WINOVIZ: Probing Visual Properties of Objects Under Different States

Woojeong Jin, Tejas Srinivasan, Jesse Thomason, Xiang Ren
Department of Computer Science, University of Southern California, USA
{woojeong.jin,tejas.srinivasan,jessetho,xiangren}@usc.edu

Abstract

Humans interpret visual aspects of objects based on contexts. For example, a banana appears brown when rotten and green when unripe. Previous studies focused on language models' grasp of typical object properties. We introduce WINOVIZ, a text-only dataset with 1,380 examples of probing language models' reasoning about diverse visual properties under different contexts. Our task demands pragmatic and visual knowledge reasoning. We also present multi-hop data, a more challenging version requiring multi-step reasoning chains. Experimental findings include: a) GPT-4 excels overall but struggles with multi-hop data. b) Large models perform well in pragmatic reasoning but struggle with visual knowledge reasoning. c) Vision-language models outperform language-only models.

1 Introduction

Language models (LMs) face challenges in developing intuitive reasoning and acquiring knowledge from experience, similar to humans. Human knowledge acquisition from the visual world is effortless but poses difficulties for LMs, as such knowledge is often not explicitly described in text. Overcoming these challenges requires visual grounding, connecting language and visual information for comprehension.

Previous studies have predominantly aimed at investigating language models in relation to object prototypical visual properties such as color, shape, and affordance, and transferring such knowledge from vision-language models (Norlund et al., 2021; Paik et al., 2021; Zhang et al., 2022; Li et al., 2023b). In this work, we study language models' reasoning ability on associations between objects and their visual properties across different object states. The task requires a model to reason about different states of an object where the object may exhibit different properties.

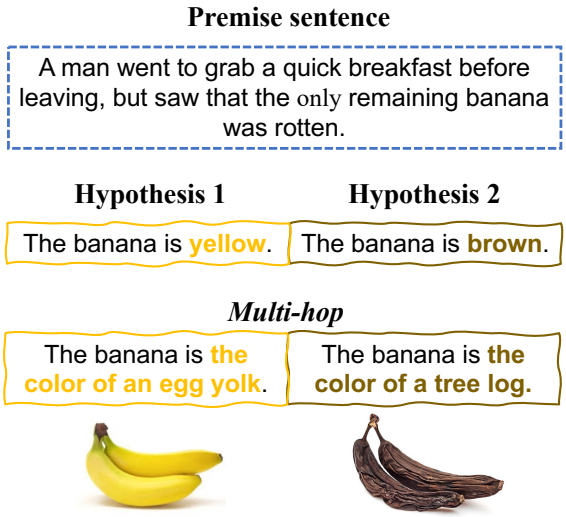


Figure 1: **The WINOVIZ task.** We investigate the divergent properties of an object and explore the reasoning abilities of language models pertaining to object attributes.

In this work, we investigate the divergent properties of an object and explore the reasoning abilities of language models pertaining to object attributes. Annotators create a premise sentence portraying a scene with a banana and two hypothesis sentences highlighting its visual properties as depicted in Fig. 1. The goal is to choose a more plausible hypothesis, requiring comprehension of the banana's properties in different states. A more challenging multi-hop version replaces the visual attribute word with another object word sharing a similar visual attribute.

Benchmarking zero-/few-shot performance includes text-only models like BERT (Kenton and Toutanova, 2019), T5 (Raffel et al., 2020; Chung et al., 2022), and GPT variants (Brown et al., 2020), ranging from 110 million to 175 billion parameters. Models incorporating visual information, such as VL-BERT (Su et al., 2019) and Oscar (Li et al., 2020), are explored.

Proceedings of the Fifth Workshop on Insights from Negative Results in NLP, pages 110–123

Key findings from experiments with the WINOVIZ benchmark include: a) GPT-4 performs effectively but degrades on multi-hop data. b) Large models excel in pragmatic reasoning but face challenges in visual knowledge reasoning. c) Vision-language models outperform language models.

2 The WINOVIZ Task

The WINOVIZ task entails the need for a model to deduce whether objects can demonstrate prototypical behaviors in various scenarios. More precisely, when provided with a natural language sentence describing an object engaged in a particular behavior (*premise sentence*), the model must determine between two sentences presenting contrasting visual attributes of the object (*hypothesis sentences*). Fig. 2 includes dataset collection (details are in the appendix)

Challenges. The WINOVIZ task assesses a machine's reasoning ability about daily objects, focusing on their varied properties. Models often struggle with visual knowledge related to common objects due to limited explicit details in training text, attributed to reporting bias (Norlund et al., 2021; Jin et al., 2022). The task is challenging as it requires pragmatic reasoning and visual knowledge reasoning, involving finding intended meanings in the text and reasoning about object properties. A more challenging version, *multi-hop data*, requires multi-step reasoning chains.

3 Experiments

We first describe the experimental setup used in our analysis and share experimental results.

Language Models. We experiment with 7 language models in total (Table 5). We include encoder-only, encoder-decoder, decoder-only models. The sizes of LMs vary from 109M to 175B. We include large LMs, GPT-3, GPT-3.5, and GPT-4 (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023).

Vision-language Models. We experiment with a total of 5 vision-language models (see Table 5). Our task involves understanding visual information about objects in various states, derived from image-caption datasets. We investigate whether vision-language models surpass language models in our task. For evaluation, we deliberately exclude

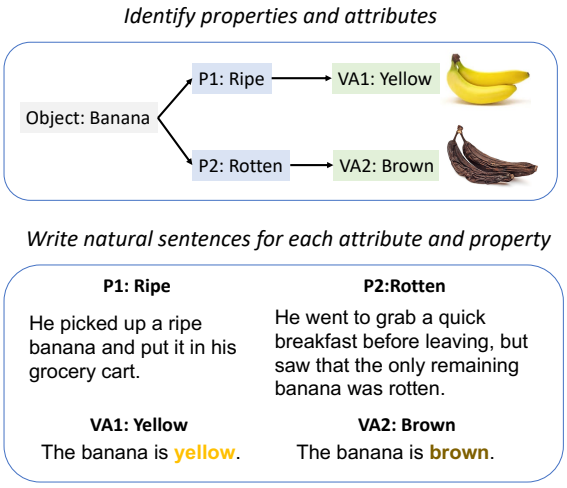


Figure 2: **Dataset Collection.** We collect our data through crowdsourcing efforts. The first step is to identify properties and visual attributes for an object and the second step is to write natural sentences for each property and attribute. Sentences with properties will be used as premise sentences and sentences with visual attributes will be used as hypothesis sentences.

image inputs and focus solely on the language components of the models, using encoder-only models (VL-BERT (Su et al., 2019) and Oscar (Li et al., 2020)), a decoder-only model (LLaVA-v1.5 (Liu et al., 2023)), and a bi-encoder model (CLIP 'clip-vit-large-patch14'(Radford et al., 2021)).

Inference. In our analysis, we rely on zero-shot inference and few-shot in-context learning for encoder-decoder, decoder-only models. Our prompt design for the zero-shot inference is as follows: *"You will be given a sentence, and two options. Output either Option 1 or Option 2, depending on which option is more likely to be true given the sentence."* For the few-shot in-context learning, we use 4 examples. We also adopt chain-of-thought prompting (Wei et al., 2022) for the few-shot inference. In addition to the encoder-decoder and decoder-only models, we explore encoder-only models. Encoder-only models cannot do zero-shot inference for multi-choice tasks since it requires a task-specific head for unseen tasks. Thus, we fine-tune the encoder-only models with SNLI (Bowman et al., 2015) and ANLI (Nie et al., 2019) datasets and we use only 'contradiction' and 'entailment' labels in fine-tuning.

Evaluation Setup. We evaluate models with two different metrics: individual accuracy (Ind.) and pair accuracy (Pair). Individual accuracy refers to accuracy on each individual question, while pair

Model	Single-hop		Multi-hop	
	Ind.	Pair	Ind.	Pair
FLAN-T5-XXL	86.24	72.71	68.09	40.43
LLaMA2	73.28	48.85	52.84	20.45
LLaVA	79.47	59.63	56.82	17.05
GPT-3	84.17	69.24	58.5	22
GPT-3.5	86.58	75.62	58	20
GPT-4	90.25	81.19	72	45

Table 1: **Results on WINOVIZ in a zero-shot manner.** We evaluate large models using 0 examples on both our single-hop and multi-hop datasets. We observe that these models performed well on the single-hop data; however, their performance is significantly degraded on the multi-hop data.

accuracy refers to the accuracy on each pair of questions. In WINOVIZ, two premise sentences are paired and they share the same set of hypothesis options. We measure the model's performance based on its ability to accurately predict both premise sentences. If the model's prediction is correct for only one of the premise sentences in the pair, we consider the prediction less robust.

3.1 Analysis Questions

In our empirical analysis, we try to answer the following questions:

1. How good are large models on our task? When it comes to multi-hop data, how good are they? (Section 3.2)

2. Do few-shot prompting and CoT prompting improve the results? (Section 3.3)

3. Which reasoning step between pragmatic reasoning and visual knowledge reasoning is main bottleneck in our task? (Section 3.5)

4. Do vision-language models outperform language-model counterparts? (Section 3.2)

3.2 Zero-shot Results

We evaluate language models and vision-language models in a zero-shot way, without utilizing any training data (Table 1). Overall, large models perform well on the single-hop data, but their performance is significantly degraded on the multi-hop data. Among them, GPT-4 exhibits the best overall performance on both single-hop and multi-hop tasks. Surprisingly, FLAN-T5-XXL, the smallest model among the comparison, yields comparable results to larger models, including GPT-3. Moreover, it outperforms GPT-3 and GPT-3.5 on the multi-hop dataset. LLaVA, built upon LLaMA2 and trained with image-caption datasets, shows noteworthy performance. As indicated in the table, LLaVA surpasses LLaMA2 on both single-hop and multi-hop data, suggesting that image-caption datasets enhance reasoning in our task.

Model	Single-hop		Multi-hop	
	Ind.	Pair	Ind.	Pair
FLAN-T5 (0)	86.35	73.17	68.09	40.43
FLAN-T5 (4)	87.84	76.15	69.32	42.05
FLAN-T5 (4 CoT)	87.16	74.77	67.05	38.64
GPT-3.5 (0)	86.58	75.62	58	20
GPT-3.5 (4)	88.42	77.75	62.5	28.41
GPT-3.5 (4 CoT)	77.18	59.63	65.34	34.09

Table 2: **Results on WINOVIZ with 4-shot in-context learning.** We use FLAN-T5-XXL and GPT-3.5 in this analysis. Standard prompting marginally improves the performance of them, while chain-of-thought prompting is beneficial for GPT-3.5 in the multi-hop task.

Method	Single-hop		Multi-hop	
	Ind.	Pair	Ind.	Pair
BERT-Large	67.31	39.44	54	16
VL-BERT-Large	69.61	42.88	56	18
Oscar-Large	72.93	50.22	64.5	32

Table 3: **Results on WINOVIZ after NLI training.** We train encoder-only models on NLI datasets and choose an option by the highest probability of the 'entailment' class.

3.3 Few-shot Results

Table 2 displays the results with 4 in-context examples for FLAN-T5-XXL and GPT-3.5. We conduct tests using standard prompting and chain-of-thought prompting in this experiment. Initially, standard prompting with 4 in-context examples marginally improves the performance of FLAN-T5 and GPT-3.5 on both single-hop and multi-hop tasks. It's surprising that chain-of-thought prompting appears to negatively impact the performance of GPT-3.5. However, it proves beneficial for GPT-3.5. in the multi-hop task. We speculate that the effectiveness of chain-of-thought prompting increases when the task is more challenging.

Model	Pragmatic	Visual	Combined
FLAN-T5-XXL	93.04	82.91	79.75
LLaMA2	86.71	70.25	69.62
LLaVA	92.41	74.05	73.25
GPT-3.5	91.14	82.28	79.75
GPT-4	95.57	88.61	85.44

Table 4: **Results on pragmatic reasoning, visual knowledge reasoning, and our original data (combined).** We study different types of reasoning in our data. We report individual accuracy.

3.4 Results of Encoder-only Models

Encoder-only models cannot be applied to our task without fine-tuning. Thus, we fine-tune the encoder-only models on natural language inference datasets instead. By doing this, our task is framed into the NLI setup and choose an option by the highest probability of the 'entailment' class. We fine-tune the encoder-only models with SNLI (Bowman et al., 2015) and ANLI (Nie et al., 2019) datasets and we use only 'contradiction' and 'entailment' labels. Table 3 shows the results of encoder-only models. VL-BERT and Oscar are BERT-based vision-language models, and they are trained on image-caption datasets. In our experiments, we observe that the vision-language models consistently surpass the BERT model on our dataset.

3.5 Pragmatic and Visual Knowledge Reasoning

We investigate whether models genuinely understand visual knowledge for our task. Our task requires pragmatic reasoning and visual knowledge reasoning. We decouple our task into pragmatic reasoning and visual knowledge reasoning and analyze which step is a bottleneck. Table 4 shows the results on pragmatic reasoning (pragmatic), visual knowledge reasoning (visual), and our original data (combined), utilizing the same subset. Firstly, results on pragmatic reasoning are better than others, suggesting that large models do well on pragmatic reasoning. For example, GPT-4 achieves 95.57% on pragmatic reasoning. Main bottleneck in our task is on visual knowledge reasoning; results on visual knowledge reasoning are lower than those on pragmatic reasoning. When comparing LLaMA2 and LLaVA, LLaVA demonstrates superior abilities in both pragmatic reasoning and visual knowledge reasoning. Interestingly, FLAN-T5-XXL performs comparably to a proprietary model, GPT-3.5, in terms of pragmatic reasoning and visual reasoning.

4 Conclusion

Examining real-world object properties requires a visual understanding that language models lack. In our study, we introduced a text-only WINOVIZ focused on question-answering tasks, comprising 1,380 examples exploring language models' reasoning capabilities across various visual properties of objects in diverse contexts. Our findings revealed that large language models demonstrate effective performance overall but struggle particularly with the multi-hop version of our dataset. It became apparent that the bottleneck in our task lies in the reasoning of visual knowledge. Vision-language models surpass their language-only counterparts, although image-generation approaches prove ineffective for our specific task. Future endeavors will delve into how to efficiently transfer visual knowledge from images or captions.

5 Limitations

Our work is focused on a specific subset of language models and vision-language models. We adopt vision-language models in which the language backbones are pre-trained using image-caption datasets. Additionally, we employ Stable Diffusion for image generation, although the current output may not directly benefit our task. Utilizing state-of-the-art diffusion models could enhance image quality, yet the challenge of generating images useful for our task persists. Moreover, our observations indicate that large language models excel in our single-hop task, achieving up to 90% accuracy. This suggests that these large models can effectively reason over visual knowledge even in the absence of explicit visual signals. Nonetheless, how visual signals can be harnessed to enhance language models is underexplored, and we defer it to future research endeavors.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches: IFIP WG 8.2, Working Conference, Tampa, FL, USA, December 13-14, 2012. Proceedings*, pages 210–221. Springer.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Yuling Gu, Bhavana Dalvi Mishra, and Peter Clark. 2022. Do language models have coherent mental models of everyday things? *arXiv preprint arXiv:2212.10029*.

Lovisa Hagström and Richard Johansson. 2022. What do models learn from training on more than text? measuring visual commonsense knowledge. *arXiv preprint arXiv:2205.07065*.

Woojeong Jin, Dong-Ho Lee, Chenguang Zhu, Jay Pujara, and Xiang Ren. 2022. Leveraging visual knowledge in language tasks: An empirical study on intermediate pre-training for cross-modal knowledge transfer. *arXiv preprint arXiv:2203.07519*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Qi Liu, Lingpeng Kong, and Xu Sun. 2023b. Can language models understand physical concepts? *arXiv preprint arXiv:2305.14057*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. *arXiv preprint arXiv:2203.08075*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Tobias Norlund, Lovisa Hagström, and Richard Johansson. 2021. Transferring knowledge from vision to language: How to achieve it and how to measure it? *arXiv preprint arXiv:2109.11321*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The world of an octopus: How reporting bias influences a language model's perception of color. *arXiv preprint arXiv:2110.08182*.

Ehsan Qasemi, Filip Ilievski, Muhao Chen, and Pedro Szekely. 2021. Paco: Preconditions attributed to commonsense knowledge. *arXiv preprint arXiv:2104.08712*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Shikhar Singh, Ehsan Qasemi, and Muhao Chen. 2022. Viphy: Probing" visible" physical commonsense knowledge. *arXiv preprint arXiv:2209.07000*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *Advances in Neural Information Processing Systems*, 34:24468–24481.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022. Visual commonsense in pretrained unimodal and multimodal models. *arXiv preprint arXiv:2205.01850*.

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Visualize before you write: Imagination-guided open-ended text generation. *arXiv preprint arXiv:2210.03765*.

A Appendix

A.1 Data Collection

The data collection is broken down into three sections: (1) collecting candidate objects, (2) annotating premise and hypothesis sentences, (3) verifying the quality of the annotated dataset, and (4) human evaluation.

Object Collection. To begin with, we gather a collection of objects along with their potential properties or attributes for constructing our data. These objects and attributes are obtained by scraping information from reliable sources such as Memory Colors (Norlund et al., 2021), Visual Property Norms (Hagström and Johansson, 2022), and McRae feature norms (McRae et al., 2005). Through this process, we manage to collect a total of 800 unique objects and 302 unique attributes. However, it is necessary to refine our dataset by filtering out attributes that are either too abstract or non-visual in nature. To accomplish this, we employ specific heuristics to ensure the inclusion of only concrete and visually relevant attributes. As a result of this filtering process, we successfully obtain a final dataset comprising 775 objects and 156 attributes.

Dataset Annotation. We utilized Amazon Mechanical Turk (Crowston, 2012) for data annotation, as depicted in Figure 1. The data annotation process involves several steps. Initially, annotators are given an object, and are instructed to identify two properties for the object and corresponding visual attributes for those properties. For example, for the object *banana*, the annotator may come up with two properties *ripe* and *rotten*, which would have corresponding visual attributes *yellow* and *brown*, respectively. After identifying a pair of object properties and visual attributes, they are tasked with composing natural language sentences for each attribute and property. The properties are associated with premise sentences, while the attributes were linked to hypothesis sentences.

Annotators were selected from a small pool of Mechanical Turkers that the authors had previously worked with. The Turkers had to further pass a qualification task that tested their understanding of the annotation task. The authors manually examined the annotations to ensure quality of the collected data.

Model	# Params	Public	VL model
BERT-Base	109M	✓	✗
BERT-Large	335M	✓	✗
VL-BERT-Large	335M	✓	✓
Oscar-Large	335M	✓	✓
CLIP-Large	427M	✓	✓
FLAN-T5-XXL	11B	✓	✗
InstructBLIP	11B	✓	✓
LLaMA2	13B	✓	✗
LLaVA	13B	✓	✓
GPT-3	175B	✗	✗
GPT-3.5	Unknown	✗	✗
GPT-4	Unknown	✗	✗

Table 5: **A list of models used in the experiments:** BERT (Kenton and Toutanova, 2019), CLIP (Radford et al., 2021), VL-BERT (Su et al., 2019), Oscar (Li et al., 2020), FLAN-T5 (Chung et al., 2022), InstructBLIP (Dai et al., 2023), LLaMA2 (Touvron et al., 2023), LLaVA (Liu et al., 2023), GPT-3 (Brown et al., 2020; Ouyang et al., 2022), and GPT-4 (OpenAI, 2023). We use the 'text-davinci-003' API for GPT-3, 'gpt-3.5-turbo-instruct' for GPT-3.5, and 'gpt-4-0314' for GPT-4.

Model	Ind.	Pair
FLAN-T5-Base (No imgs)	67.89	40.37
CLIP-Large	64.67	36.46
FLAN-T5-XXL (No imgs)	86.24	72.71
FLAN-T5-XXL (Captions)	85.83	71.88
InstructBLIP	53.21	22.93

Table 6: **Results on WINOVIZ with generated images.** We use Stable Diffusion (Rombach et al., 2022) to generate 5 images per premise sentence. We adopt majority voting at inference time to choose an option. FLAN-T5-Base (No imgs) refers to a model without any generated images, with a size comparable to CLIP-Large. FLAN-T5-XXL (No imgs) refers to a model without any generated images, while FLAN-T5-XXL (Captions) refers to a model with captions generated by BLIP2 on the generated images. Instead of directly inputting images into FLAN-T5, we extract captions from the generated images and use them as additional context. InstructBLIP uses generated images.

A.2 Versions of WINOVIZ

We now collect our WINOVIZ data. We also propose the multi-hop data, a more challenging version of WINOVIZ, and a dataset for probing visual knowledge. For the multi-hop data, we create new hypothesis options that require more intermediate steps while we simplify the premise sentences to measure the ability of models about visual knowledge.

Multi-hop Data. To create a more challenging task, we introduce a multi-hop version of our data, which requires more intermediate steps. The basic idea of the multi-hop data is to replace a visual attribute word in hypotheses with another object word which has a similar visual attribute. This requires one more reasoning step to find out the visual attribute. For example, one hypothesis option is 'The banana is yellow.'. Then 'yellow' can be replaced with 'the color of an egg yolk.' So the new hypothesis option for the multi-hop version is 'The banana is the color of an egg yolk.' The multi-hop version is more challenging since a model has to find out what color is an egg yolk. We focus on color, shape, material on the multi-hop data and curate prototypical objects for each visual property word. We get 200 samples for the multi-hop data.

Pragmatic Reasoning vs. Visual Knowledge Reasoning. Another important aspect of this work is that models genuinely understand and know visual knowledge. Our task requires pragmatic reasoning, the process of finding the intended meaning, and visual knowledge reasoning but models may fail in one of the reasoning steps. Thus, we decouple the premise sentence into pragmatic reasoning step and visual knowledge reasoning step to analyze which step is a bottleneck. Pragmatic reasoning involves finding the intended meaning and finding key phrases for the next step, visual knowledge reasoning. For example, a model should first find 'the banana is ripe' given the premise sentence in the pragmatic reasoning step (Figure 1). Given the simplified sentence, a model should choose a better option, 'the banana is yellow', in the visual knowledge reasoning step. We obtain 160 samples to study this (Section 3.5).

A.3 Using Image Generation for WINOVIZ Task.

Another approach for our task is to utilize image generation. We generate images based on premise sentences and employ these generated images for our task. The generated images may contain useful information that assists in identifying a correct hypothesis. We utilize an image generation approach, Stable Diffusion (Rombach et al., 2022), to generate images. We use the generated images to guide the LMs inspired by imagination-guided text generation (Zhu et al., 2022). Given the generated

images, there are three ways to use them. The first method involves using CLIP (Radford et al., 2021) on both the images and hypothesis sentences to identify a superior hypothesis option. Specifically, we calculate the cosine similarity between the embedding of a generated image and the embedding of a hypothesis option, selecting the hypothesis with a higher cosine similarity score. The second approach is to generate captions for the generated images using a caption model. Since language models cannot directly process images, we generate captions and utilize them as additional context for the task. BLIP2 (Li et al., 2023a) is employed for caption generation. The third strategy is to reframe our task as a visual question-answering task and employ a vision-language model to identify a better option. In this setup, we use InstructBLIP (Dai et al., 2023). For image generation, we use Stable Diffusion (Rombach et al., 2022), generating 5 images per premise sentence. A better hypothesis option is determined through majority voting.

Table 6 displays the outcomes related to image generation. The first approach utilizing CLIP falls short compared to FLAN-T5-Base which is slightly smaller than CLIP-Large. In the second approach involving BLIP2 captions, we opt for FLAN-T5-XXL as the benchmark, comparing one scenario with no additional data and another incorporating captions from generated images. Our experiment reveals a notable decline in performance when captions are employed. The third approach significantly underperforms FLAN-T5-XXL by a large margin. These experiments collectively indicate that generated images offer limited utility for our task. Furthermore, a manual assessment of 100 generated images reveals that 66% of them do not contribute meaningfully to our objectives. Examples of generated images with premise sentences are shown in Figure 3. In the figure, the bananas in both images are yellow; the generated images do not provide any clues to choose a more plausible option.

Figure 3: **Examples of generated images.** We generate images using Stable Diffusion (Rombach et al., 2022). In the second example, the bananas in both images are yellow, leading the model to select the incorrect option. The generated image examples don't assist in selecting a more plausible hypothesis option.

A.4 Related Work

There are multiple perspectives on how our contributions relate to previous work, and we elaborate on this in the subsequent sections.

Visual Knowledge Probing. Several attempts have been made to assess the reasoning ability of language models regarding objects, primarily through natural language benchmarks (Norlund et al., 2021; Hagström and Johansson, 2022; Paik et al., 2021; Zhang et al., 2022; Singh et al., 2022; Qasemi et al., 2021). Norlund et al. (2021) introduced a task involving querying a multimodal model for visual commonsense knowledge related to memory colors, which are the typical colors associated with well-known objects. Hagström and Johansson (2022) expanded on this work by proposing visual property norms as a measure of visual commonsense knowledge in both language models and multimodal models. Paik et al. (2021) evaluated the color perception of language models using a color dataset called CoDa, revealing that reporting bias negatively affects model performance and that multimodal training can alleviate these effects. Zhang et al. (2022) confirmed these findings and extended the evaluation to a wider range of visually salient properties. Similarly, Singh et al. (2022) evaluated vision-language models on a visually accessible commonsense knowledge dataset. Liu et al. (2022) explored spatial commonsense, the knowledge about spatial position and relationship between objects, finding that image synthesis

models are more capable of learning accurate and consistent spatial knowledge than other models. Gu et al. (2022) proposed a probing dataset for physical knowledge about everyday things. In contrast, we present a challenging dataset that probes the reasoning abilities of language models regarding variant visual properties of objects under different context.

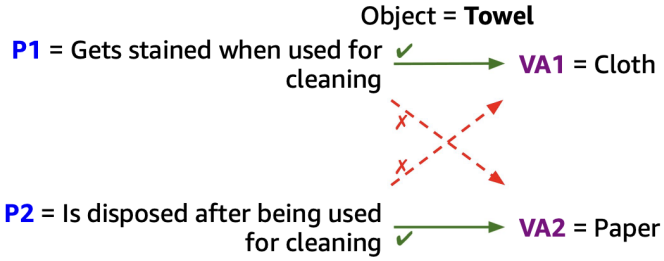
Vision-Language Modeling Recent advances in vision-language (VL) models have led to success on vision-language tasks such as visual question answering, captioning, and grounding (Antol et al., 2015; Lin et al., 2014; Mao et al., 2016). Existing VL models jointly learn image and text representations through cross-modal alignments including VL-BERT (Su et al., 2019), LXMERT (Tan and Bansal, 2019), Oscar (Li et al., 2020). Recent approaches have introduced visual instruction tuning, which involves fine-tuning a VL model using instruction-following data (Liu et al., 2023).

While these VL models have shown significant improvement in VL tasks, the exploration of how to transfer visual knowledge from VL modeling to language tasks remains underexplored. Vokenization (Tan and Bansal, 2020) utilized token-level text-to-image retrieval to transfer visual knowledge to language models. VidLanKD (Tang et al., 2021) employd contrastive learning to train a teacher model on video datasets and uses distillation approaches to transfer visual knowledge from the teacher to a student model. CMKT (Jin et al., 2022) investigated two types of knowledge transfer: text knowledge transfer (e.g., captions) and visual knowledge transfer (e.g., images and captions). Their findings demonstrate that such transfer can enhance performance on commonsense reasoning tasks.

A.5 Annotation Interfaces

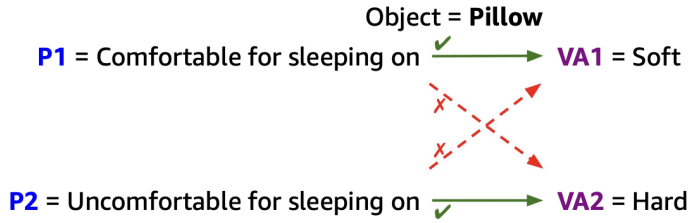
We provide Turking interfaces: qualification task in Figure 4, and annotation task in Figures 5, 6, 7, 8.

1. **Select the appropriate reason for which the below contrast set is invalid.**



- ○ One of the properties does not match with its corresponding visual attribute.
- ○ One of the properties matches with both visual attributes.

2. **Select the appropriate reason for which the below contrast set is invalid.**



- ○ The visual attributes are not completely visual (cannot be completely observed just from an image).
- ○ The visual attributes are visual, but are not always strongly associated with their corresponding properties.

Figure 4: **The Interface of the qualification task.** We provide 12 questions to find quality workers.

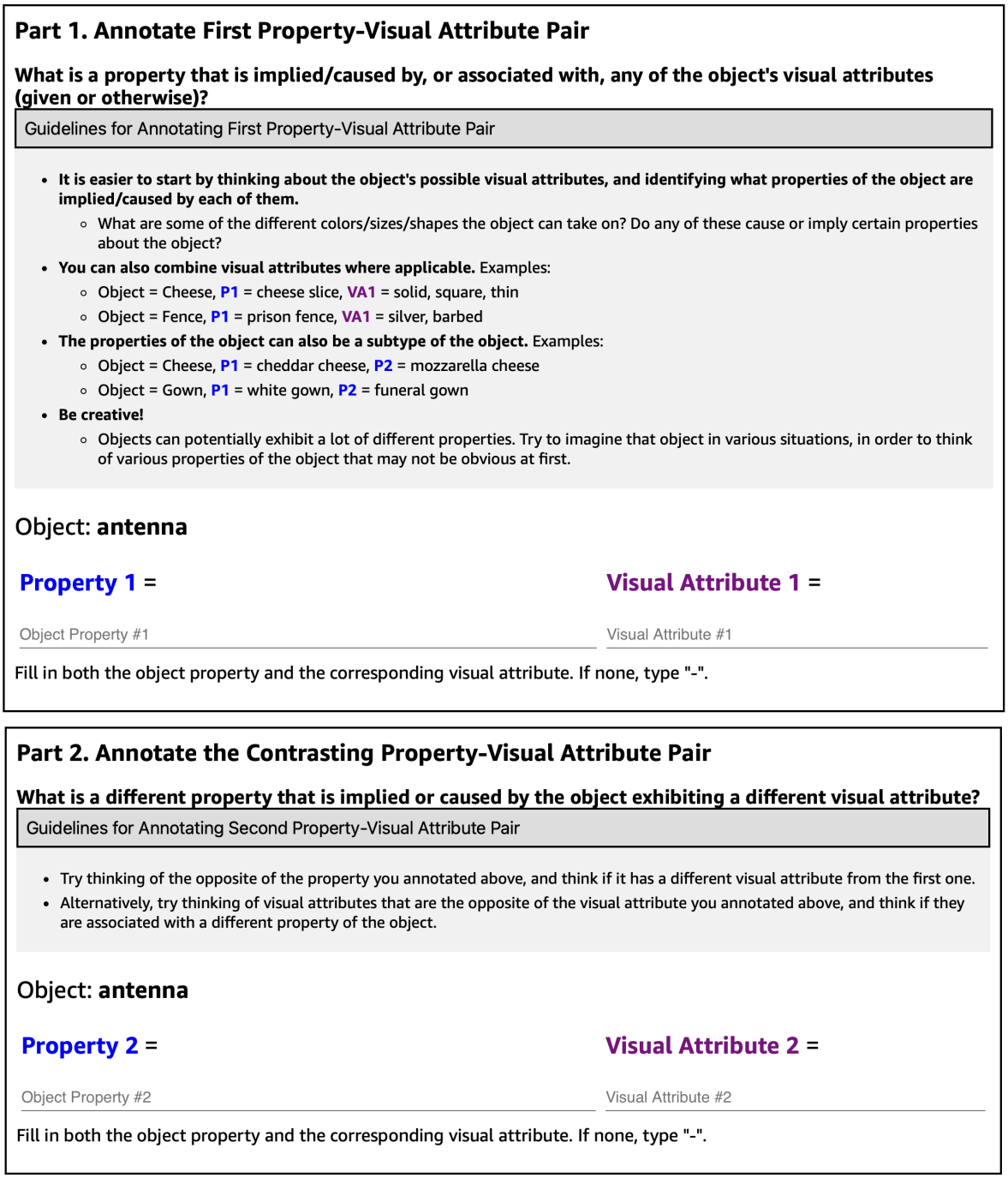
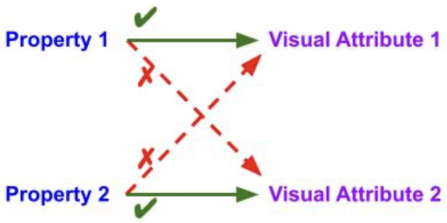


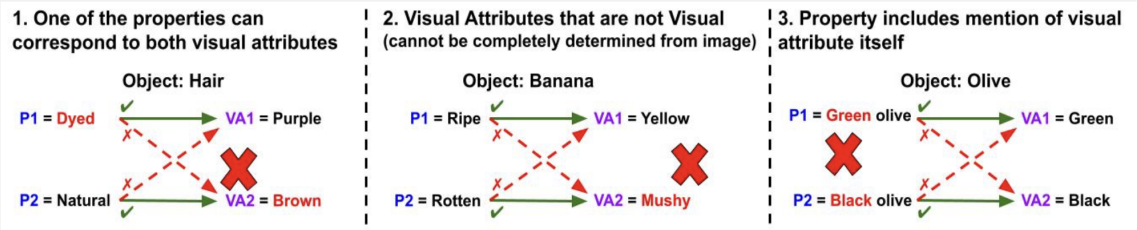
Figure 5: **Interfaces of annotating visual contrast sets (parts 1 and 2).**

Part 3. Ensure Validity of Visual Attributes and Properties



Your answers should be such that if the object has property **P1**, it is understood (by any human) that the object has visual attribute **VA1** rather than **VA2** (and vice versa)

Common mistakes resulting in invalid contrast sets



1. **Ensure properties and visual attributes across pairs do not match:**
   - If object **antenna** has property **property1**, then it is highly unlikely to have visual attribute **visualAttribute2**.
   - If object **antenna** has property **property2**, then it is highly unlikely to have visual attribute **visualAttribute1**.

2. **Ensure visual atributes are visual in nature:**
   - The visual attribute **visualAttribute1** is visual in nature, and can be completely observed/determined from an image of the object.
   - The visual attribute **visualAttribute2** is visual in nature, and can be completely observed/determined from an image of the object.

3. **Ensure properties do not contain mention of the corresponding visual attribute:**
   - The property **property1** does not contain a mention of the corresponding visual attribute **visualAttribute1**.
   - The property **property2** does not contain a mention of the corresponding visual attribute **visualAttribute2**.

Figure 6: **Interfaces of annotating visual contrast sets (part 3).**

Part 1: Create Sentences about the Object Properties

For each of the two properties you annotated above, write a sentence about the object in a real-world situation where it is exhibiting that property.

- The sentence should specifically mention both the object, and its property. **It should be clear from the sentence that the object in this sentence has that property**.
- The sentence should **NOT** mention the visual attribute corresponding to this property.
- Sentence should be **at least 10 words long**, and grammatically correct (begin with capital letter, end with punctuation).
- **Make the sentence as free-form and creative as possible.** Involve one or more characters in the sentence if possible.
- **DO NOT** make very short and simple sentences, that directly mention the object having a property and nothing else.Examples of bad sentences:
  - Object = **plate**, Property = **folded**, Sentence = *The plate was folded.*
  - Object = **cinnamon**, Property = **ground**, Sentence = *The cinnamon was ground.*
- If the two properties are opposites of each other (e.g. ripe vs rotten), you can write near-identical sentences with just the properties switched. However, do not force this if it does not make sense for the selected properties.

Example Sentences about Object Properties

- Object = **banana**, Property = **rotten**. Example Sentence: *He went to grab a quick breakfast before leaving, but saw that the only remaining banana was rotten.*
- Object = **box**, Property = **can be carried in one hand**. Example Sentence: *She already had the carpet in one hand, but picked up another box before heading up to the apartment.*
- Object = **cheese**, Property = **cheese slice**. Example Sentence: *He added a slice of cheese to his turkey sandwich.*
- Object = **napkin**, Property = **worn on lap at restaurants**. Example Sentence: *She placed the napkin across her lap before starting to eat her dinner.*

Object = antenna

Property 1 =

Property Sentence #1

Property 2 =

Property Sentence #2

Part 2: Create Sentences about the Visual Attributes

For each of the two visual attributes you annotated above, write a simple sentence that explicitly states that the object has that visual attribute.

- The sentence should specifically mention both the object and its visual attribute, and nothing else.
- The sentence should **NOT** mention the property corresponding to that visual attribute.
- Make the sentence as simple as possible. For e.g., "The banana was yellow", "The cheese was solid", "The nail was made of metal".
- Make the two sentences as identical to each other as possible, with only the visual attribute being different.
- Match the tense of the sentence to the tense of the corresponding property sentence - if the property sentence is in past tense, make the visual attribute sentence in past tense as well.

Example Sentences about Visual Attributes

- Object = **banana**, Visual Attribute = **black**. Example Sentence: *The banana was black.*
- Object = **box**, Visual Attribute = **small**. Example Sentence: *The box was small.*
- Object = **cheese**, Visual Attribute = **solid, square**. Example Sentence: *The cheese was solid and square.*

Suggested sentence format: *The **OBJECT** is/was **VISUAL ATTRIBUTE**.*

Object = antenna

Visual Attribute 1 =

Visual Attribute Sentence #1

Visual Attribute 1 =

Visual Attribute Sentence #2

Figure 7: **Interfaces of converting contrast sets into sentence puzzles (parts 1 and 2).**

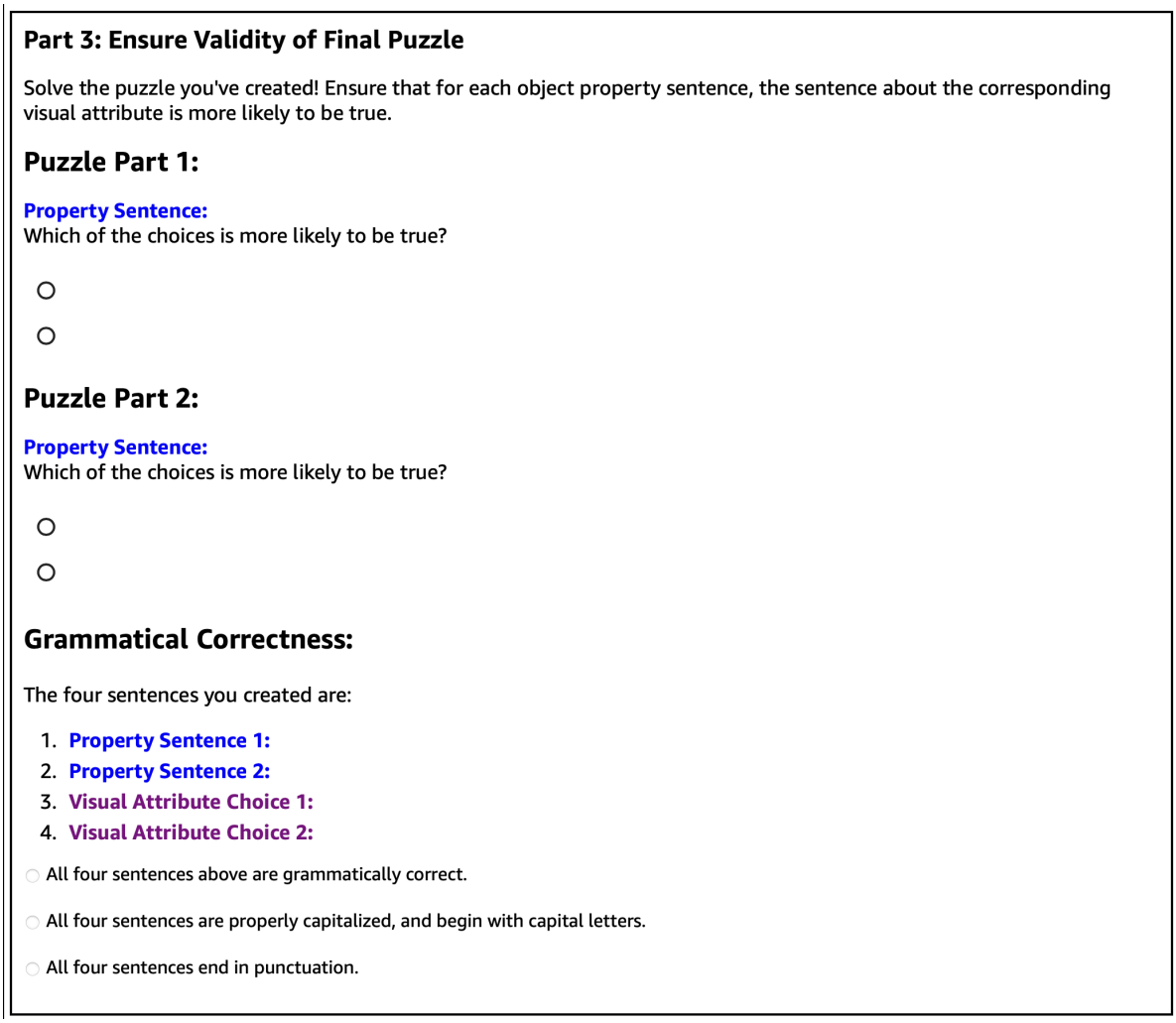


Figure 8: **Interfaces of converting contrast sets into sentence puzzles (part 3).**