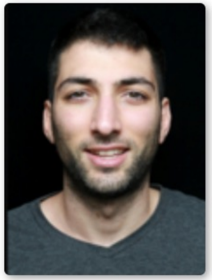
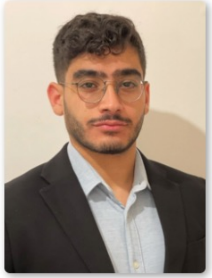
WinoGAViL: Gamified Association Benchmark to Challenge Vision-and-Language Models

YONATAN BITTON*



NITZAN GUETTA*

RON YOSEF



YUVAL ELOVICI

MOHIT BANSAL

GABRIEL STANOVSKY

ROY SCHWARTZ

NEURAL INFORMATION PROCESSING SYSTEMS

* Equal contribution

Vision-and-language models in tasks like VQA

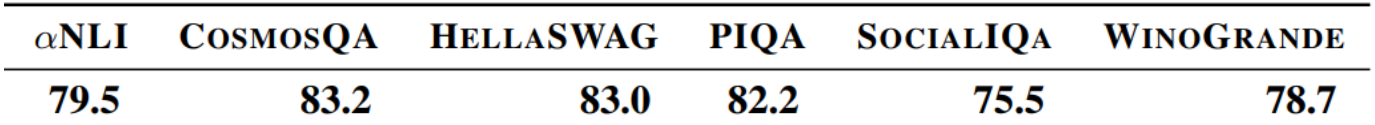
Models in tasks that require human commonsense

Commonsense Benchmarks

Winograd Schema Challenge

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

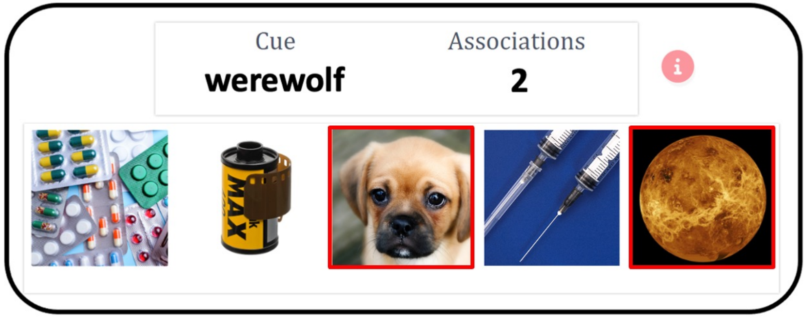
UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark



αNLI	CosmosQA	HellaSwag	PIQA	SocialIQa	WinoGrande
79.5	**83.2**	**83.0**	**82.2**	**75.5**	**78.7**

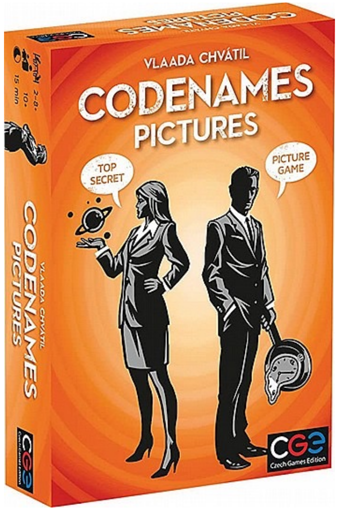
Overview

- WinoGAViL: an online game to collect vision-and-language associations
  - Dynamic benchmark
  - Less saturable
- A “spymaster” gives a textual cue related to several visual candidates, and another player has to identify them
- We use the game to collect 3.5K instances
- Finding that they are intuitive for humans (>90% Jaccard index)
  but challenging for state-of-the-art AI models (<52%)

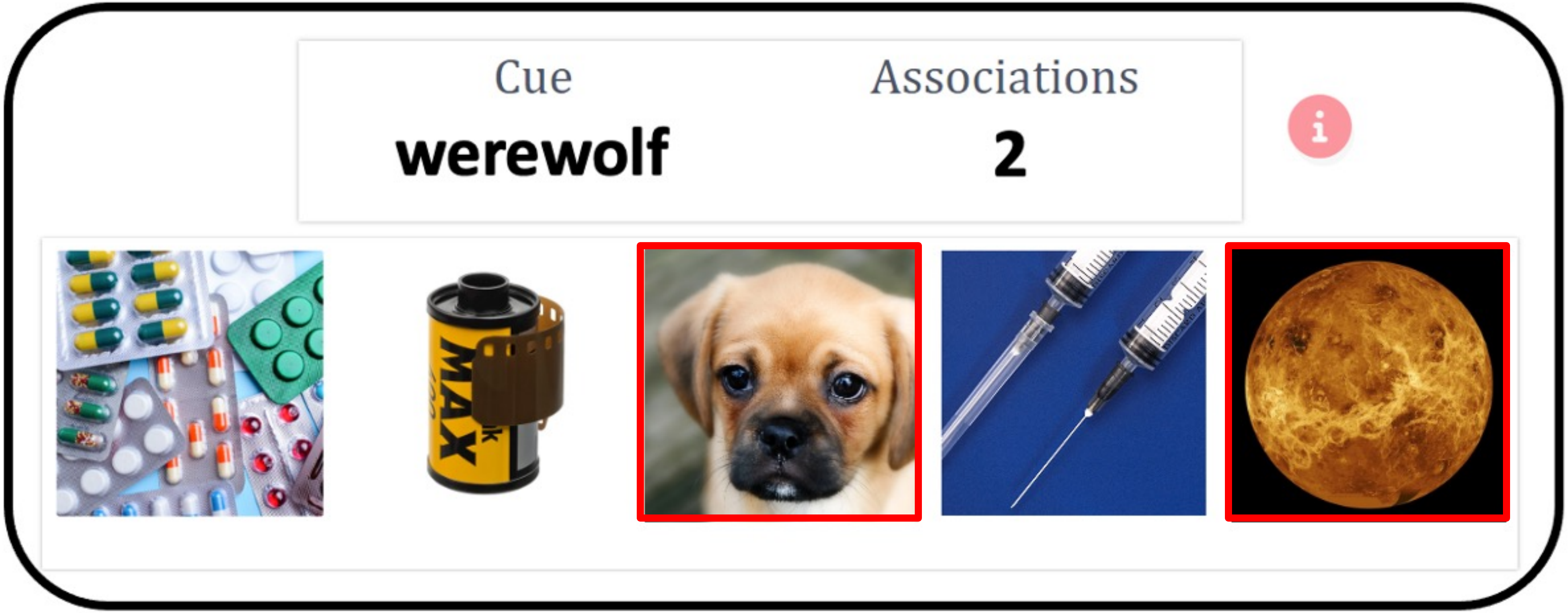


Cue: werewolf

Associations: 2

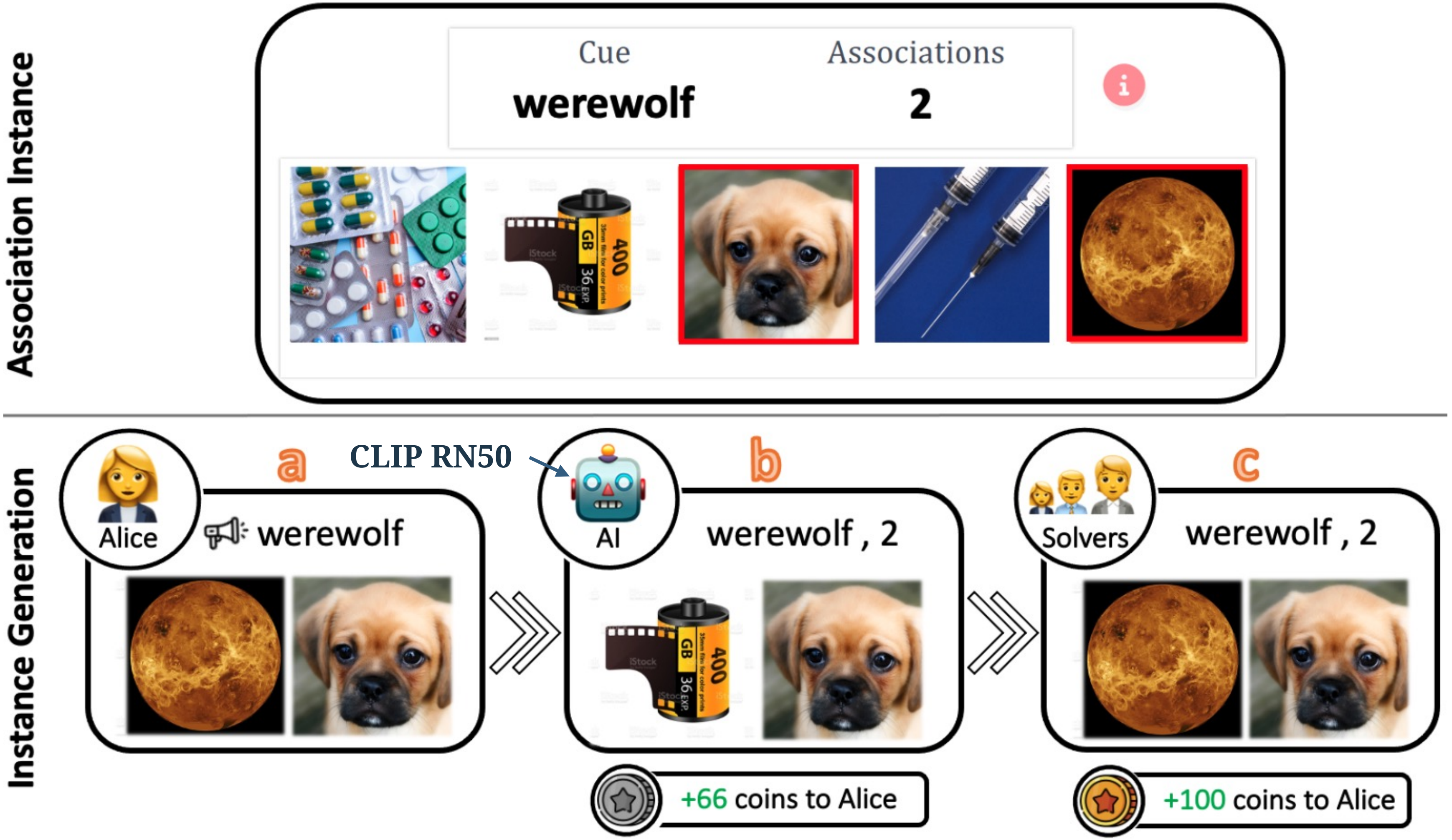


The Game



Cue	Associations
werewolf	2

The Game



Association Instance

Cue

werewolf

Associations

2

Instance Generation

a

Alice

werewolf

CLIP RN50

b

AI

werewolf , 2

+66 coins to Alice

c

Solvers

werewolf , 2

+100 coins to Alice

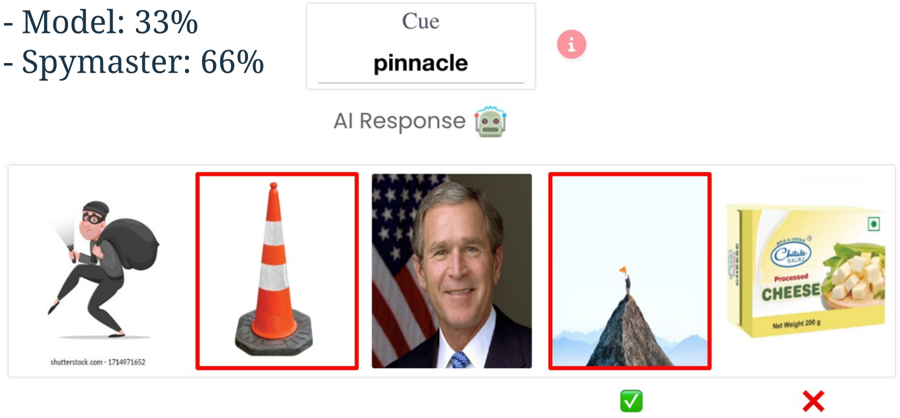
The Game

- Scoring metric: Jaccard index

1. A spymaster creates a challenging association
2. A rival AI model makes a prediction
   – fool-the-AI score
3. Three human players validate the created association
   – solvable-by-humans score
4. The spymaster becomes a solver

- Automatic validation

- Model: 33%
- Spymaster: 66%



Cue

pinnacle

Human Annotation

- Amazon Mechanical Turk (AMT)
- Total budget 2,000$ (~12-15/h)
- Qualification tests
- Bonus if “solvable-by-humans” score > 80%, which grows according to the “fool-the-AI” score (max 0.61$)

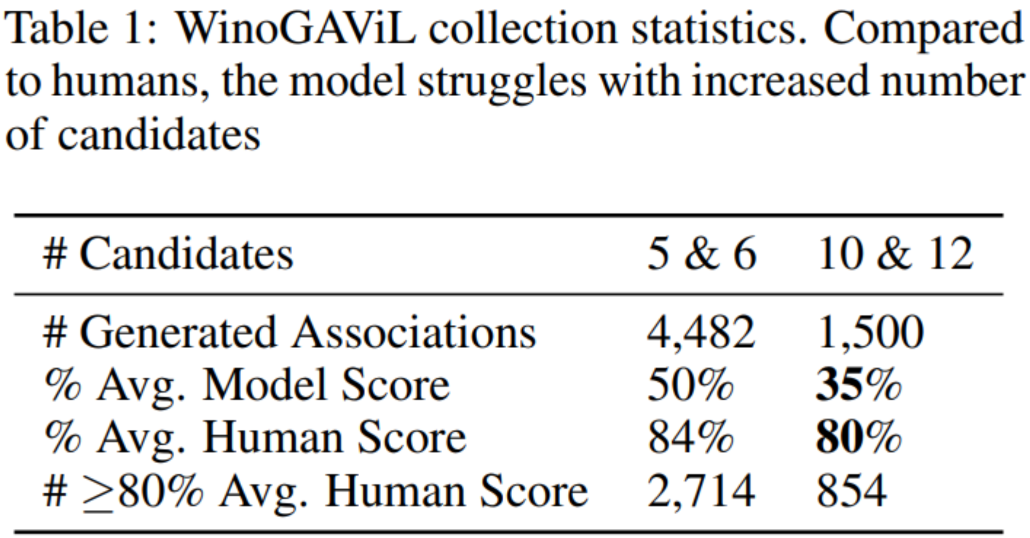
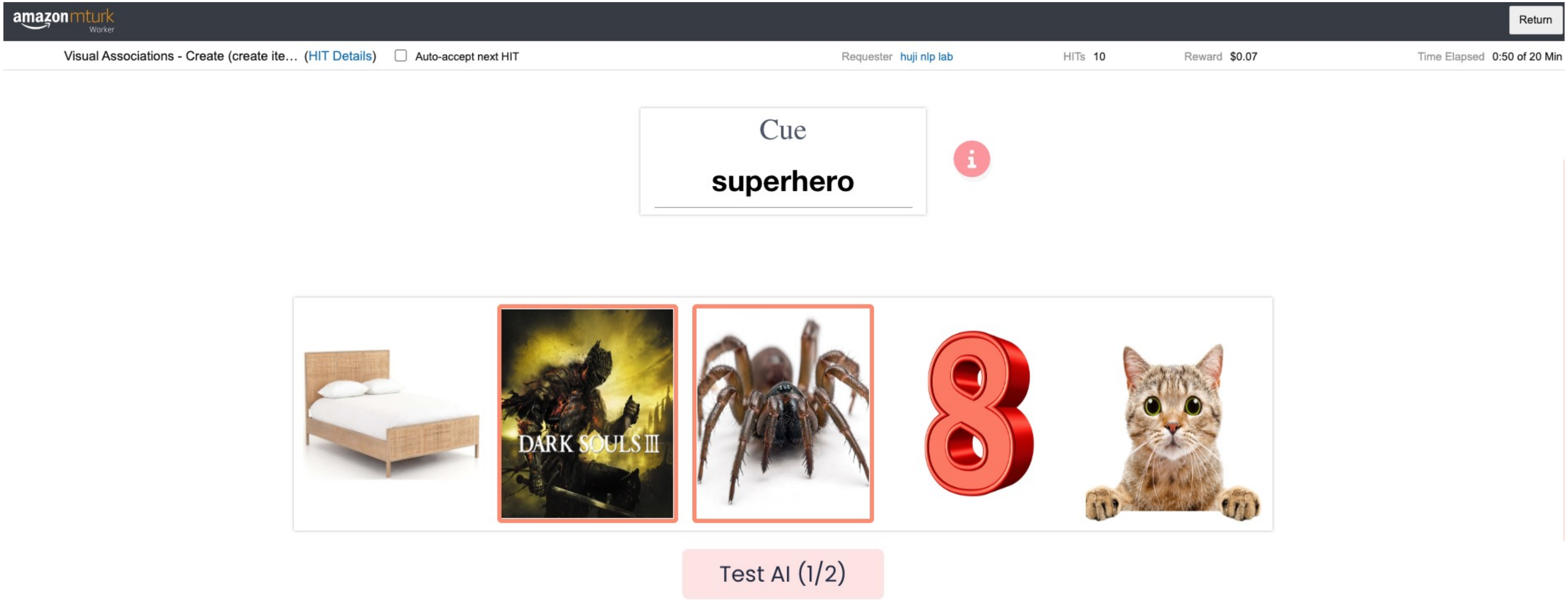


Table 1: WinoGAViL collection statistics. Compared to humans, the model struggles with increased number of candidates

# Candidates	5 & 6	10 & 12
# Generated Associations	4,482	1,500
% Avg. Model Score	50%	**35%**
% Avg. Human Score	84%	**80%**
# ≥80% Avg. Human Score	2,714	854

Human Annotation



Visual Associations - Create (create ite… (HIT Details) □ Auto-accept next HIT

Requester huji nlp lab

HITs 10

Reward $0.07

Time Elapsed 0:50 of 20 Min

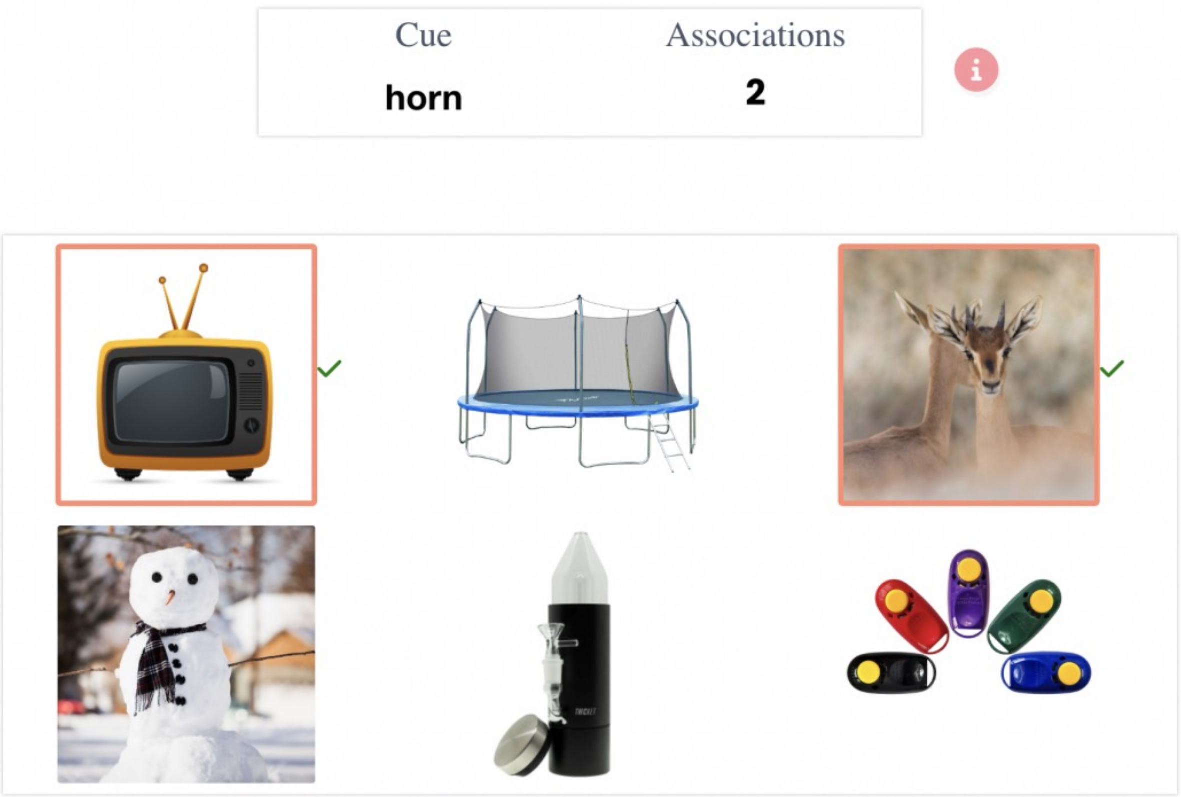
Cue

superhero

WinoGAViL Analysis

Reasoning Skills

Analogy

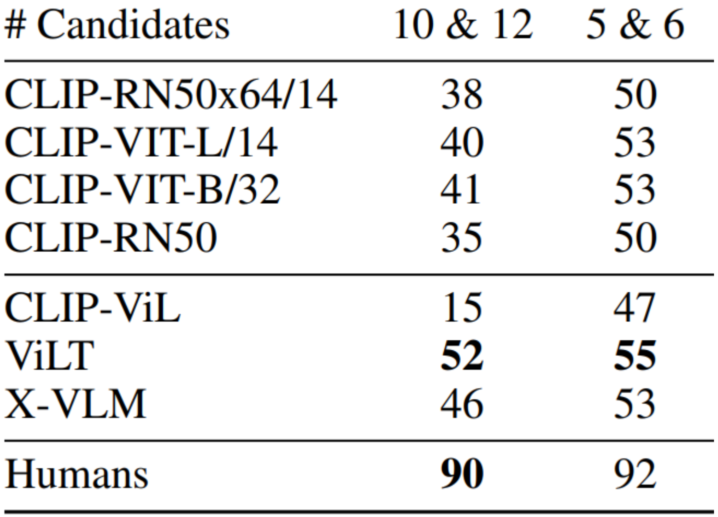


Cue	Associations
horn	2

Results

Zero-Shot

- The game allows collection of associations that are easy for humans and challenging for models



# Candidates	10 & 12	5 & 6
CLIP-RN50x64/14	38	50
CLIP-VIT-L/14	40	53
CLIP-VIT-B/32	41	53
CLIP-RN50	35	50
CLIP-ViL	15	47
ViLT	**52**	**55**
X-VLM	46	53
Humans	**90**	92

The paper contains many additional results

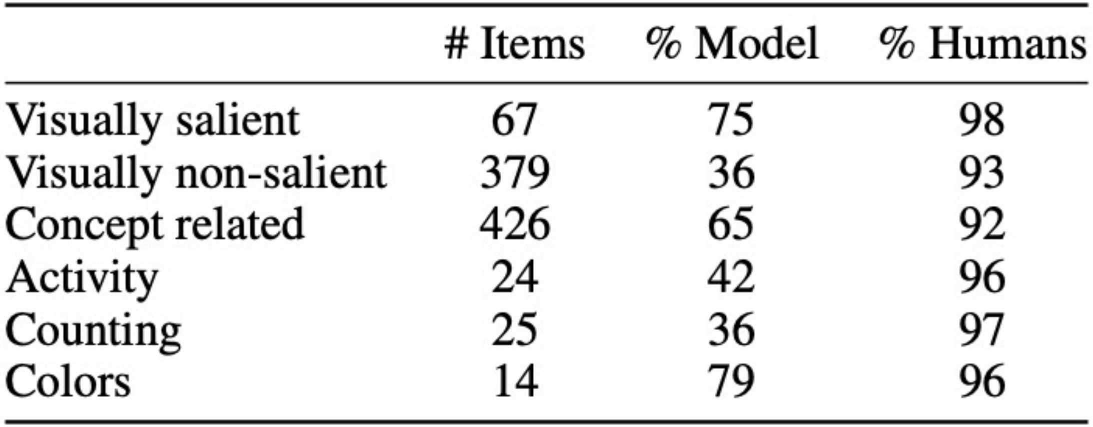
- Alternative data generation baseline
- Supervised experiments
- Textual models experiments

and more…

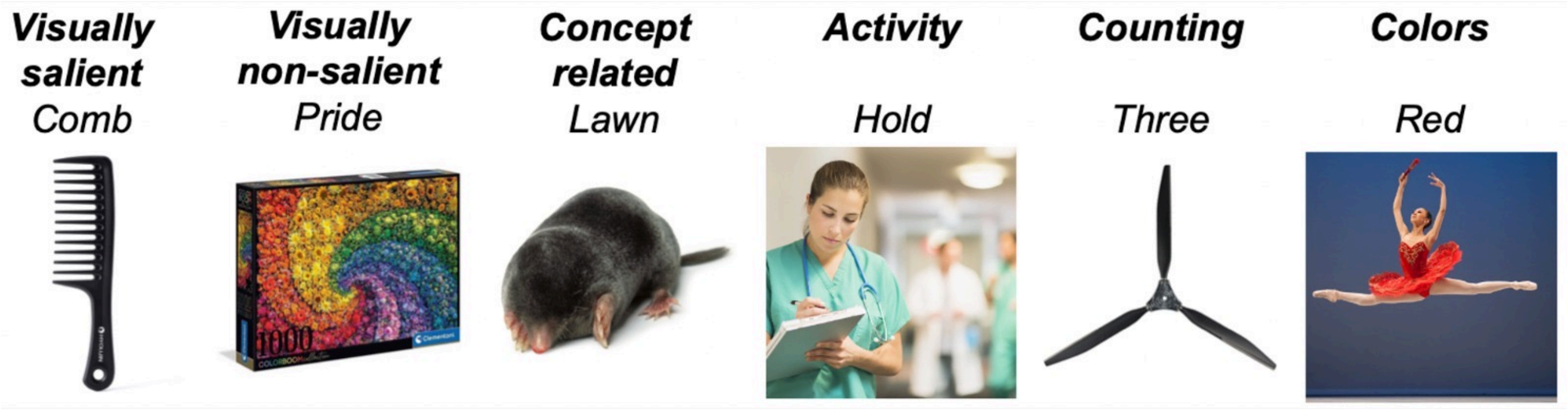
Model Analysis

Model performance varies between different association types

- Annotators classifier 1K (cue-image) pairs to the following association types

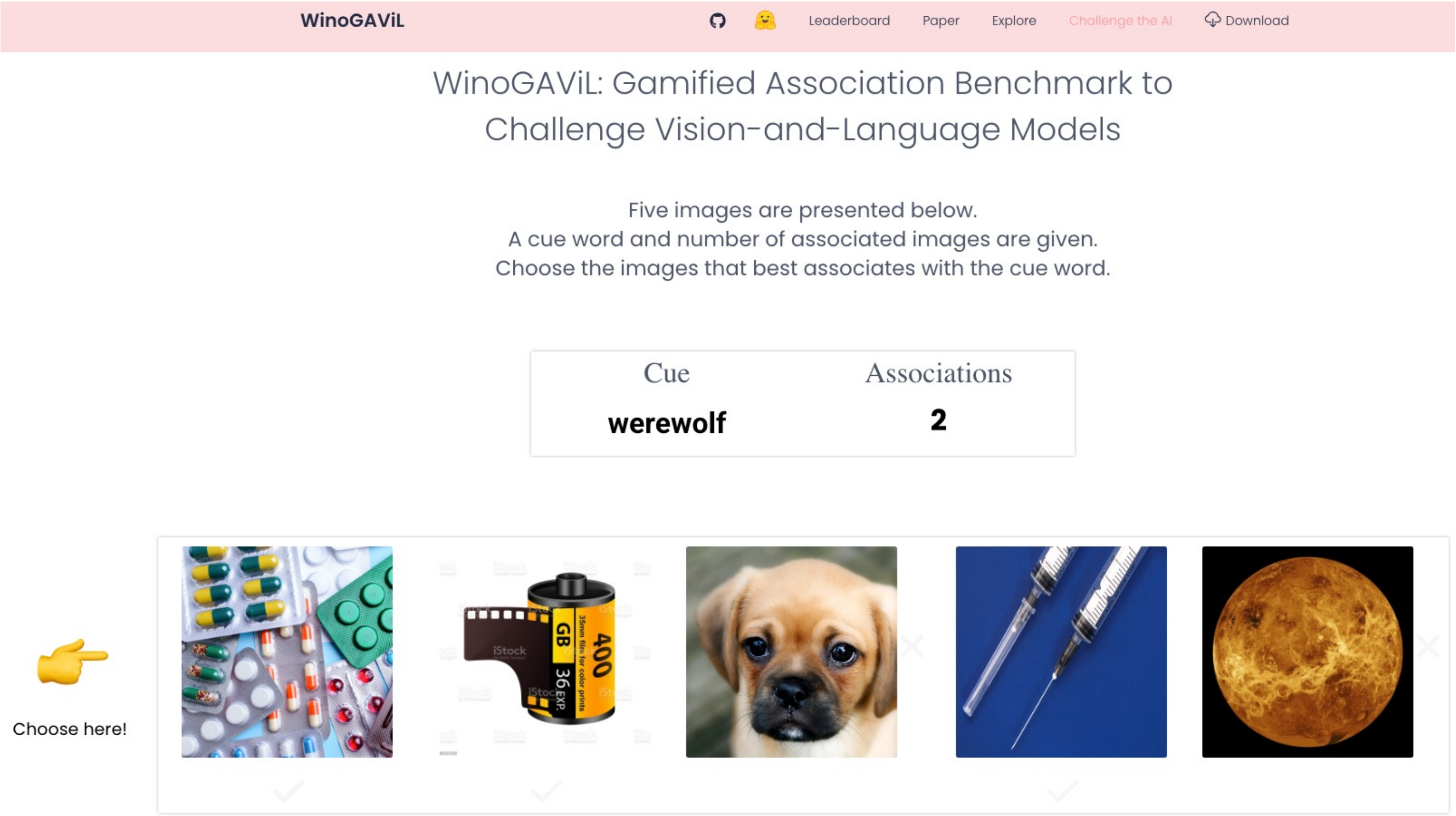


	# Items	% Model	% Humans
Visually salient	67	75	98
Visually non-salient	379	36	93
Concept related	426	65	92
Activity	24	42	96
Counting	25	36	97
Colors	14	79	96



Check out the project website!

winogavil.github.io



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

Five images are presented below.
A cue word and number of associated images are given.
Choose the images that best associates with the cue word.

Cue	Associations
werewolf	**2**

👉 Choose here!