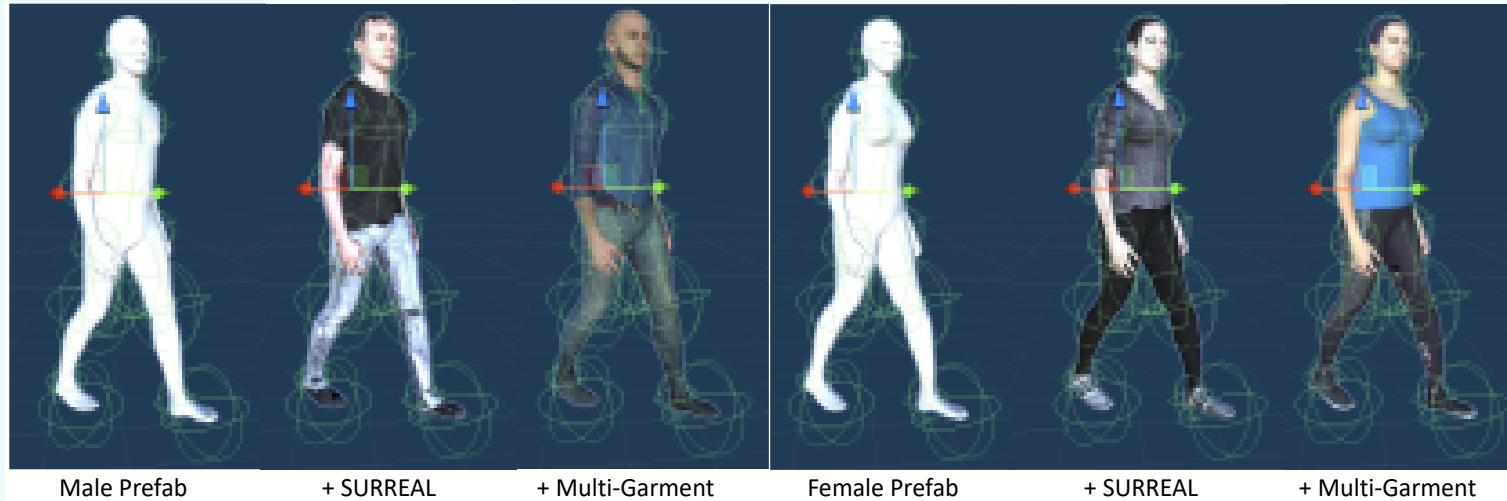
An *old* woman *talking*
to the *boy* *with the*
pale turquoise shirt

Going Beyond Nouns...



Going Beyond Nouns...

Digital Humans



Going Beyond Nouns...



synthesize

SyViC Dataset 767K
image+text pairs

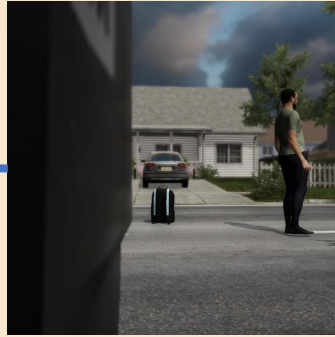


This scene contains a gift wrapping and two humans. They are in a street with grey floor, green plants on the side, and houses around. The first human is to the right of the gift wrapping. The first human is walking. The first human wears a red shirt and solid black pants. The first human has brown hair. The second human stands straight. The second human has brown hair. The second human is wearing blue jeans pants, brown shoes, and a white shirt. The second human is male.



Caption

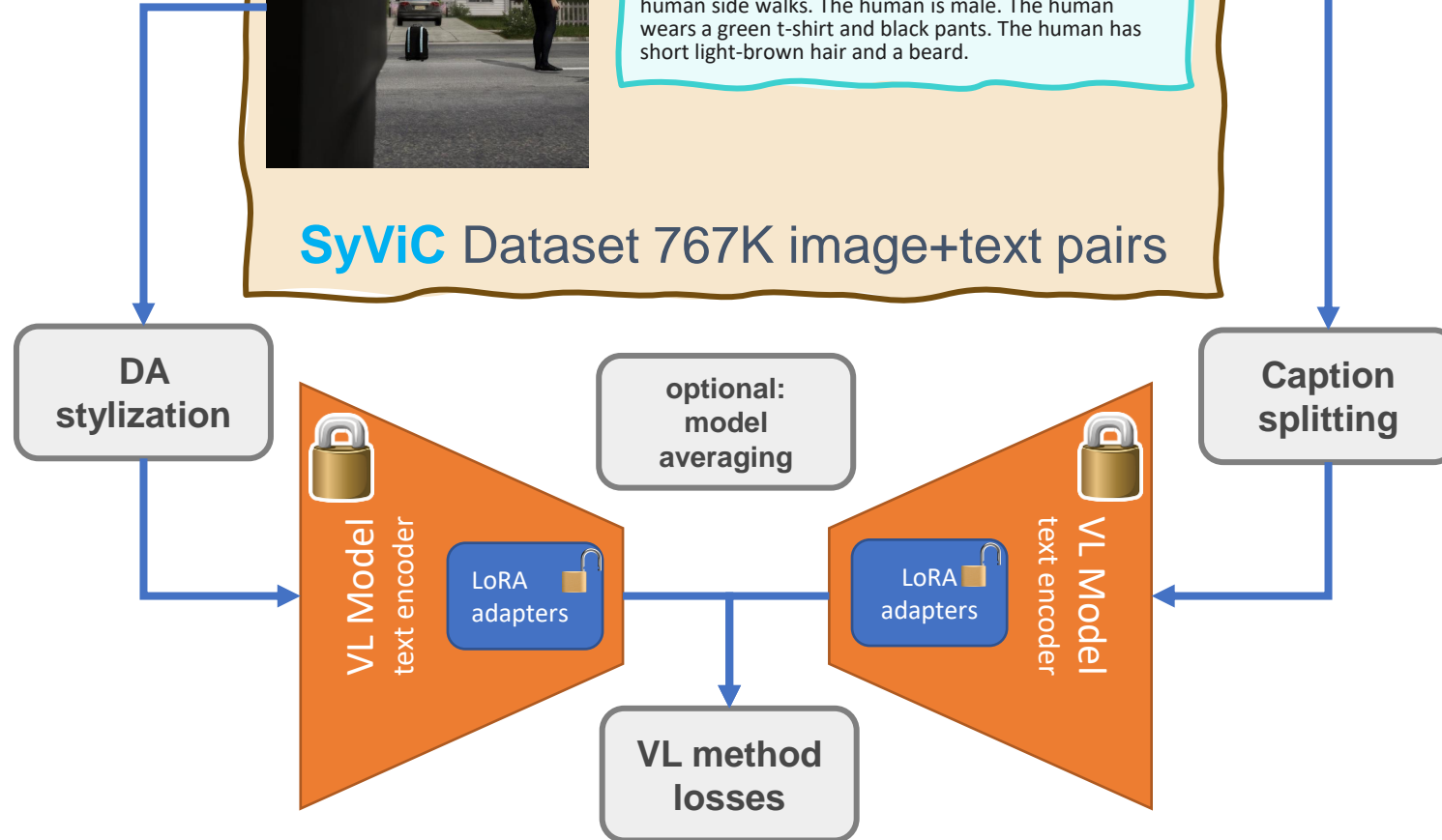
This scene contains a gift wrapping and two humans. They are in a street with grey floor, green plants on the side, and houses around. The first human is to the right of the gift wrapping. The first human is walking. The first human wears a red shirt and solid black pants. The first human has brown hair. The second human stands straight. The second human has brown hair. The second human is wearing blue jeans pants, brown shoes, and a white shirt. The second human is male.



Caption

This scene contains a backpack and one human. The human is to the right of the backpack. The backpack is behind the human. The backpack is to the left of the human. The human is in front of the backpack. The human side walks. The human is male. The human wears a green t-shirt and black pants. The human has short light-brown hair and a beard.

SyViC Dataset 767K image+text pairs



Some Results on Compositional Evaluations

VL-Checklist Benchmark

Attribute



Color

[POS]: sheep is **white**.
[NEG]: sheep is **golden brown**.

Material

[POS]:sheep is **furry**.
[NEG]:sheep is **hardwood**.

Relation



Action

[POS]: child **brushing** teeth.
[NEG]: child **photographing** teeth.

Spatial

[POS]:shirt **on** boy.
[NEG]:shirt **under** boy.

CLIP 63.57% ➔ 69.39% ↑ 4.34%
 67.51% ➔ 70.37%

Some Results on Compositional Evaluations

VL-Checklist Benchmark

Attribute



Color

[POS]: sheep is **white**.
[NEG]: sheep is **golden brown**.

Material

[POS]:sheep is **furry**.
[NEG]:sheep is **hardwood**.

Relation



Action

[POS]: child **brushing** teeth.
[NEG]: child **photographing** teeth.


Spatial

[POS]:shirt **on** boy.
[NEG]:shirt **under** boy.

CLIP * 63.57% → 78.31% ↑ 10.77%
67.51% → 74.31%

* SyViC + Text Augmentation

Learning from *better* Synthetic Data

CLIP			syn-CLIP		
		<p>1) there is a table below someone</p> <p>2) there is someone below a table</p>			<p>1) there is a table below someone</p> <p>2) there is someone below a table</p>
					
"a table"	"someone"		"a table"	"someone"	

Detailed Gains

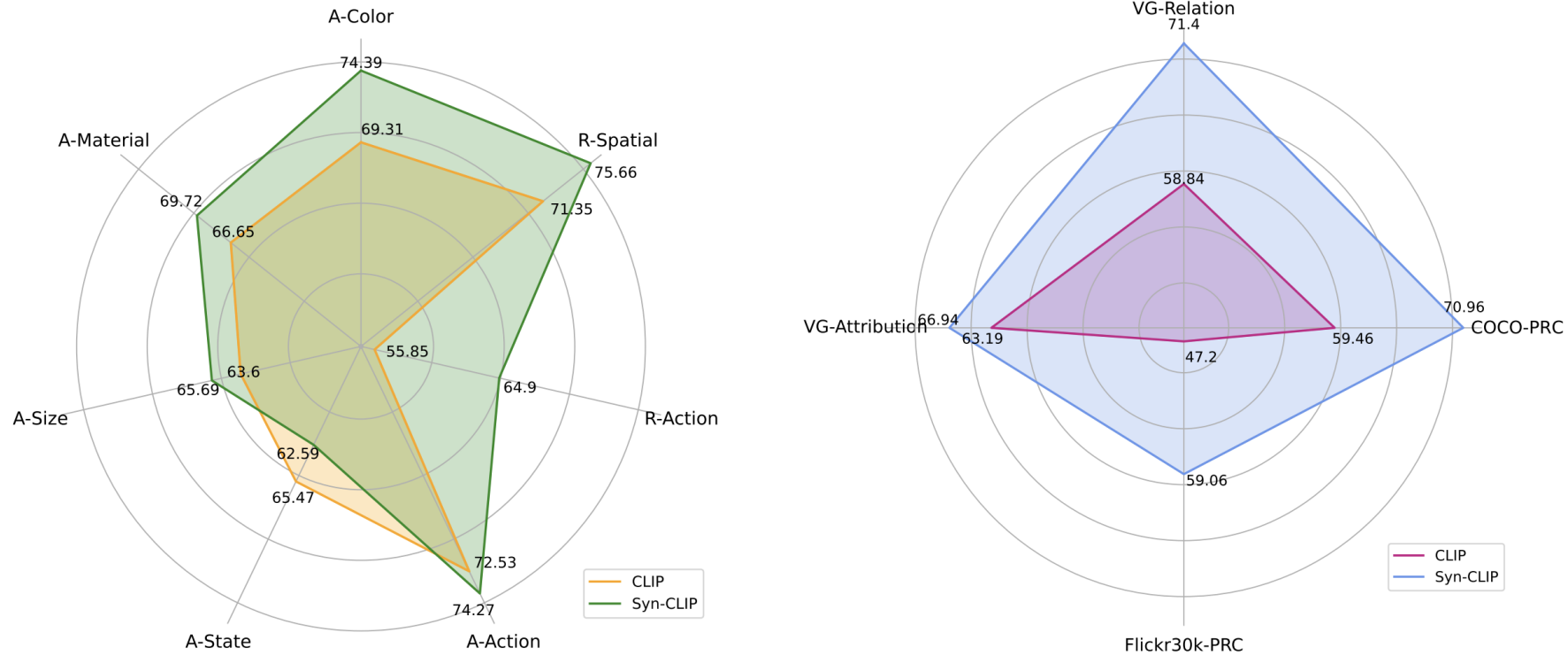
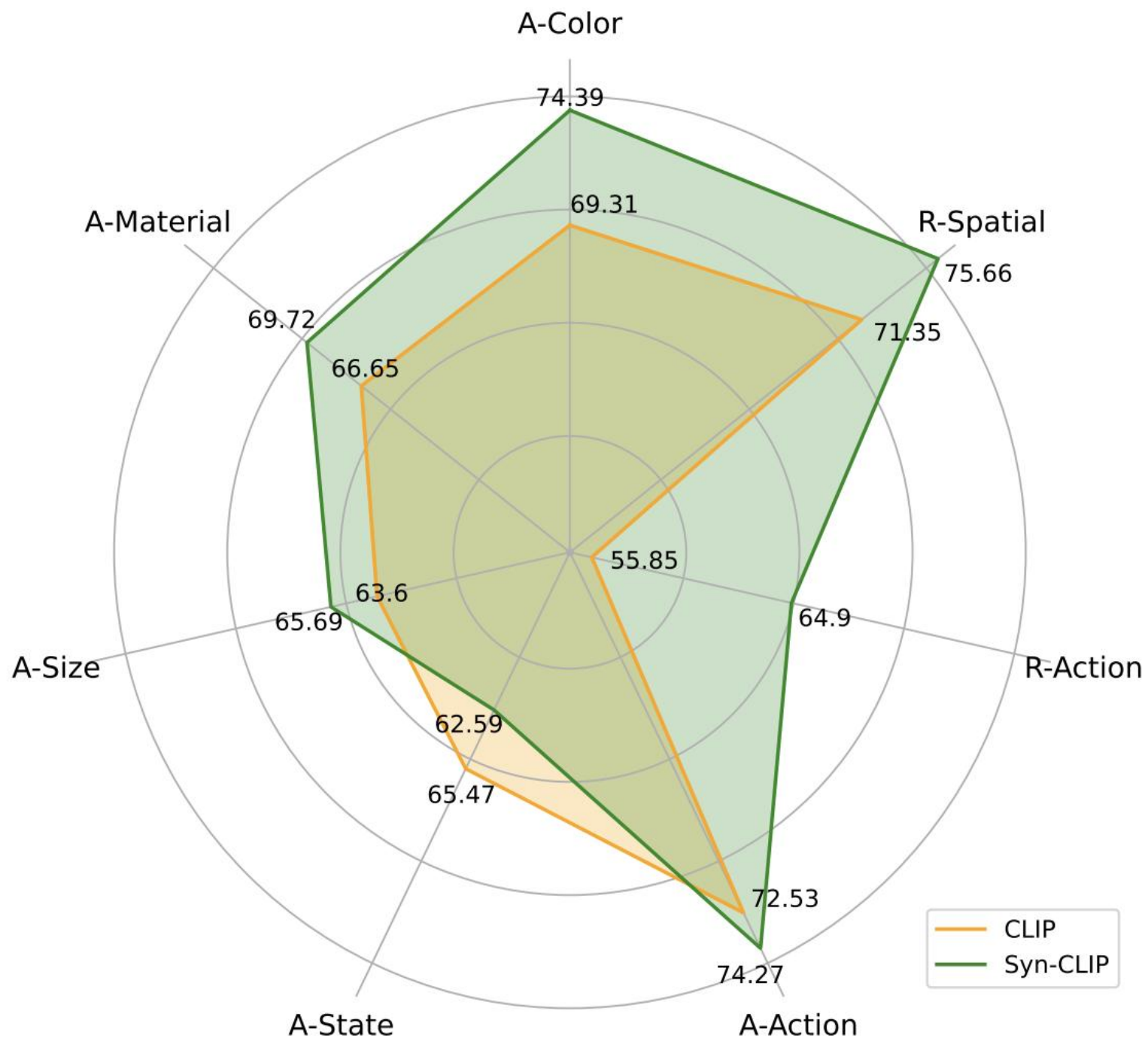


Figure 3. **(left)** detailed evaluation of syn-CLIP on the 7 separate VL-Checklist [70] metrics; **(right)** detailed evaluation of syn-CLIP on all the Compositional Tasks proposed in ARO [66].



Detailed Gains

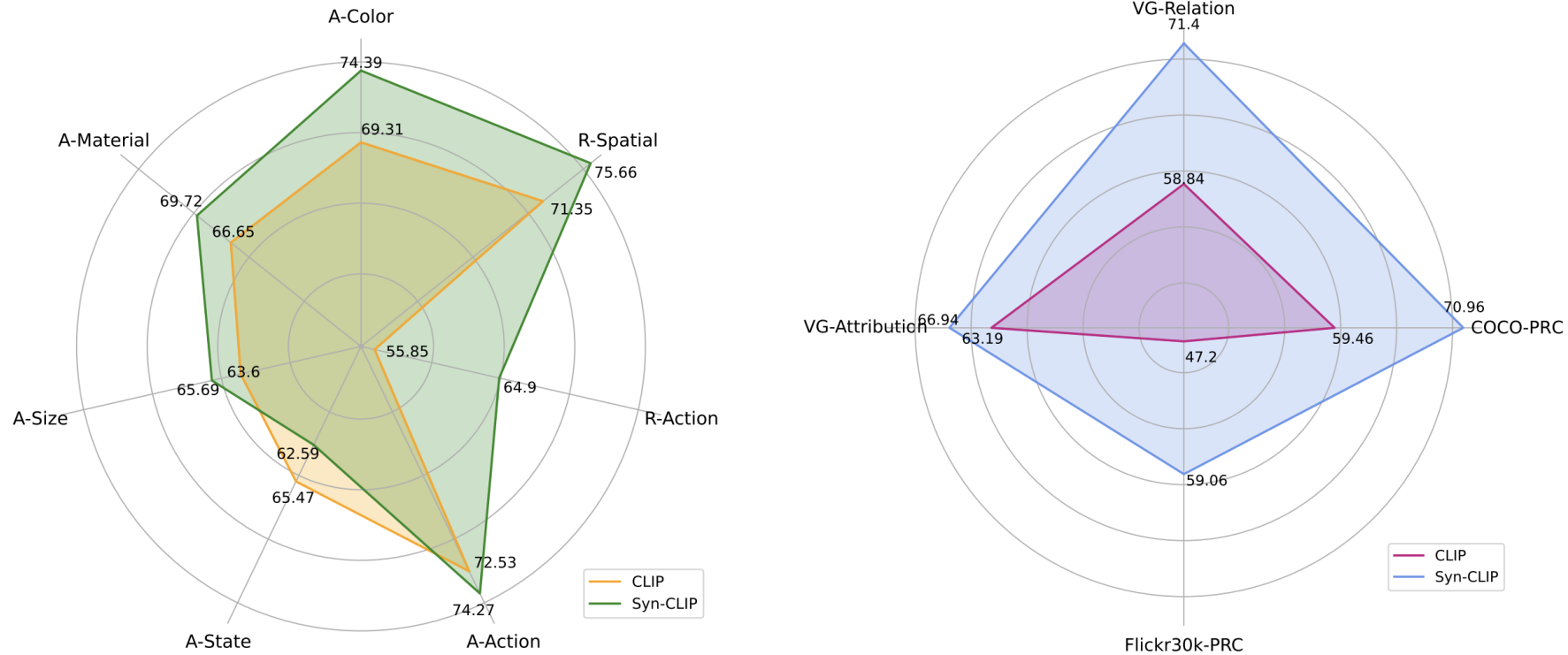
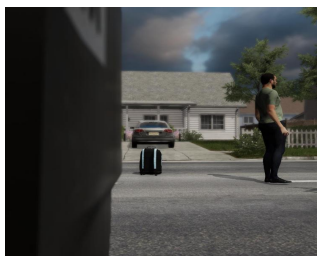


Figure 3. **(left)** detailed evaluation of syn-CLIP on the 7 separate VL-Checklist [70] metrics; **(right)** detailed evaluation of syn-CLIP on all the Compositional Tasks proposed in ARO [66].

Contributions:

Generation Pipeline

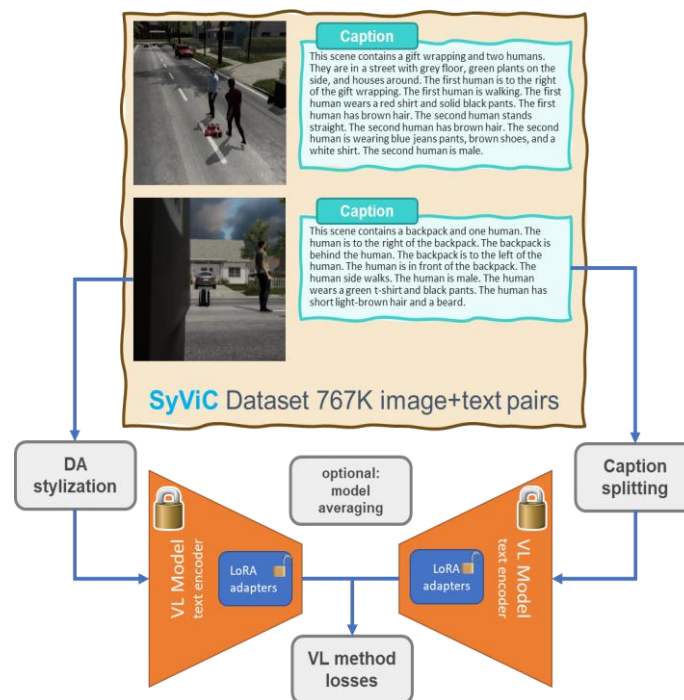


This scene contains a backpack and one human. The human is to the right of the backpack. The backpack is behind the human. The backpack is to the left of the human. The human is in front of the backpack. The human side walks. The human is male. The human wears a green t-shirt and black pants. The human has short light-brown hair and a beard (...)



This scene contains a gift wrapping and two humans. They are in a street with grey floor, green plants on the side, and houses around. The first human is to the right of the gift wrapping. The first human is walking. The first human wears a red shirt and solid black pants. The first human has brown hair. The second human stands straight. The second human has brown hair. The second human is wearing blue jeans pants, brown shoes, and a white shirt. The second human is male. (...)

Transferable Knowledge



Compositionality without catastrophic forgetting

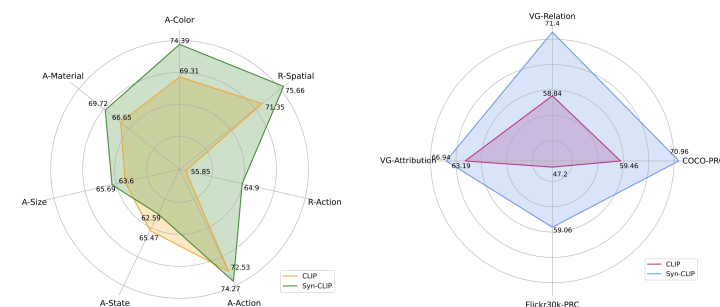
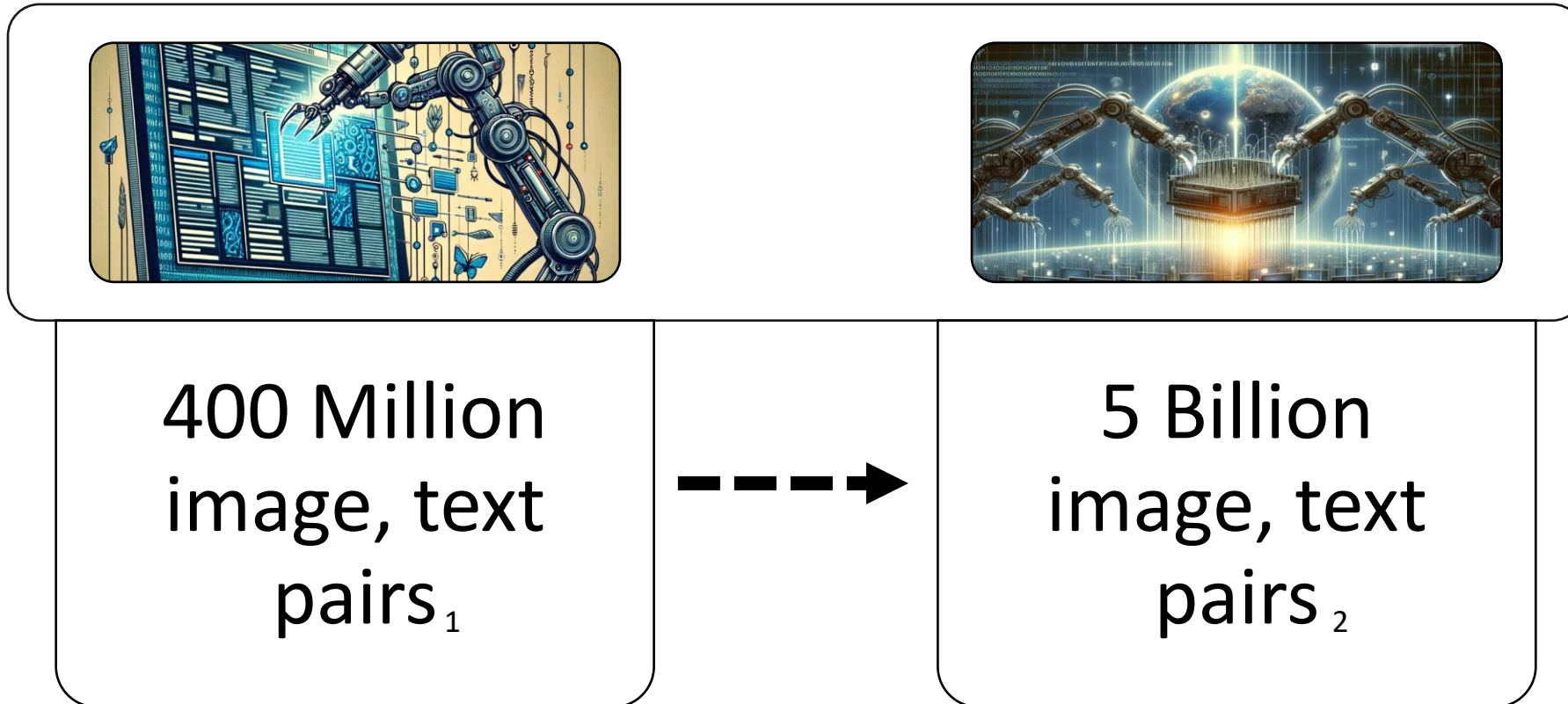


Figure 3. **(left)** detailed evaluation of syn-CLIP on the 7 separate VL-Checklist [70] metrics; **(right)** detailed evaluation of syn-CLIP on all the Compositional Tasks proposed in ARO [66].

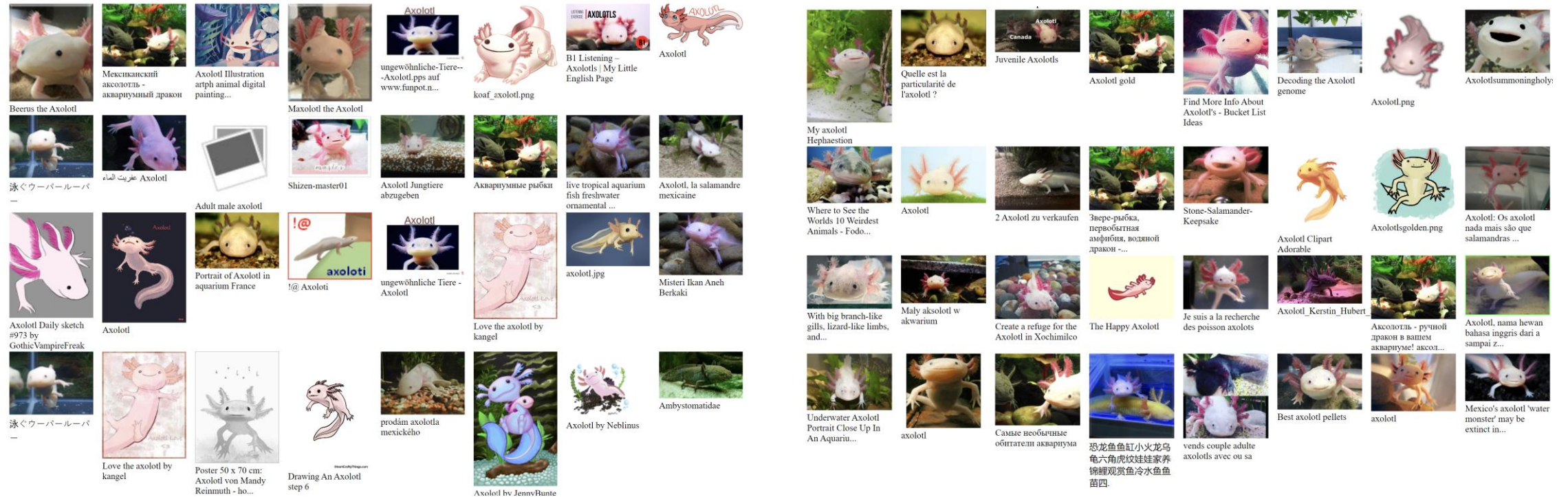
Large Vision & Language Models are trained
with **Massive** amounts of data

Large-scale Multimodal Models (LMMs)

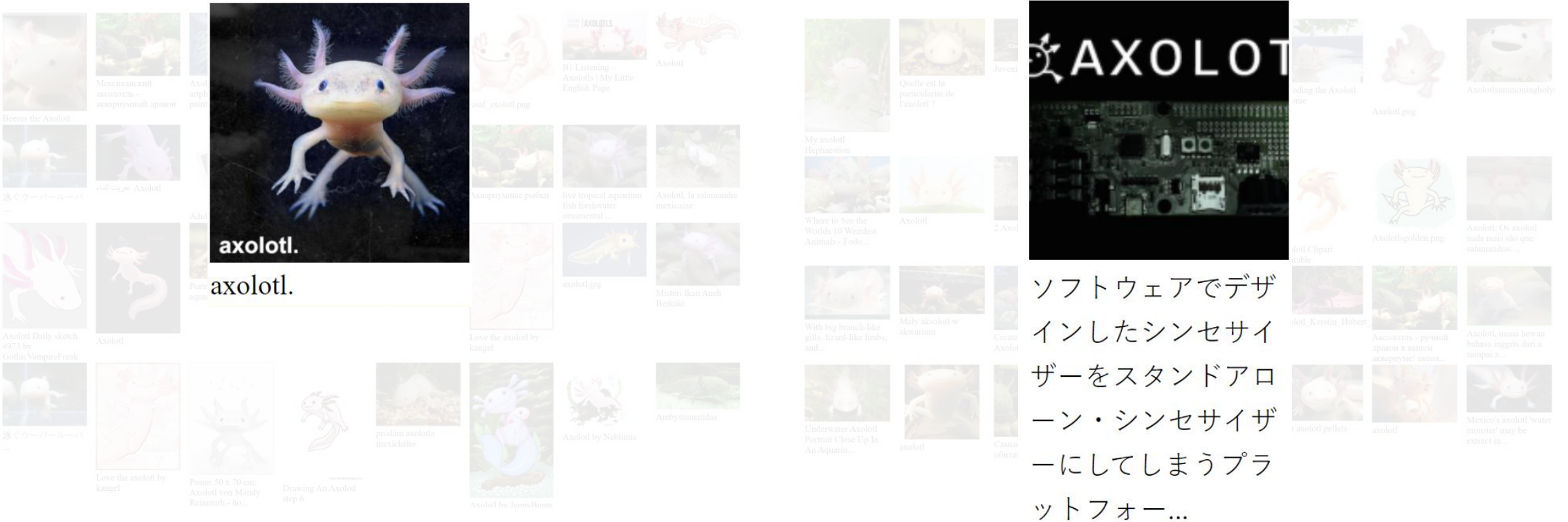


- 1) Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.
- 2) Schuhmann, Christoph, et al. "Laion-5b: An open large-scale dataset for training next generation image-text models." Advances in Neural Information Processing Systems 35 (2022): 25278-25294.

Internet-scale data



Internet-scale data



VLMs can output detailed textual descriptions



This image shows a warm, intimate moment between a couple in a modern kitchen. The couple is standing close to each other near a dishwasher, with the woman barefoot and leaning in, holding a cup while gently touching the man's face. The man is smiling and has one arm around her waist, creating a cozy and affectionate scene. Natural light is coming through the windows behind them, and the kitchen has a clean, contemporary design with white upper cabinets, dark lower cabinets, and a kitchen island. There's a bottle of olive oil and some kitchen items on the counter.

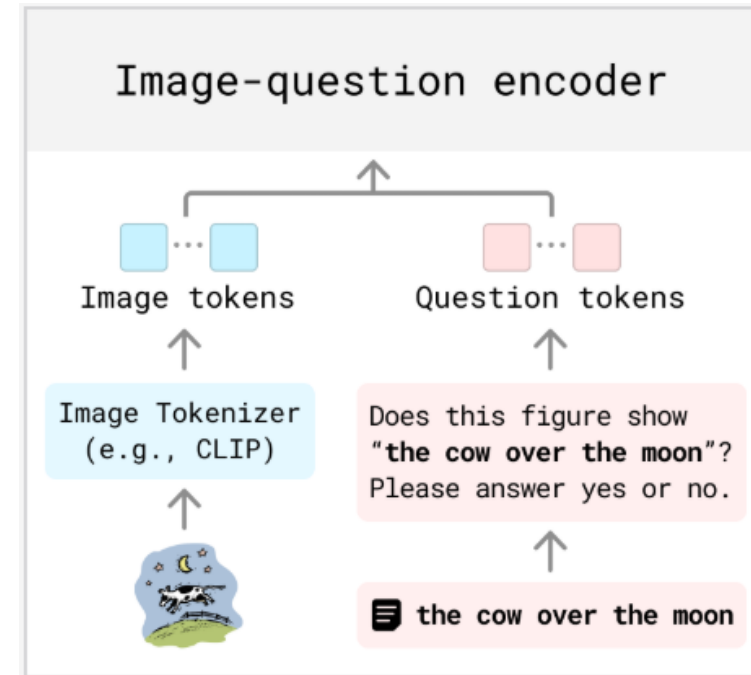
VLMs hallucinate / outputs are wrong



This image shows a warm, intimate moment between a couple in a modern kitchen. The couple is standing close to each other near a dishwasher, with the woman barefoot and leaning in, holding a cup while gently touching the man's face. **The man is smiling and has one arm around her waist, creating a cozy and affectionate scene.** Natural light is coming through the windows behind them, and the kitchen has a clean, contemporary design with white upper cabinets, dark lower cabinets, and a kitchen island. There's a bottle of olive oil and some kitchen items on the counter.

Is there a better way to unlock compositionality?

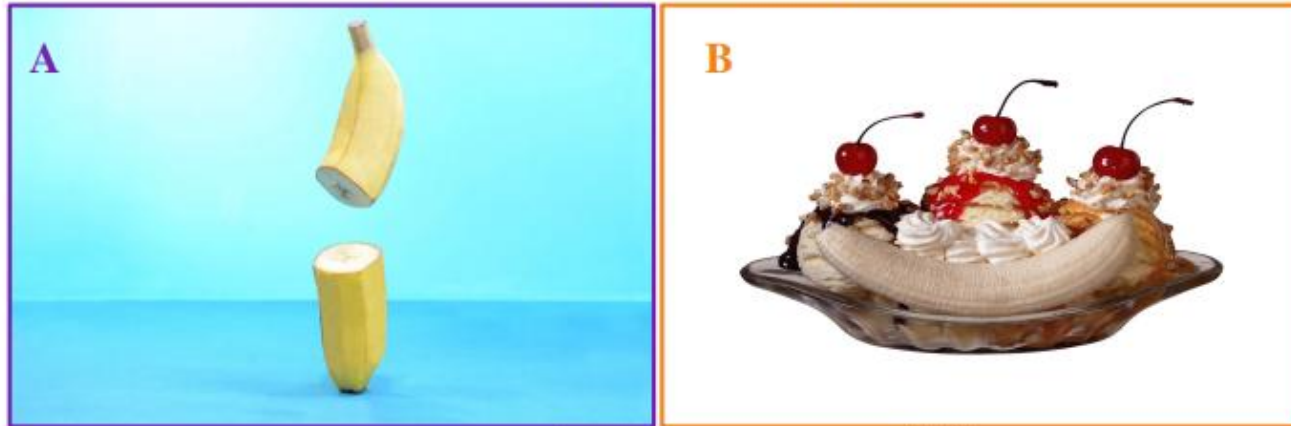
VQAScore: Compute the probability of 'Yes' conditioned on the image and a simple question.



(ECCV 2024) Evaluating Text-to-Visual Generation with Image-to-Text Generation. Lin et al.

Common Sense Reasoning is Hard

Original Caption/Image Pair



Text (T)	$P(\text{Yes} \mid T, \text{A})$	$P(\text{Yes} \mid T, \text{B})$
there is a split banana	60.6%	64.6%



VQAScore

Decomposing Sentences via Semantics?

Original Caption/Image Pair





Text (T)	P(Yes T, A)	P(Yes T, B)
there is a split banana	60.6%	64.6%

Sentence Decomposition via Semantics (SDS)

Text (T)	P(Yes T, A)	P(Yes T, B)
there is a banana	90.4%	93.1%
the banana is split	45.0%	79.9%
SDS Final Score	63.8%	86.2%

Caption Expansion via Natural Language Inference!

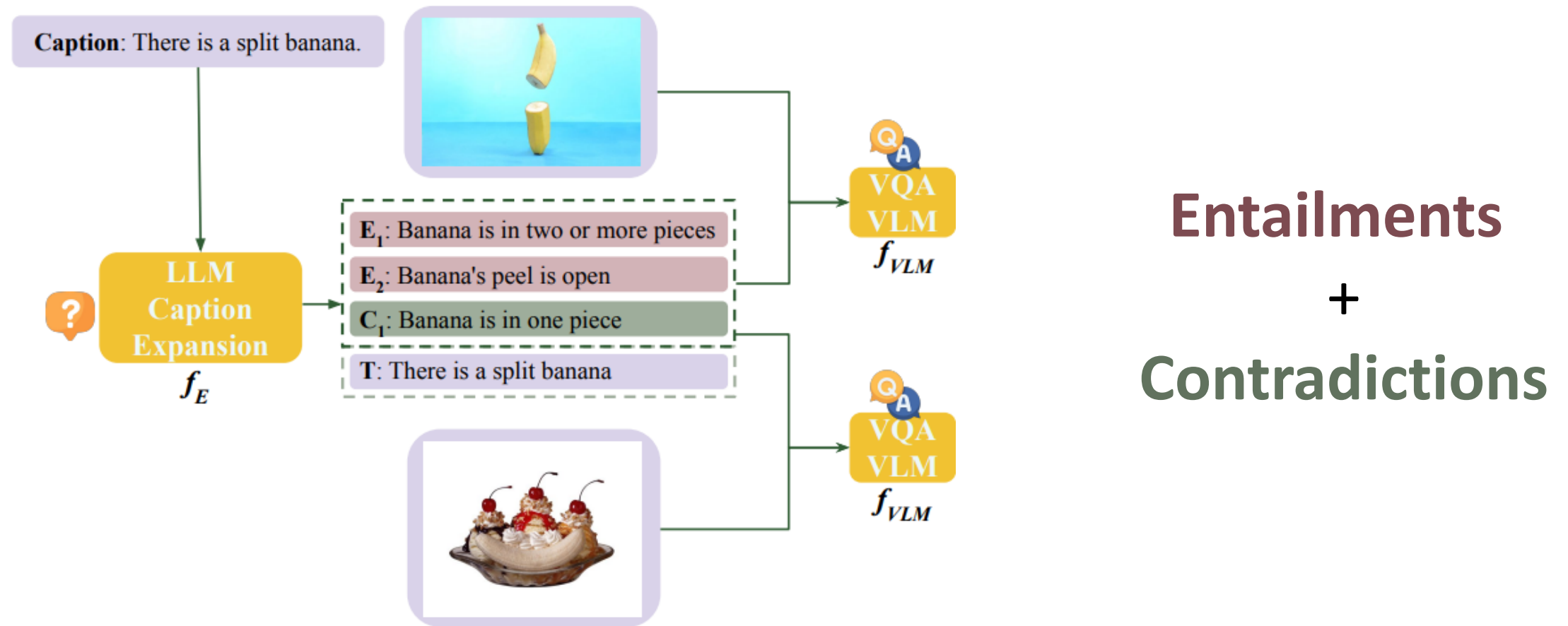
Original Caption/Image Pair

<div><div><div>A</div></div><div><div>B</div></div></div>		
Text (T)	P(Yes T, A)	P(Yes T, B)
there is a split banana	60.6%	< 64.6%

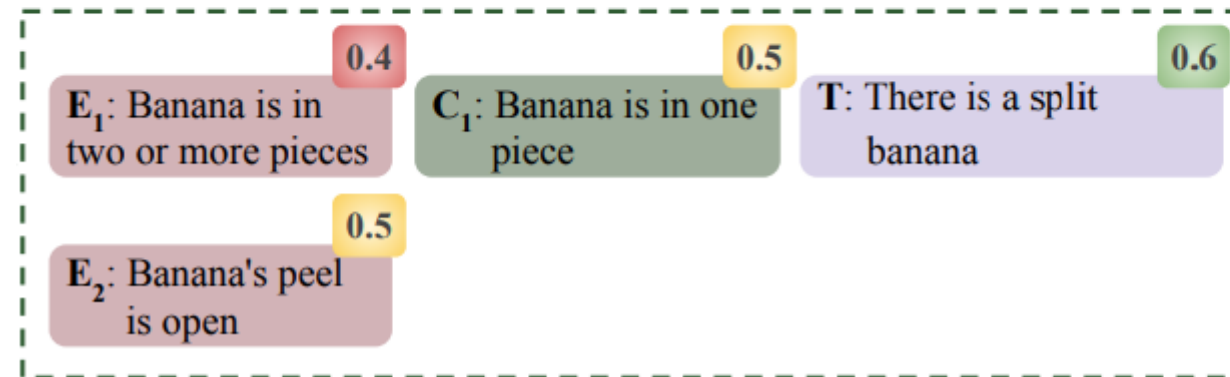
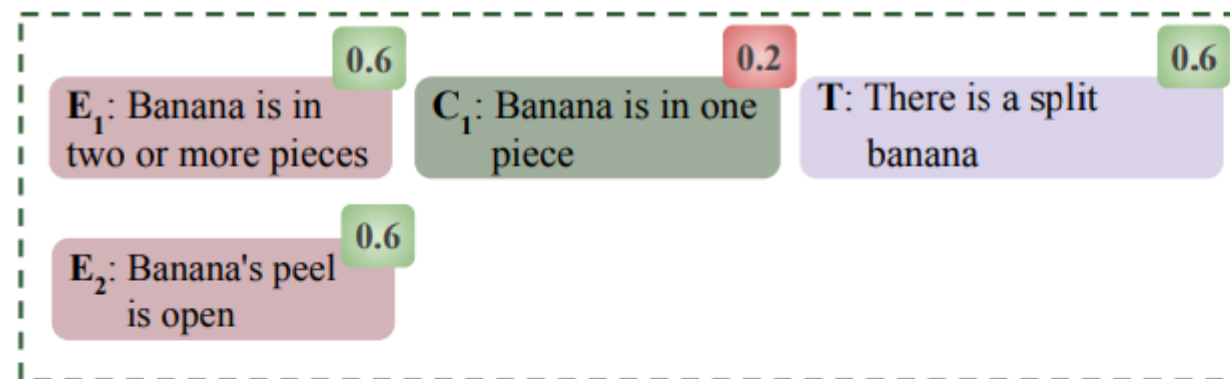
Caption Expansion via CECE

Text (T)	P(Yes T, A)	P(Yes T, B)
banana is in two or more pieces	62.1%	43.0%
banana's peel is open	61.9%	46.9%
CECE Final Score	62.0%	> 44.9%

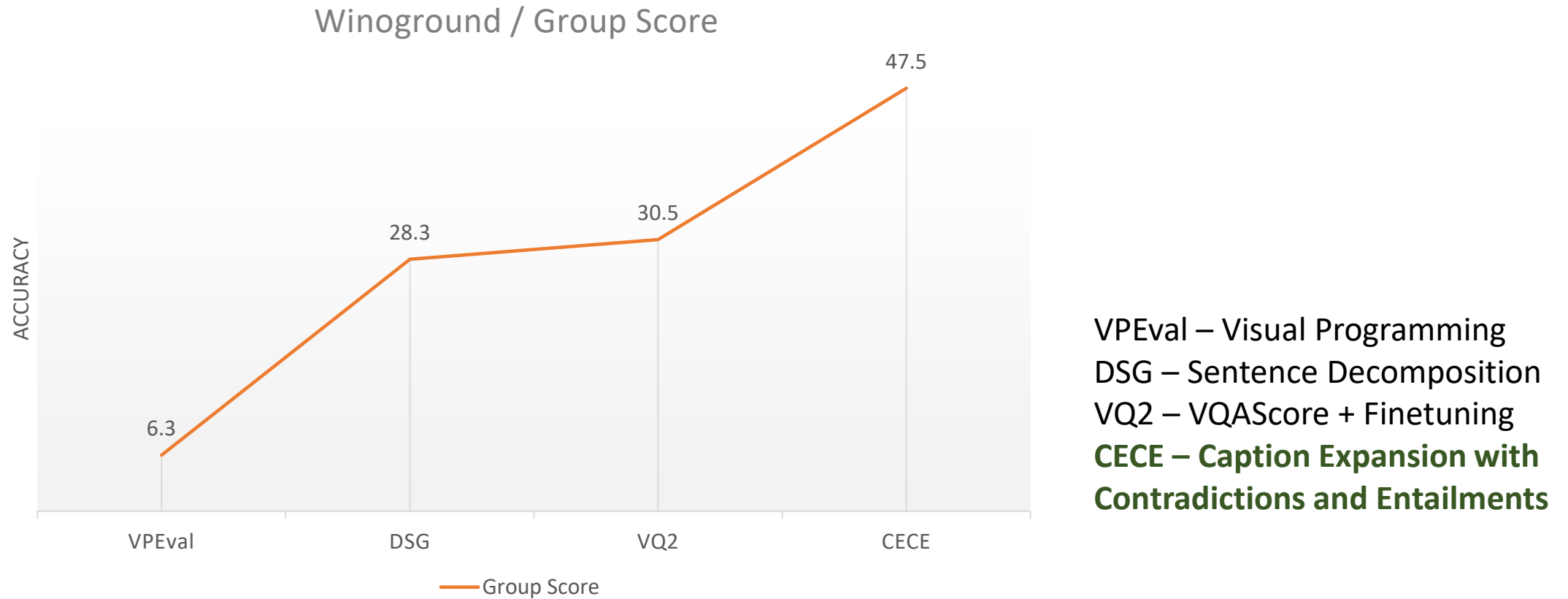
Natural Language Inference Improves Compositionality in VLMs



Diversity uncovers existing vision-language alignment



CECE works... without finetuning!





(ICLR 2025) Natural Language Inference Improves Compositionality in VLMs. Cascante-Bonilla et al.

Score agreement with human judgments of alignment for image-text alignment

Method	Tools- f_{VLM}	LLM- f_E	DrawBench	EditBench	COCO-T2I	TIFA160	Pick-a-Pic
<i>End-to-end models</i>							
CLIPScore (Radford et al., 2021)	CLIP-L-14	–	49.1	60.6	63.7	54.1	76.0
BLIP2 _{ITM} (Li et al., 2023)	BLIPv2	–	60.5	68.0	70.7	57.5	80.0
VQAScore (Lin et al., 2024)	InstructBLIP	–	82.6	75.7	83.0	70.1	83.0
VQAScore (Lin et al., 2024)	LLaVA-1.5	–	82.2	70.6	79.4	66.4	76.0
VIEScore (Ku et al., 2023)	GPT4-Vision	–	–	–	–	63.9	78.0
GPT4V-Eval (Zhang et al., 2023)	GPT4-Vision	–	–	–	–	64.0	74.0
<i>Sentence Decomposition via Semantics (SDS)</i>							
VQ2 (Yarom et al., 2023)	LLaVA-1.5	FlanT5	52.8	52.8	47.7	48.7	73.0
DSG (Cho et al., 2023a)	LLaVA-1.5	ChatGPT	78.8	69.0	76.2	54.3	70.0
VQ2 (Yarom et al., 2023)	PaLI-17B	FlanT5	82.6	73.6	83.4	–	–
TIFA (Hu et al., 2023)	PaLI-17B	Llama2	73.4	67.8	72.0	–	–
<i>Caption Expansion with Contradictions and Entailments (CECE)</i>							
CECE (Ours)	InstructBLIP	Llama3.1	85.4	76.7	81.4	69.3	84.0
CECE (Ours)	LLaVA-1.5	Llama3.1	87.3	75.6	81.3	68.9	86.0
CECE (Ours)	LLaVA-1.6	Llama3.1	86.3	75.9	83.8	70.4	83.0
CECE (Ours)*	LLaVA-1.5, LLaVA-1.6	Llama3.1	88.2	76.4	83.0	69.8	85.0

Contributions:

Decompose the Problem

Original Caption/Image Pair		
A 	B 	
Text (T)	P(Yes T, A)	P(Yes T, B)
racing it over	8.7%	24.3%

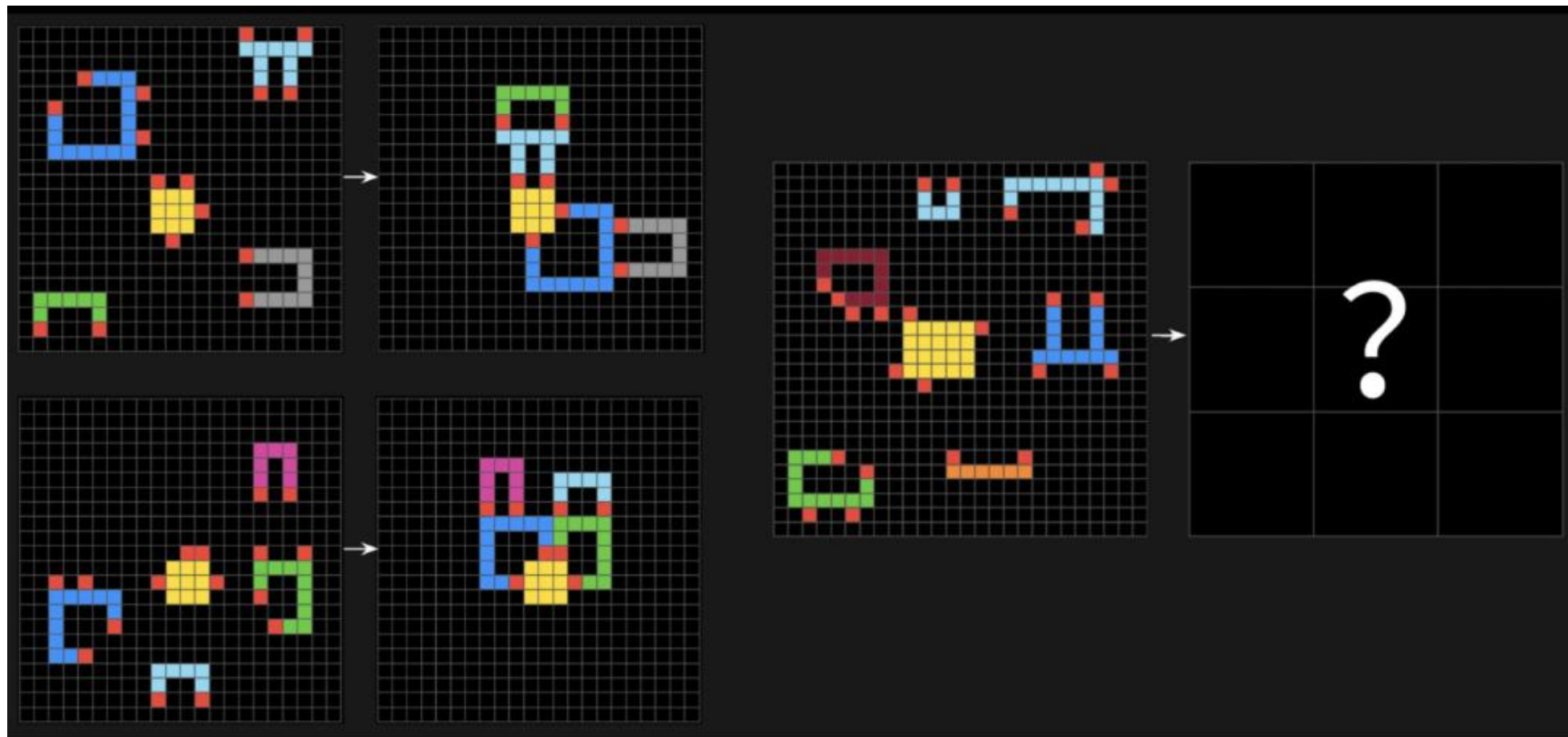
Caption Expansion via Natural Language Inference

Sentence Decomposition via Semantics (SDS)		
Text (T)	P(Yes T, A)	P(Yes T, B)
something is racing it	25.8%	45.3%
a person is racing a car	0.5%	1.3%
SDS Final Score	3.6%	7.7%

Interpretability

Caption Expansion via CECE		
Text (T)	P(Yes T, A)	P(Yes T, B)
a person is in motion	94.1%	97.1%
an object is being moved from one place to another	84.2%	58.7%
CECE Final Score	89.0%	75.5%

Compositional Reasoning – are we on the right path?



INTRODUCING ARC-AGI-2

A new benchmark that challenges frontier AI reasoning systems.

ARC-AGI-1 was created in 2019 (before LLMs even existed). It endured 5 years of global competitions, over 50,000x of AI scaling, and saw little progress until late 2024 with test-time adaptation methods pioneered by [ARC Prize 2024](#) and [OpenAI](#).

ARC-AGI-2 - the next iteration of the benchmark - is designed to stress test the **efficiency** and **capability** of state-of-the-art AI reasoning systems, provide useful signal towards AGI, and re-inspire researchers to work on new ideas.

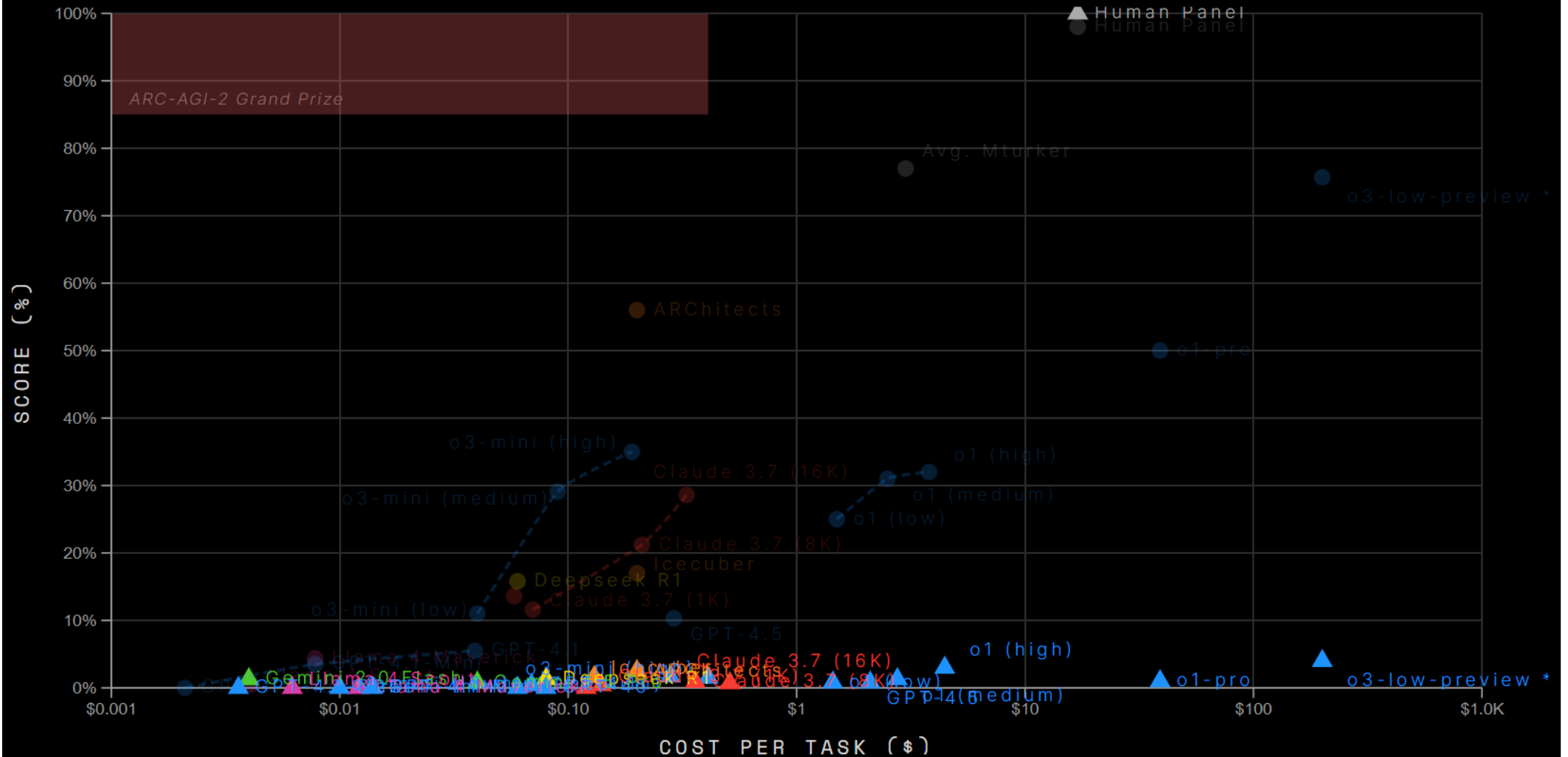
Pure LLMs score 0%, AI reasoning systems score only single-digit percentages, yet extensive testing shows that humans can solve every task.

Can you create a system that can reach 85% accuracy?

> [LEARN MORE](#)

<https://arcprize.org/>

ARC-AGI LEADERBOARD



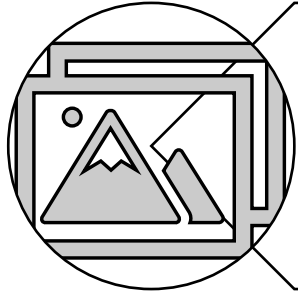
LEADERBOARD BREAKDOWN

AI System	Organization	System Type	ARC-AGI-1	ARC-AGI-2	Cost/Task	Code / Paper
Human Panel	Human	N/A	98.0%	100.0%	\$17.00	—
o3-low-preview *	OpenAI	CoT + Synthesis	75.7%	4.0%	\$200.00	
o1 (high)	OpenAI	CoT	32.0%	3.0%	\$4.45	
ARChitects	ARC Prize 2024	Custom	56.0%	2.5%	\$0.200	 
o3-mini (medium)	OpenAI	CoT	29.1%	1.7%	\$0.280	
Icecuber	ARC Prize 2024	Custom	17.0%	1.6%	\$0.130	
o3-mini (high)	OpenAI	CoT	35.0%	1.5%	\$0.410	
Gemini 2.0 Flash	Google	Base LLM	N/A	1.3%	\$0.004	
o1 (medium)	OpenAI	CoT	31.0%	1.3%	\$2.76	
Deepseek R1	Deepseek	CoT	15.8%	1.3%	\$0.080	
Gemini-2.5-Pro-Exp-03-25 **	Google	CoT	12.5%	1.3%	N/A	
o1-pro	OpenAI	CoT + Synthesis	50.0%	1.0%	\$39.00	—
Claude 3.7 (8K)	Anthropic	CoT	21.2%	0.9%	\$0.360	
Gemini 1.5 Pro	Google	Base LLM	N/A	0.8%	\$0.040	
GPT-4.5	OpenAI	Base LLM	10.3%	0.8%	\$2.10	

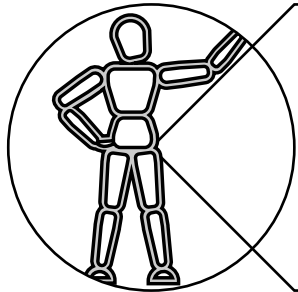
Cost: \$200
(per task)

Score: 4%

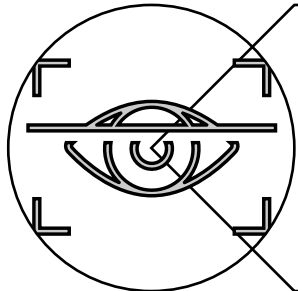
Questions?



Data **quality** and
distribution



Dynamic evaluations
and real-world
applications



Compositionality
and common-sense



generate an image representing a virtual world