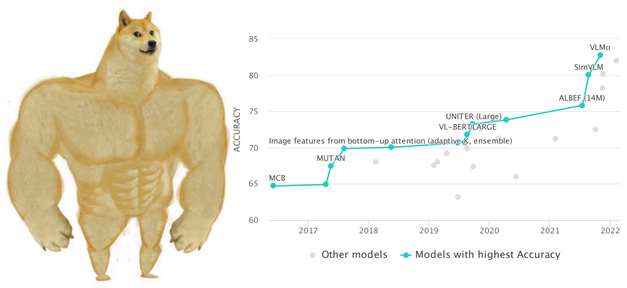
WinoGAViL: Gamified Association Benchmark to Challenge Vision-and-Language Models

Yonatan Bitton, **Nitzan Bitton Guetta**, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, Roy Schwartz

I Motivation

While vision-and-language models perform well on tasks such as visual question answering, they struggle when it comes to basic human commonsense reasoning skills.

Vision-and-language models in tasks like VQA



"What color is the banana? Yellow"

Models in tasks that require commonsense

The Curious Case of Commonsense Intelligence. Yejin Choi, 2022
Can Computers Learn Common Sense? Matthew Hutson, 2022
Why AI is harder than we think? Melanie Mitchell, 2021

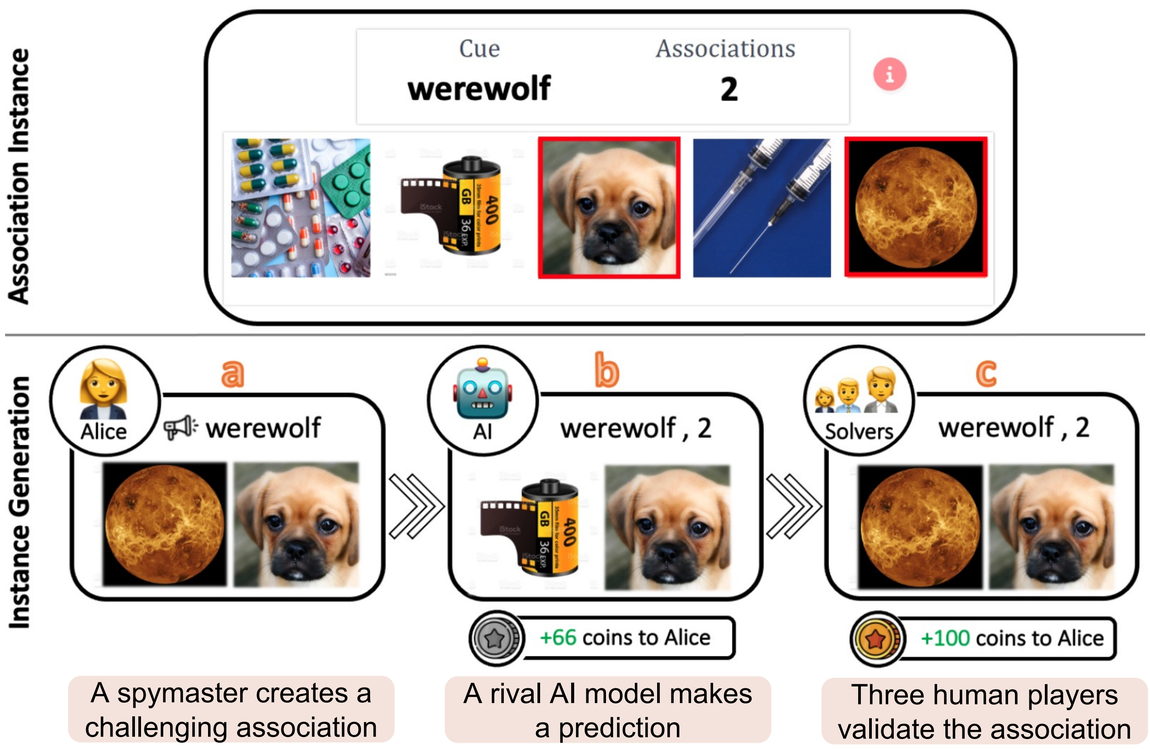
Winograd Schema Challenge (WSC)

"The **city councilmen** refused the **demonstrators** a permit because **they** feared violence."

WINOGRANDE: An adversarial WSC at Scale

Twin sentences	Options (**answer**)
The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**

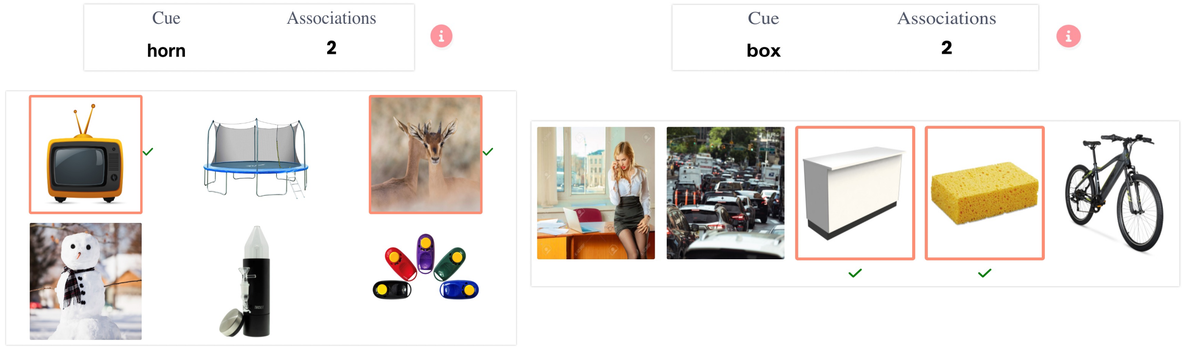
The Game



II Benchmark Analysis

Reasoning Skills

Skill	Observed Pattern	Description	Example	%
Non-Visual	Attribute	Cue has attributes of Association Cue is Association	iguana has green color miners are dirty	14%
	Use-Of	Cue uses the Association Association is used in relation to Cue	miner uses tractor tupperware is used to store food	9%
	General Knowledge	Cue is a name for Association Association is used in a relation to Cue	ford is a name of a car oats for horses increase their performance	13%
Visual	Activity	Associations perform a Cue in the image	deer & snowman looks like they stare	6%
	Analogy	Cue can be seen/used like/with Association Cue is usually related with object of another type	TV antenna looks like a horn waffle maple syrup can be dripped	4%
	Visual Similarity	Cue appears in the Association image Association is visually similar to the Cue	horns appears on the head of the deer earth is circular in the image	20%



User Feedback

Rate for the following skills how much you found them required while performing the task						
Role	Visual Reasoning	General Knowledge	Associative Thinking	Commonsense	Abstraction	Divergent Thinking
Spymaster	4.4	3.6	4.5	3.9	4.3	4.5
Solver	4.4	4	4.7	4.3	4.1	4.1
Role	Interest in play and recommend it as an online game		Level of enjoyment while doing the task		How clear was the UI	
Spymaster	3.8		3.7		4.7	
Solver	4.1		4.4		4.9	

"I used the model's guesses to make my associations better. I went after associations that the model frequently got wrong."

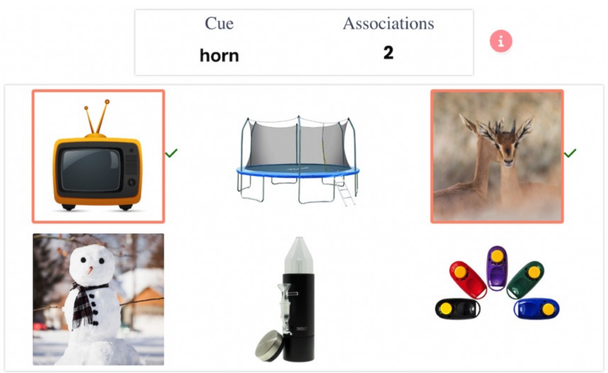
"Bonus keep motivation up when it was hard to come up with connections."

III Experiments

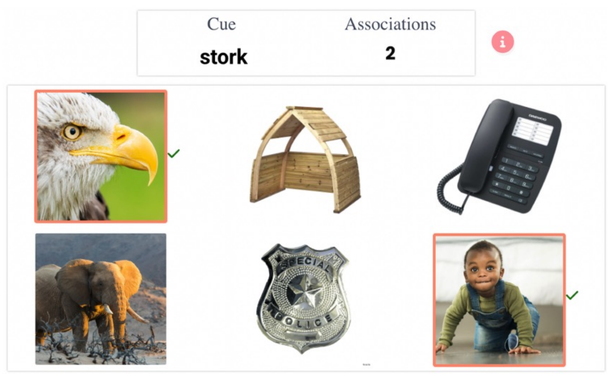
Baseline

We show the value of our gamified framework by comparing it to an alternative data generation baseline based on SWOW, an existing resource of textual associations.

WinoGAViL



SWOW



Models

- Diverse state-of-the-art vision-and-language models
- $Model(cue, image)$
- Taking k images with the top scores

Supervised

- Training is effective when the task is difficult

# Candidates	10 & 12	5 & 6
Zero-Shot	42 ± 3	53 ± 2
Supervised	49 ± 3	52 ± 1

Zero-Shot

- Easy for humans and challenging for models
- More challenging associations compared to the SWOW based method

Model	Game		SWOW
# Candidates	10 & 12	5 & 6	5 & 6
CLIP-RN50x64/14	38	50	70
CLIP-VIT-L/14	40	53	**74**
CLIP-VIT-B/32	41	53	74
CLIP-RN50	35	50	73
CLIP-ViL	15	47	66
ViLT	**52**	**55**	59
X-VLM	46	53	68
Humans	**90**	92	95

IV Model Analysis

Model performance varies between different association types

	# Items	% Model	% Humans
Visually salient	67	75	98
Visually non-salient	379	36	93
Concept related	426	65	92
Activity	24	42	96
Counting	25	36	97
Colors	14	79	96
OCR	20	50	98

Performance of textual models is close to vision-and-language models, but still far from human

Model	Game		SWOW
# Candidates	10 & 12	5 & 6	5 & 6
MPNet	39	52	72
MPNet QA	47	55	75
Distil RoBERTa	37	50	65
Humans	90	92	95