A Appendix

A.1 Dataset Supplementary Materials

1. Dataset documentation, metadata, and download instructions: `https://winogavil.github.io/download`.
2. Intended uses: we hope our benchmark will be used by researchers to evaluate machine learning models. We hope that our benchmark will be played by users, leading to new associations collection.
3. Author statement: We bear all responsibility in case of violation of right in using our benchmark.
4. Licenses: Code is licensed under the MIT license `https://opensource.org/licenses/MIT`. Dataset is licensed under CC-BY 4.0 license `https://creativecommons.org/licenses/by/4.0/legalcode`.
5. Hosting & preservation: our website is deployed and all data is accessible and available. We encourage researchers to send us model predictions on the created test sets. We will update a model and players leader-board with this results periodically.
6. Code repository: `https://github.com/WinoGAViL/WinoGAViL-experiments`

A.2 Reasoning Skills

Table 8 lists the full reasoning and observed patterns annotated to solve the *WinoGAViL dataset* (§2.3), and Figure 6 shows an example of each visual pattern we annotated.

Table 8: Some of the observed patterns and reasoning skills required to solve WinoGAViL associations. Each association instance may require multiple skills

Skill	**Observed Pattern**	**Description**	**Example**	**%**
Non-Visual	Kind-Of	Cue is a kind of Association Association and Cue are kinds of Something	a bathtub is a shower a croissant & bread are pastries	4%
	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%
	Word Sense Meaning	Cue has word sense meaning with Association	skin ↔ object: iguanas have scales, but it is also used to measure weight	3%
			visible trail ↔ body part: comets have tail, but it is also an animal body part	3%
	Locations	The location of a Cue is Association Cue and Associations are located Somewhere	comet is in the sky polar bears live in an ice environment	5%
	Outcome	Cue is an outcome of Association Association is an outcome of Cue	oboe creates music birth & baby is the outcome of a pregnancy	6%
Visual	Activity	Associations perform a Cue in the image	deer & snowman looks like they stare (Figure 6b)	6%
	Humor/Sarcasm	Cue is related to Association in a funny way	pigpen is a dirty place, tide can make it cleaner (Figure 6e) a man that looks neglected is described as trim	1%
	Analogy	Cue can be seen/used like/with Association, although they are from a different concept map	TV antenna looks like a horn (Figure 6d)	4%
	Visual Similarity	Association is visually similar to the Cue	The sponge shape is similar to a box (Figure 6a)	20%
	Abstraction	Cue is related to Association in an abstract way	discovery is when a bulb turns on (I got it!) (Figure 6c)	5%
	Generalization	-	bread dough becomes fresh bread when baked raven is a bird that can be found in a backyard	8%

A.3 Human Annotation

Figure 7 shows an example of the Mechanical Turk user-interface. Section A.4 describe the annotator qualifications we required. Section A.4.1 describes the designed bonus reward, aiming to receive generated data that is challenging for models and easy for humans. Section A.4.2 describes the player feedback we collected. Finally, Section A.5 describes additional analysis such as players statistics and the generated textual cues analysis.

A.4 Qualifications

The basic requirements for our annotation task is percentage of approved assignments above 98%, more than 5,000 approved HITs, the location from the US, UK, Australia or New Zealand. To be a 'solver' or a 'spymaster', we required additional qualification tests: We selected 10 challenging examples from SWOW based dataset as qualification test. In each qualification test, a new worker entered demographic information: age, gender, level of education and whether he is a native English speaker. To be qualified as a 'solver', we accepted annotators that received a mean jaccard score over 80%. To be qualified as a 'creator', we require "fool-the-AI" score above 40%, and "solvable-by-humans" score above 80%. To obtain "solvable-by-humans" score, we sent the created associations to solvers (who have passed to solve qualification). The players received instructions, presented in 8 and could do an interactive practice in the project website.[11]. We do not collect or publish players personal information. We presented anonymous demographic statistics, and we do not publish the demographic information.

A.4.1 Bonus Reward

If the score is between [50,60), the bonus is 0.03$. If the score is between [60,67), the bonus is 0.07$. If the score is between [67,80), the bonus is 0.18$. Finally, if the score is at least 80, the bonus is 0.27$. The payment can thus reach up to 0.61$ for a single annotation when creating two cues for the same image instances that completely fool the AI model and are still solvable by humans.

A.4.2 Players Feedback

Here we list some of the open text feedback we received from our crowd workers. It is not cherry-picked - we chose five representative responses with positive and negative insights.

Q: Describe what did you like and dislike while performing the task.
Spymasters:

1. I liked the chance to improve my creativity and brainstorm. It was fun.
2. I liked the mental challenge, especially on the larger 10-12 ones. It was frustrating when the AI clearly guessed and got it right on the 5-6.
3. I liked that I got immediate feedback and it was something different than what I usually do on mturk. I did not like that sometimes it seemed like the objects had nothing in common and it took me too long to think of a word to try and associate the objects.
4. I liked that it was a very creative-focused task, even more so on the creator's side. It was fun to think of what I could come up with to link these words/images and fool the AI/other people.
5. Creating was exponentially harder for me than the solving. I felt frustration and I kind of felt stupid because I struggled with it. (But the solving was a blast.)

Solvers:

1. I liked how easy and straightforward they were, and that they were also super fun and different from other typical HITs I have done. The only thing I disliked was probably the pay but it was not a big deal.
2. I like the fact that I got to be creativity. Nothing to dislike about this task.
3. I liked that the correct answers were sometimes abstract and required a little thinking.
4. I liked that it was a puzzle. I really enjoy puzzles. I did not like that some of them seemed unsolvable. But all in all, I enjoyed it and did much more than I usually do.
5. I liked trying to figure out what the creator was thinking

Q: Are there additional reasoning skills you feel that were required from you?

Spymasters:

[11] `https://winogavil.github.io/beat-the-ai`

1. I find things like common sense and general knowledge mattered less for creating than when solving, because the AI was very good at cracking anything using general knowledge. You had to go more for abstract, metaphorical, or otherwise really ‘out there’ associations to get past it.

Solvers:

1. This is probably covered under “general knowledge” but I found that a lot of answers required a basic understanding of Pop Culture references.
2. Luck, of course, but also a fair bit of pop culture wisdom, which is separate from general knowledge.
3. Seeing a different perspective.

Q: Did seeing the model’s predictions affect you in any way? If so, how? (For spymasters only)

1. I was impressed at some of the AI ideas, admired the programmers and learning.
2. Yes, it helped but it was also kind of discouraging as it seemed like the AI was able to guess nearly all of my associations, which made me feel like I had even more limited options.
3. I used the model’s guesses to make my associations better. I went after associations that the model frequently got wrong.
4. Yes, it either increased my confidence or made me think harder about cues.
5. Sometimes the model was very off especially in detecting emotions.

Q: Have you been affected by the performance bonus? In what way? (For spymasters only)

1. It was nice to have a little extra pay. It helped to keep my motivation up when it was hard to come up with connections.
2. The bonus did make me sometimes give up on making a “good” cue and make a “performance” cue. Performance cue being a cue that utilizes a quirk of the AI that I know and almost guarantee that it will get wrong and will generally be easy for humans to guess. But it’s not a creative or interesting cue. Notable words are human, male and female or sometimes features like eyes, noses, ears, hands, etc.
3. Yes, it made me try harder to fool the AI.
4. The performance bonus motivated me to try harder to beat the AI, so I could justify the time investment.
5. Not really, it wasn’t enough of a bonus for me to be motivated to do more

Q: Anything else that you want to say?

1. I enjoyed this a lot and hope to participate in similar tasks for you in the near future!
2. It was fun and I hope the best for this project! If you make an online game I would 100% suggest a leaderboard for “creators” for people to create the cues. Introduce categories so people can focus on specific things. If you’re also so inclined, build something to work with Twitch.tv so streamers can play with their audience. There are some pictionary like games that do this where the streamer draws and the people in chat try to guess.
3. This would be a super interesting online if you include things like leaderboards for creators, categories, more images (although be sure to get rights to images!) and letting people rate the cues. I can definitely see game like this being popular with streamers on Twitch.tv to play with their audience (streamer https://twitch.tv/itshafu is pretty known to like games like these and sometimes streams her playing code names with other streamers) or with a group of people online.
4. This was something different to do and was fun, thank you for the opportunity. I also really appreciated how you communicated with us!
5. I liked creating, more than solving, even though I think I was a better solver than creator; I’m hoping to read the paper that results from this research.

A.5 Additional Analysis

Annotators statistics. Table 9 in Appendix A presents statistics for the Amazon Mechanical Turk workers that were involved in WinoGAViL annotation, both as spymasters and as solvers. A total of 58 crowd workers, mostly English native speakers (≥95%), of a variety of ages (26–65), genders, and levels of education (high school to graduate school). Figure 9 in Appendix A presents the spymaster's score plots, which include the number of annotations, fool-the-AI score, and solvable-by-humans score for each spymaster.

Table 9: WinoGAViL Workers Statistics

	Solvers	Creators
# Workers	41	18
# Avg. Annotations	567	332
% Avg. Performance (5-6 candidates split)	85.1	fool-the-AI: 50 solvable-by-humans: 83
Avg. Age	41 ±10	43 ±9
# High School Education	13	6
# Bachelor Education	19	11
# Master Education	8	1
% Native English Speakers	98	95

Generated cues statistics. For the final 3,568 test instances, 2,215 different cues were collected. We measure the concreteness of cue words using the concreteness dataset described [62], in which human annotated concreteness scores on a scale of 1-5 were collected. This dataset covers over 88% of the collected cues, indicating a 12% upper bound for out-of-vocabulary words. We see a diversity of both abstract and concrete generated cues in Figure 11, Appendix A. Additionally, we measure how often different annotators compose the same cues for the same group of images. Since we asked three different annotators to provide two different cues for each group of images, we have six annotations for each image group. We find that almost always (98%) they combine different cues.

A.6 Full Results

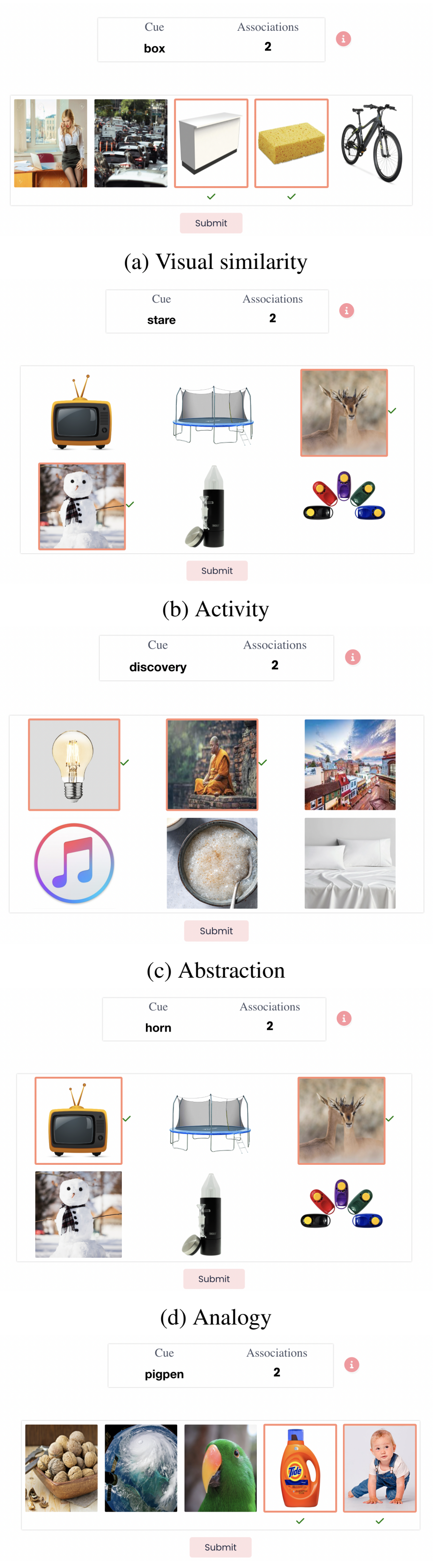
Table 10 show results for all cases of generated data, with different number of candidates and generated associations. We observe that spymasters usually selected two associations, and that performance (both human and model) are similar between 5 and 6, and between 10 and 12. When comparing human to model performance, we see that the generated data is challenging for models and easy for humans.

Table 10: *WinoGAViL dataset* Human and model (CLIP RN50) for different candidates and distractors

# Candidates	# Associations (k)	# Items	% Human Performance	% Model Performance
5	2	1,091	90	52
	3	234	92	57
6	2	1,087	90	48
	3	259	88	51
	4	43	100	57
10	2	338	87	37
	3	83	93	35
	4	5	92	29
12	2	328	90	37
	3	84	93	33
	4	16	100	28

A.7 Model Analysis on Different Association Types

A sample of 1,000 associations were annotated by three different annotators. We defined the final category as the annotators' majority vote, that was reached in 98% of the cases, and discarded the other 2%. We reported the accuracy per category, which is the proportion of the model success in each given category. The full annotation guidelines are available in the following link: https://github.com/WinoGAViL/WinoGAViL-experiments/tree/main/assets/association_types_annotations_guidelines.pdf The annotated data is available in the following link: https://github.com/WinoGAViL/WinoGAViL-experiments/blob/main/assets/model_analysis_on_different_association_types.csv We show examples of the annotated data categories in Figures 12,13,14,15, 16, 17. A screenshot from the annotation task is presented in Figure 18.



Cue	Associations
box	2

(a) Visual similarity

Cue	Associations
stare	2

(b) Activity

Cue	Associations
discovery	2

(c) Abstraction

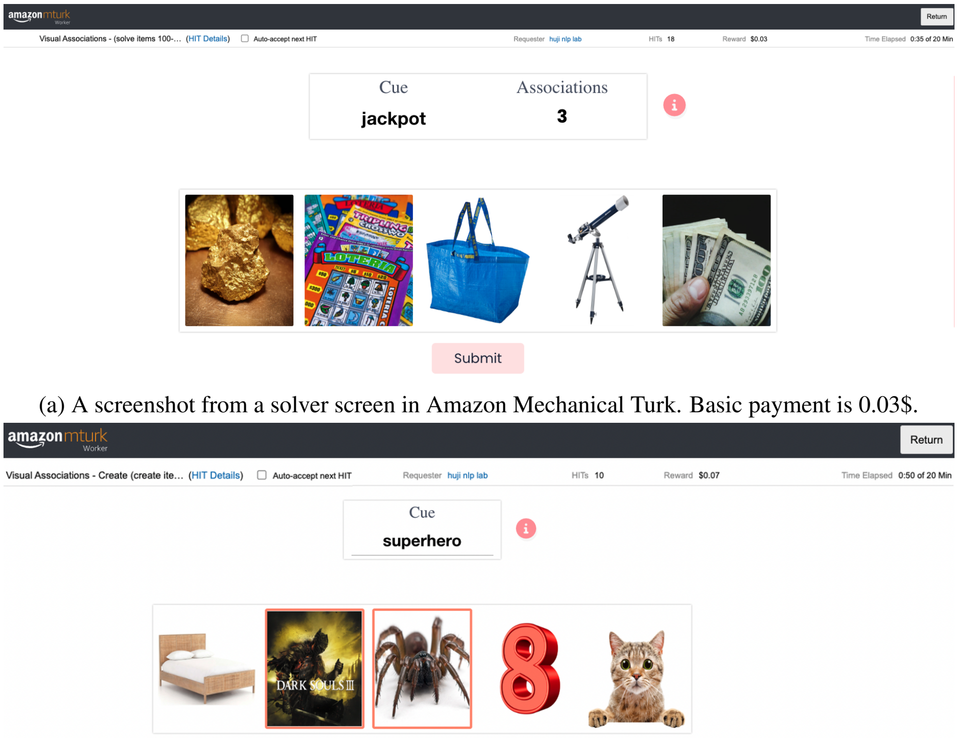
Cue	Associations
horn	2

(d) Analogy

Cue	Associations
pigpen	2

(e) Sarcasm

Figure 6: Visual Reasoning Skills Examples



Visual Associations - (solve items 100-... (HIT Details) □ Auto-accept next HIT Requester huji nlp lab HITs 18 Reward $0.03 Time Elapsed 0:35 of 20 Min

Cue: jackpot
Associations: 3

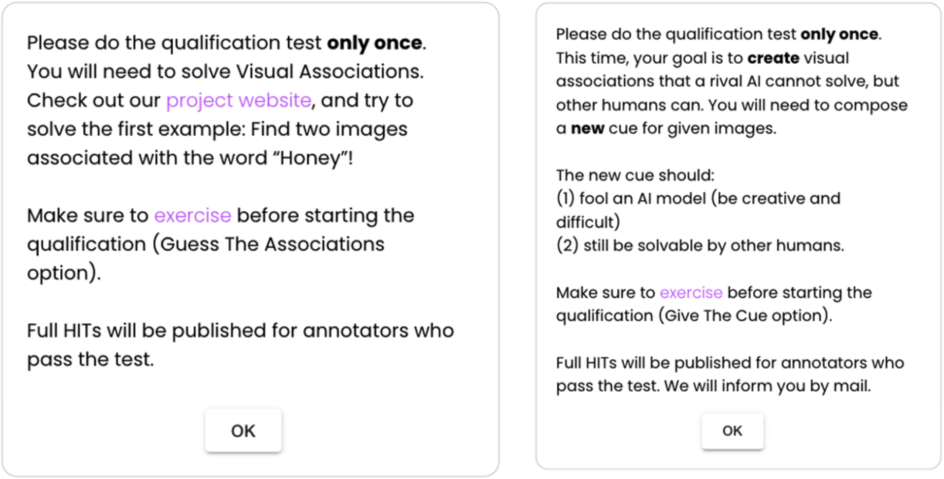
(a) A screenshot from a solver screen in Amazon Mechanical Turk. Basic payment is 0.03$.

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

(b) A screenshot from a spymaster screen in Amazon Mechanical Turk. Basic payment is 0.07$.

Figure 7: Examples of the Mechanical Turk user-interface, which referred the crowd workers to the WinoGAViL website



Please do the qualification test **only once**. You will need to solve Visual Associations. Check out our project website, and try to solve the first example: Find two images associated with the word “Honey”!

Make sure to exercise before starting the qualification (Guess The Associations option).

Full HITs will be published for annotators who pass the test.

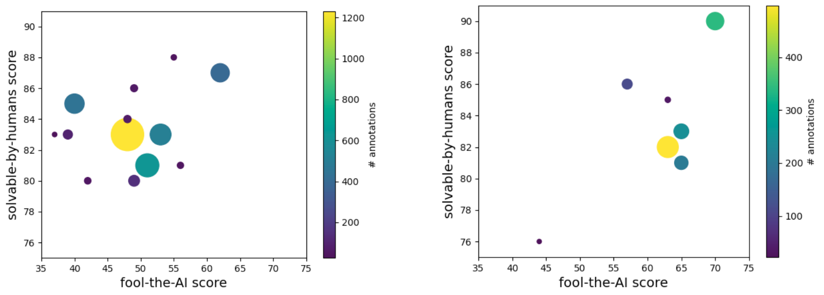
Please do the qualification test **only once**. This time, your goal is to **create** visual associations that a rival AI cannot solve, but other humans can. You will need to compose a **new** cue for given images.

The new cue should:
(1) fool an AI model (be creative and difficult)
(2) still be solvable by other humans.

Make sure to exercise before starting the qualification (Give The Cue option).

Full HITs will be published for annotators who pass the test. We will inform you by mail.

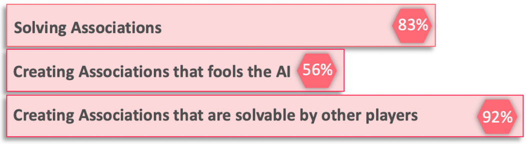
Figure 8: A screenshot of the instructions given to the annotators.



(a) 5 & 6 Candidates

(b) 10 & 12 Candidates

Figure 9: Spymasters fool-the-AI and solvable-by-human scores. Each point represents a spymaster. The best spymaster on the top right achieved fool-the-AI score of 62 and solvable-by-humans score of 87 on the case of 5 & 6 candidates; and a fool-the-AI score of 70 and solvable-by-humans score of 90 on the case of 10 & 12 candidates



Solving Associations 83%

Creating Associations that fools the AI 56%

Creating Associations that are solvable by other players 92%

Figure 10: A screenshot from the player dashboard, aiming to increase players motivation. It contains different statistics measuring the performance in beating the AI, creating novel associations, and solving other player's associations.

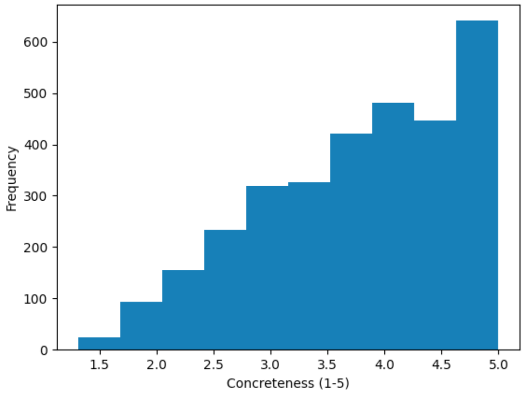
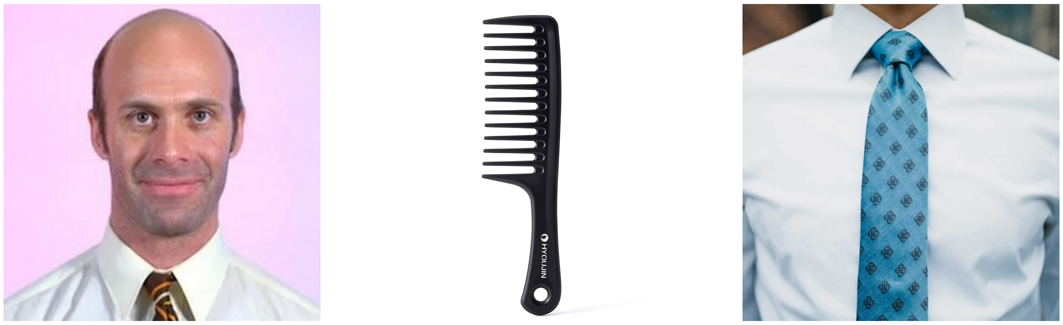
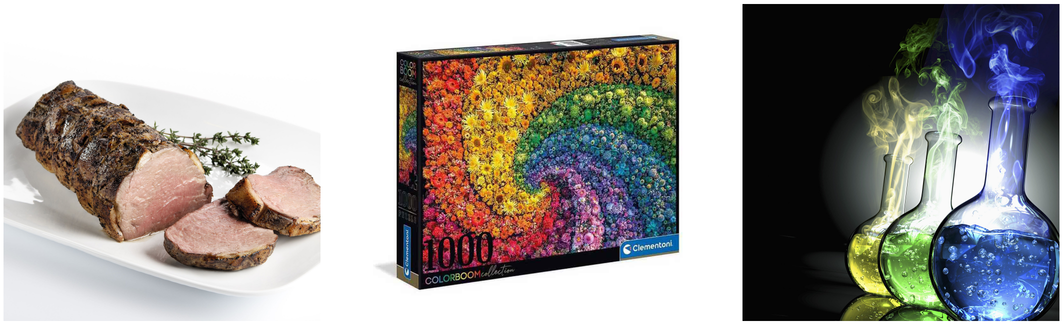


Figure 11: Generated cues concreteness distribution.



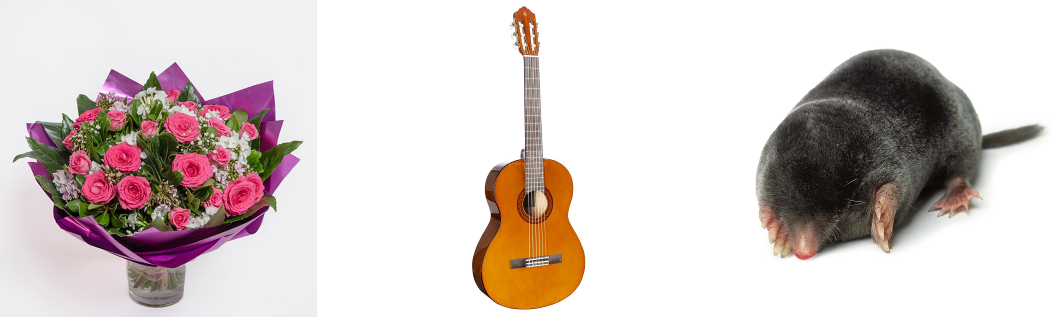
(a) human (b) comb (c) tie

Figure 12: Visually salient



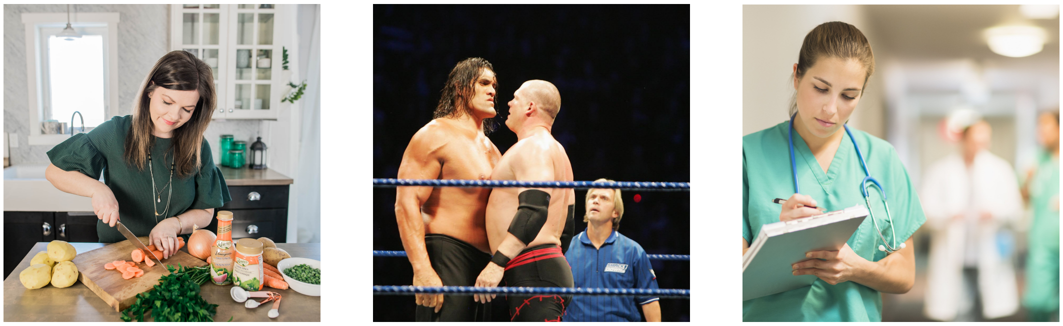
(a) log (b) pride (c) vapors

Figure 13: Visually non-salient



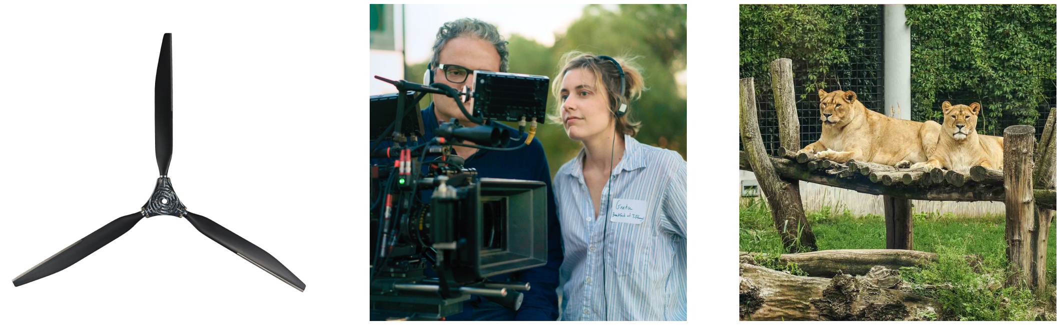
(a) pollinate (b) loud (c) lawn

Figure 14: Concept related



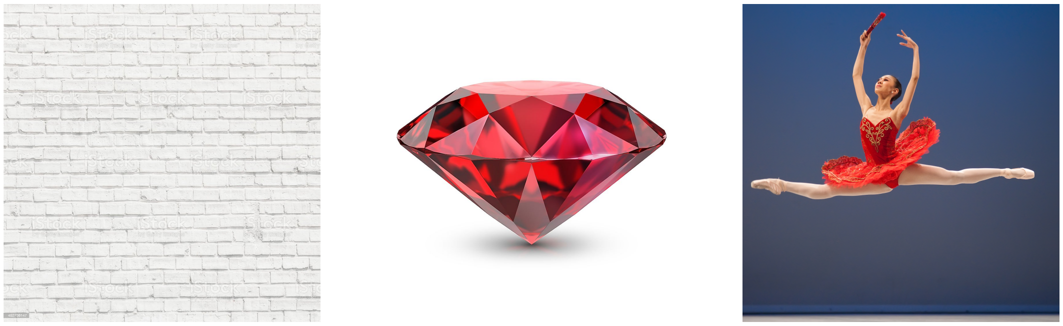
(a) cook (b) confront (c) hold

Figure 15: Activity



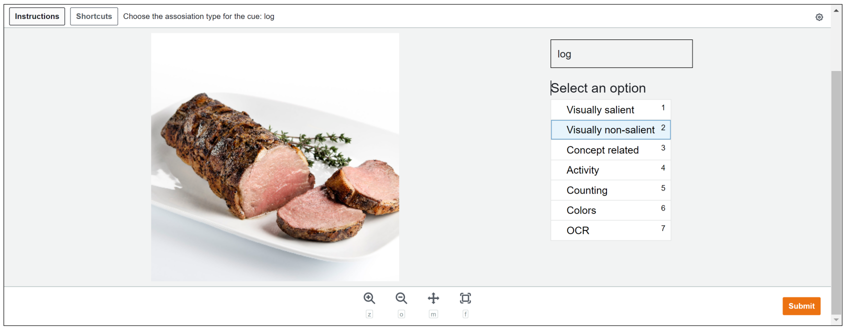
(a) three (b) two (c) multiple

Figure 16: Aggregation / Counting



(a) white (b) sanguine (c) red

Figure 17: Colors



Choose the assosiation type for the cue: log

log

Figure 18: A screenshot from the task of annotating different analogy types.

A.8 Multimodal Evaluation

The *SWOW vision baseline dataset* has four options of text-image modalities, so we evaluate all cases of models: vision-and-language, textual only and visual only.

Computer vision models when both the cue and candidates are visual we evaluate ViT [63], Swin Transformer [64], DeiT [65] and ConvNeXt [66].[12]

Visual associations are more difficult than textual Table 11 shows results for the different modalities. The performance is the highest in the all-text version, decreases when one of the cues or candidates are images, and the worst when both are images.

[12] The exact versions we took are the largest pretrained versions available in timm library: ViT Large patch32-384, Swin Large patch4 window7-224, DeiT Base patch16 384, ConvNeXt Large.

Table 11: Results on the multi-modal versions of SWOW baseline dataset. Visual associations are more difficult than textual

Model type	Model	Modalities		Jaccard Index
		Key	Candidates	
Vision and Language	CLIP-ViT-L/14	Text	Text	86
			Image	74
		Image	Text	79
			Image	65
	ViLT	Text	Image	58
		Image	Text	59
	LiT	Text	Image	37
		Image	Text	40
	X-VLM	Text	Image	68
		Image	Text	70
Vision	ViT	Image		61
	Swin			59
	DeiT			53
	ConvNeXt			56
Text Transformers	MPNet	Text		88
	MPNet QA			91
	Distil RoBERTa			77
Text Word2Vec	Spacy	Text		91
Text	CLIP-ViT-L/14	Text		87
	MPNet			88
	MPNet QA			90
	Distil RoBERTa			73
	CLIP-ViT-L/14	Text	Synthesized Text	55
	MPNet			72
	MPNet QA			76
	Distil RoBERTa			66
	CLIP-ViT-L/14	Synthesized Text	Text	81
	MPNet			77
	MPNet QA			78
	Distil RoBERTa			73
	CLIP-ViT-L/14	Synthesized Text		61
	MPNet			64
	MPNet QA			64
	Distil RoBERTa			67