Benchmarking Image Generators on Open-Vocabulary Scene Graphs

Brigit Schroeder
UC Santa Cruz
brschroe@ucsc.edu

Adam Smith
UC Santa Cruz
amsmith@ucsc.edu

Abstract

Prompt-driven image generation systems often face common problems such as missing objects, missing attributes, and blended objects. Scene graphs, which explicitly represent the relationships between objects and their attributes, hold potential to address these challenges due to their structured nature. However, previous work in scene graph to image generation relied on closed vocabularies, where having a small fixed vocabulary limited the flexibility and richness of the image generators . To overcome this limitation, we propose the idea of open vocabularies scene graphs (OVSGs) to capture the expressive power of free-form text while describing scene structures directly. We introduce new evaluation methods to better understand how existing generators fail on OVSGs, using both qualitative coding and a visual-question-and-answering (VQA) quiz to capture common failure scenarios in OVSG image generation (OVSG2IM). We find that all of the systems we evaluated (after adapting them to take OVSG inputs) demonstrate frequent flaws associated with not expressing details given explicitly in their graph inputs. However, existing image generators still struggle with OVSGs, indicating that there is room for improvement for future OVSG2IM systems.

1. Introduction

Modern image generators built on deep learning methods such as Stable Diffusion are impressive in how they create diverse and interesting images from text prompts [18, 23]. However, they are broken in the sense that they suffer from common problems like missing objects or missing attributes, blended objects, etc. in the rendered images [6, 13, 19]. In this paper, we seek to address these problems by proposing a new data structure which explicitly describes objects and their relationship while embracing the flexibility of free-form text, as a new input type to image generators.

We propose the idea of open vocabulary scene graphs (OVSGs) as a way to mitigate the problems image generators suffer from. A scene graph is a structured representation of a scene where nodes represent objects and edges represent semantic relationships between objects. This data structure has long been used in the field of computer vision [3, 15, 20]. In an open-vocabulary scene graph, arbitrary strings can be used to label the nodes and edges. An example of this can be seen in Figure 1. Unlike closed vocabulary scene graphs (CVSG), OVSGs are not limited to a fixed list of terms that a user must select from, which restricted the flexibility and expressiveness of previous generators [11].

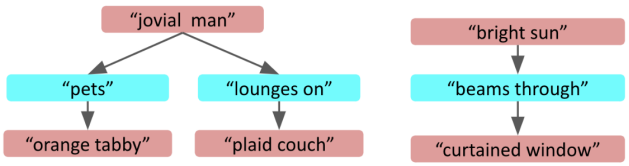


Figure 1. **Open Vocabulary Scene Graph.** An example of an open vocabulary scene graph used to describe a scene. Objects and their relationships are represented by strings containing free-form text.

This work makes a number of contributions. We build an argument for the importance of OVSGs based upon a review of the computer vision and NLP ethics literature, which points to the dangers and limitations of using closed vocabularies in computer vision and NLP models. We introduce two new methods of quickly adapting relevant graph and text-prompted image generators to OVSGs without the need to retrain models. The first method involves adapting an OVSG to a closed vocabulary scene graph using nearest neighbor word search in embedding space [5]. The second method converts OVSGs to delimited strings that can be used in prompt-driven image generators. Finally, we conduct comprehensive studies to understand how and why existing image generators fail to respect the structure and semantics of OVSGs. We use qualitative coding [2] as a way to encode a set of widely observed failure cases in OVSG image generators, giving us an easy way to describe the common errors image generators make. We create an automated evaluation in the form of a quiz which probes the accuracy of elements in images generated with OVSGs,

from an object, attribute and scene level perspective. We benchmark two well-known image generators: sg2im [11] which takes closed vocabulary scene graphs as input and Stable Diffusion [18] which takes open vocabulary-based text prompts.

2. Background

Our work tries to extend the expressiveness of text-prompted image generators to the realm of graph-prompted generators where closed vocabularies have been the basis for typical node and edge labels. This section problematizes image generation, argues that researchers should adopt open-vocabulary representations, and compares the affordances of graphs with text-only representations.

2.1. Text-driven Image Generation Systems

Large vision-language models like Stable Diffusion [18] and DALL-E2 [17] have gained attention for their remarkable ability to create high fidelity and expressive images. Both of these image generation systems use free-form text as image prompts which are based upon open vocabularies. In Figure 2, we can see an example image generated by Stable Diffusion (2.1) with the prompt "the purple bowl has green cubes." Even with a simple prompt, the image generator is prone to making errors, such as blending the attributes of objects in the prompt (the bowl is purple and green and the cubes are also purple and green). As we document later in this paper, other common failure cases are missing objects, blended objects, missing attributes and wrong attributes, although there is currently no unifying evaluation framework to benchmark these failure cases. There are existing benchmarks [14] which evaluate how well text-to-image synthesis models generalize to novel (new and unseen) compositions of concepts. However, this benchmark falls short in terms of being able to qualitatively and quantitatively capture common mistakes that image generators like Stable Diffusion make.

2.2. Scene Graphs vs. Text

A scene graph is a structured representation of a scene which encodes the objects as nodes and the relationship between them as edges. Looking specifically at objects in a scene, we would like to be able to describe the distinct attributes associated with each object. Text descriptions of scenes lack the structure and explicit representation of objects, attributes and relationships. This is evident in prompt-driven image generators, where objects and attributes can often get lost or become merged. Related to this, several recent papers [6,13] on text-driven image generators highlight the attribute binding problem attributes in a text prompts are bound to the wrong object (as seen in Figure 2). This a more specific case of the binding problem [22] found in neural networks (the network's inability to organize computational primitives in such a way that separates them into distinct objects). Compositionality is the ability of a system to combine multiple components together into a single image in a cohesive way. We want models to be compositionally competent enough to be able to generate meaningful images in a zero-shot manner, with objects unseen during training. A potential solution for addressing attribute binding (which is a compositionality challenge) is to provide the image generator with a data structure such as an open vocabulary scene graph (OVSG) as a way to reason about the compositionality of images. The OVSG has a direct representation of objects as nodes and edges as relationships, making them a much more succint and powerful way to represent an image, as opposed to the unstructured nature of text which can be a tedious way, sometimes requiring many extra words, of describing a scene.

Figure 2. **Image Generators Make Mistakes.** An image generated by Stable Diffusion with the prompt "the purple bowl has green cubes" that contains mistakes because due to the blending of color attributes. This is an example of an attribute binding problem.

2.3. Open Vocabularies

In natural language processing, closed vocabularies are represented by lists of words which almost always only represent English terms and tend to exclude words from other languages. Word choice is limited and this situation discriminates against non-English speakers [9], forcing the

user to pick the best term on a limited list of words, resulting in incomplete expression of ideas and concepts of those cultures. These vocabularies are often constructed using the highest frequency words found in the datasets they are sourced from (e.g. a large corpus of text derived from Wikipedia [5]). This is a problem as they predominantly represent commonly expressed ideas, topics and objects found in a specific widely-used language such as English (or a collection of languages). Natural language processing (NLP) and computer vision researchers have perpetuated bias against languages they described as 'low-resource' through building models that reinforce English-like word usage patterns. For example, many image classifiers are trained to use the vocabulary of the well-known ImageNet [4] dataset ends up supporting only 1000 specific English-only terms: e.g. "African crocodile" and "American lobster", and exclude common words like "thermometer" and "blanket".

Open vocabularies, on the other hand, are flexible [12] as they are not limited to a set of word tokens like closed vocabularies are. Individual concepts and ideas, limited to a single token in closed vocabularies, can be expressed by a multitude of terms found in natural language-based free-form text in the form of strings. Open vocabularies have far more versatility to deal with problems seen in closed vocabularies (e.g. support for multiple languages and dialects, slang, jargon, or accounting for imprecise spelling), While they don't necessarily solve the problems associated with closed vocabularies directly, they do provide systems with the potential to address these issues. For example, by using transfer learning, open vocabulary NLP models can be adapted for a target language by just using a few terms from the language via one-shot learning [24]

Large language models (LLMs) such as ChatGPT [7] use an open vocabulary representation and are trained with massive amounts of text data (e.g. 570 GB [10], which are inherently biased. According to ChatGPT [8], its "training data reflects the biases present in society from which that data was collected." Recent research [1] into LLMs discusses how "large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations." Open vocabularies offer the promise to address some of the biases seen in closed vocabularies, but it is important to be aware that systems built with them suffer from endemic bias problems related to the datasets they were built with. Despite this warning, findings in the field of natural language processing suggest that new systems should be built to support open vocabularies, and the computer vision community should follow this advice.

3. Adapting Existing Generators for OVSG

We have adapted a set of image generation systems to take OVSGs as input using two different approaches. In the first method, OVSGs are converted to closed vocabulary scene graphs (the format the first system sg2im natively accepts) using nearest neighbor word search in BERT [5] embedding space. Each open vocabulary term in the OVSG is matched to its nearest neighbor from the closed vocabulary term, creating a scene graph which is not very relatable to the original OVSG. For example, very specific relationships such as "lounging" may be converted to a term as generic as "on" as the closed vocabulary is very limited in its expressiveness. In the second method, the triplets (<subject, predicate, object>) found in the OVSG are concatenated as a set of comma-delimited strings which are fed to Stable Diffusion, a prompt-based image generator. This is a very naive method which doesn't leverage the structure and object definition intrinsically found in the scene graph. Figure 3 is an overview of how an OVSG is adapted into two different formats, one for each image generation system. Note how very expressive nodes containing objects and attributes (e.g. "hazy moon") are collapsed into single terms (e.g. "sky") which are nonspecific and far less descriptive.

4. Methods for Evaluating Adapted Generators

In this section, we present two evaluation methods which were used to benchmark image generators on OVSGs.

4.1. Qualitative Coding

We apply qualitative coding [2] to create a dictionary of terms that captures commonly-seen failure cases in both image generation systems. Each image in the qualitative evaluation set is labeled with a set of descriptive codes which are described in the following bulleted list:

- **Missing Object (MO)**: Objects represented by nodes in the graph are omitted from image.
- **Ambiguous Object (AO)**: Objects represented by nodes in the graph are not recognizable due to malformation, etc. and appearance is unclear.
- **Blended Object (BO)**: Multiple objects, represented by nodes of a graph, are blended into one object, creating an object with the appearance of two partial objects.
- **Missing Attribute (MA)**: Attribute that should be attached to specific object node is missing.
- **Wrong Attribute (WA)**: Attribute that should be attached to a specific object node is attached to the wrong object.

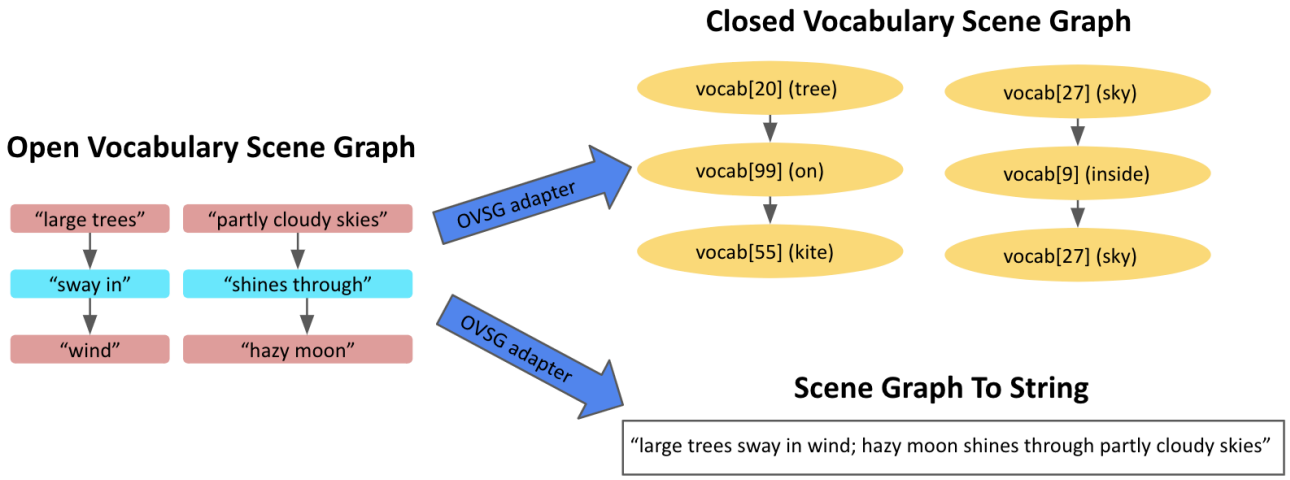


Figure 3. **Open Vocabulary Scene Graph Adapters.** In this figure we can see two open vocabulary scene graph adapters that adapt an OVSG to a closed vocabulary scene graph and a set of delimited strings.

- **Missing Relationship (MR)**: The relationship between two objects represented by nodes and linked by a graph edge is missing.
- **Wrong Relationship (WR)**: The relationship between two objects represented by nodes and linked by a graph edge is incorrect.
- **Missing Triplet (MT)**: An entire triplet relationship, represented by two graph nodes and an edge, is missing from the scene.

The codes describe specific flaws related to objects, attributes and relationships in an image, which are typically wrong or missing, reflecting how image generators fail to respect the structure of OVSGs. An example of each type of qualitative code described above can be seen in Figure 4, which shows failures in both sg2im and Stable Diffusion.

4.2. VQA-Style Probing Quiz

We quantitatively evaluate each of the benchmarked image generators by creating a CLIP-based [16] visual-question-and-answer (VQA) style quiz [14, 21], to probe the accuracy of the compositional elements in the images generated from OVSGs. We evaluate the fidelity of these elements at different levels of detail: attributed objects (e.g. "purple dog"), generic objects (e.g. "dog"), relationships (e.g. "holding") and scene-level details (e.g. "person holding dog"). In the end, our quiz will reveal to us how faithful the image generators were to the OVSGs as each question is designed to probe each of the evaluated elements of a scene. For example one simple quiz question might be to determine whether an attributed object is present in the scene or not. The CLIP model [16] can represent both images and text in a joint vector space, allowing us to measure the cosine similarity between a generated image and several simple text captions describing the image (derived from the components of the OVSGs), as seen in the example in Figure 5 Images and text that are visually and semantically similar should have high cosine similarity scores and the caption with the highest score is selected as the answer. The goal of the quiz is to measure how well and to what level the generated images align with the correct answers. In Figure 5, the quiz question probes whether there are "black swans" or just "swans" in the image (both can be considered correct answers when probing on different levels of detail). Images generated by poorly-adapted OVSG systems will tend to select the wrong answers as they will often have missing or distractor elements in the scene which do not represent the original OVSG very well.

5. Human Evaluation Results

We performed a human evaluation of two OVSG adapted image generation systems by qualitatively coding a small dataset of 20 images generated from a set of 10 OVSGs, generating one image for each system benchmarked (sg2im and Stable Diffusion). In this evaluation, a human assigns qualitative codes to each image (no limit in the number used) and then lists the details of the failure next to the code (the full set of codes are described in Section 4). Figure 6 is an example of the qualitative codes applied to a subset of the images in the evaluation dataset (the full evaluation can be found in the Appendix). When OVSGs are adapted to closed vocabulary image generation systems (e.g. sg2im),

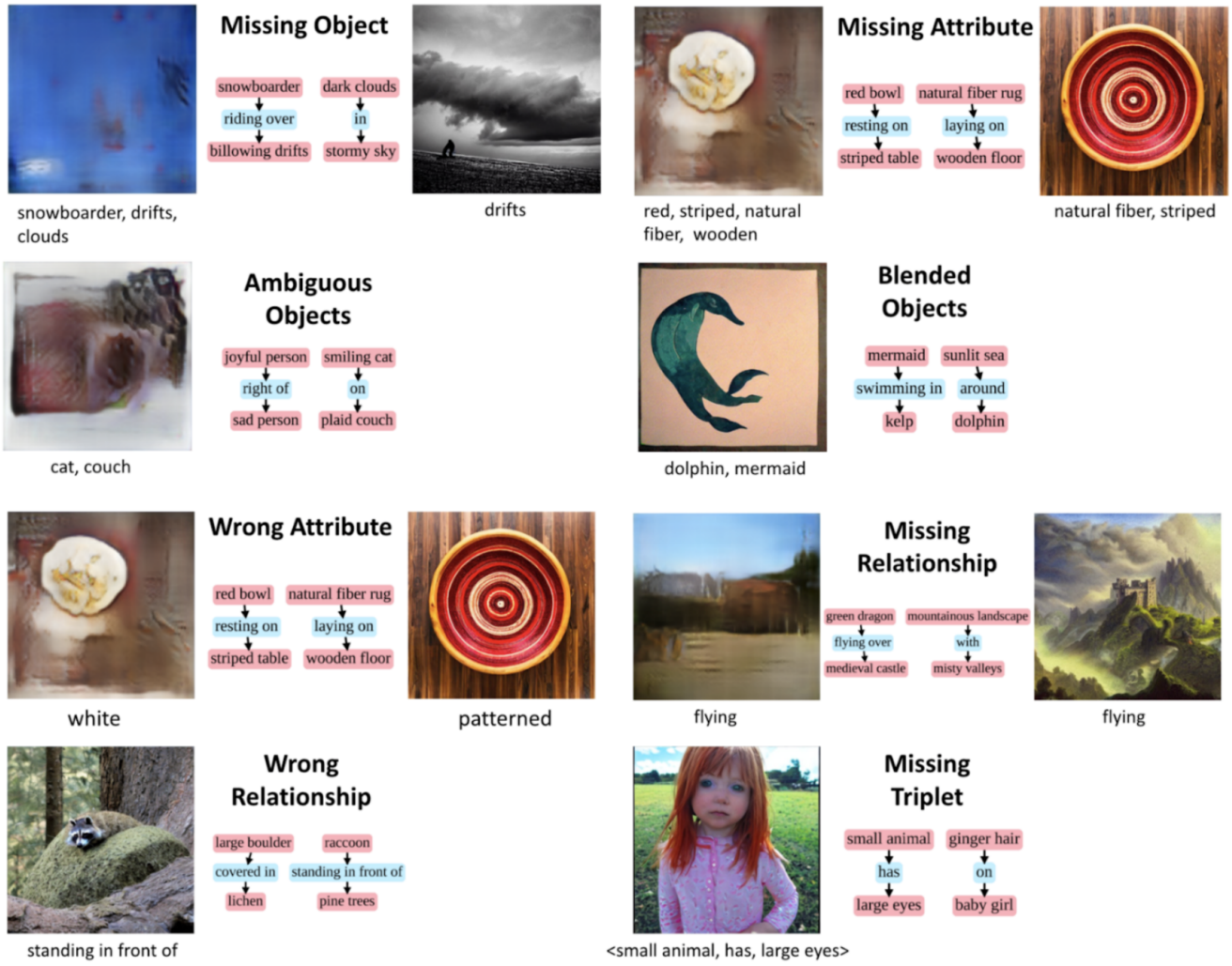


Figure 4. **Qualitative Coding of OVSG Image Generators.** Presented here are examples of each of the qualitative code previously described which capture the common failures of image generators when applied to OVSGs.

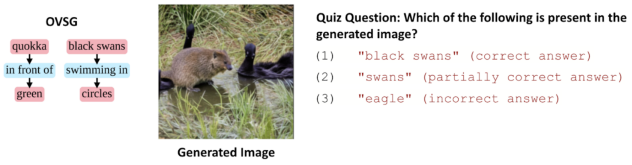


Figure 5. **Probing for Objects in an Image.** A sample VQA-style quiz which probes whether an object is in the scene or not. The image in the figure has been generated using the OVSG on the left, and simple quiz questions are derived from details in the scene graph.

we see large gaps in the actual and desired output (an image that stays true to the semantics of the OVSGs), as evidenced by the images in Figure 6. One of the most common failure cases is ambiguous objects, likely due to the overgeneralization of open vocabulary terms to very generic closed vocabulary terms (e.g. all open vocabulary person-like nodes such as "mermaid" or "joyful person" get collapsed a generic "person" term in a closed vocabulary). Objects are frequently missing because they may not have a plausible near equivalent in the closed vocabulary set of terms available (e.g. there is no "kelp" or "spaceship" in sg2im's closed vocabulary). Even though Stable Diffusion is originally an open vocabulary text-driven system, it still exhibits several failures when taking OVSGs as input. One observed failure case is blended objects (e.g. blended dolphin/mermaid in Figure 6, where object nodes in the scene graph are merged, pointing to the need for Stable Diffusion to support graph-based data structures as input types.

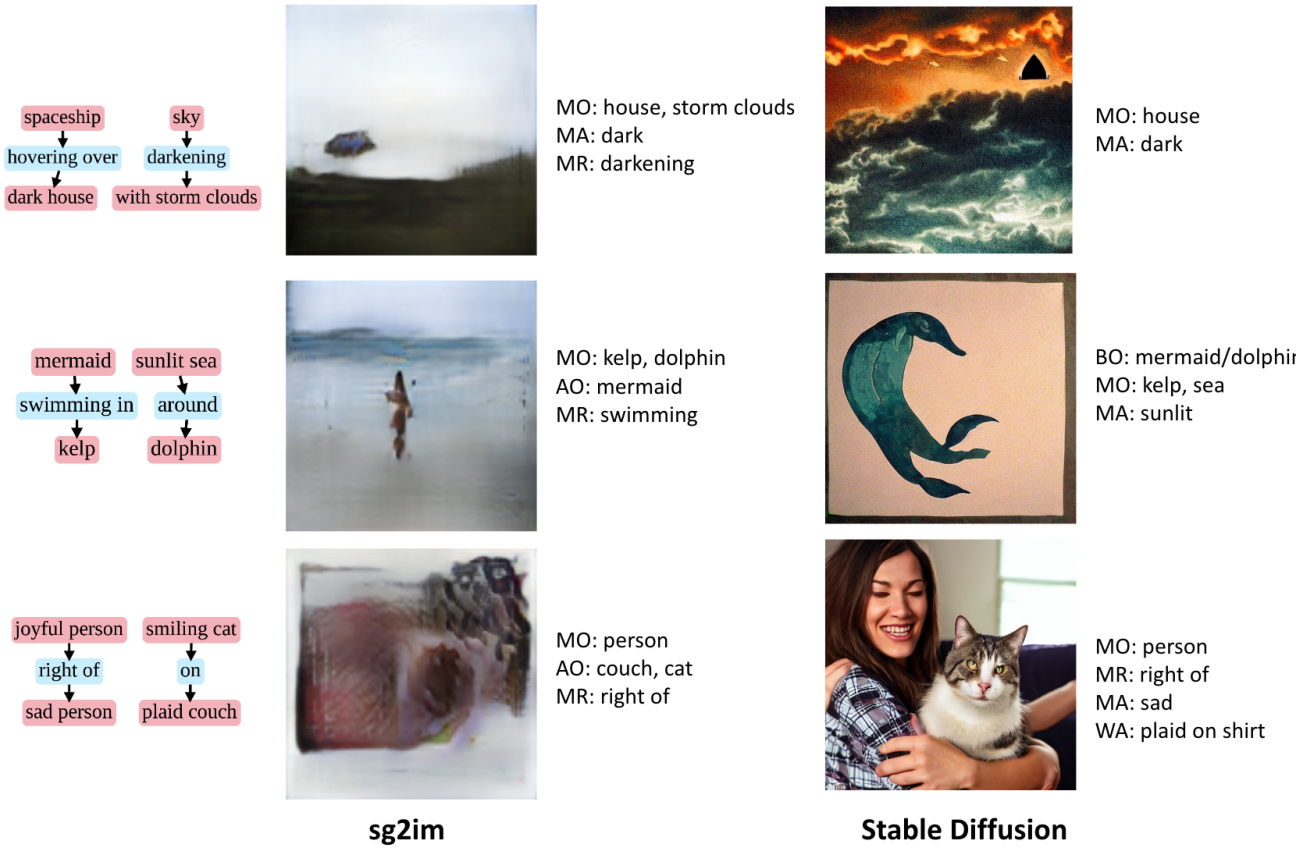


Figure 6. **Human Evaluation using Qualitative Coding.** An example of a human evaluation of two OVSG adapted image generation systems, sg2im and Stable Diffusion, using qualitative codes that capture common system failures.

Adapting OVSGs as strings does not provide Stable Diffusion with enough structure to be able to discern one object from another, causing objects to be blended or even omitted. The lack of distinct structure in the string-based OVSG also lends to the problem of attributes being bound to the wrong objects (as seen in [6]), such as where "plaid" is bound to a person rather than a couch.

6. Automated Evaluation Results

An automatic evaluation of the two OVSG adapted image generation systems was done using a CLIP-based [1] VQA-style quiz to probe the correctness of the elements in the images generated from OVSGs. We want to quantify to an element-wise level how faithful the adapted image generators are to the original OVSG. A set of 30 OVSGs was used to generate a dataset of images for evaluation for each of the benchmarked systems. 30 images were generated for sg2im (1 image per OVSG) and 90 images were generated for Stable Diffusion (3 images per OVSG using k=3 image seeds). The full dataset with the corresponding scene graphs can be found in the appendix. The results of the quiz-based evaluation can be seen in Table 1. The open vocabulary-based scene graph to image generation system (OVSG-adapted Stable Diffusion) is much more successful in producing realistic recognizable objects (46.7% vs. 4.4% ("attributed objects")), specific identifiable relationships (such as "holding" vs. generic "has" frequently seen in closed vocabularies) (32.2% vs 24.4% ("relationships")) and semantically correct scene-level details (45.6% vs. 15.6% ("scene")) than closed vocabulary systems (OVSG-adapted sg2im). Sg2im produces more generically identifiable objects (31.1% vs. 15.5% ("generic objects")) than Stable Diffusion as its closed vocabulary does not support attributes and is not very expressive (all very specific object classes get collapsed into generic classes, as discussed in the Human Evaluation section).

[1] https://huggingface.co/openai/clip-vit-base-patch32

Image Elements	sg2im [11]	Stable Diffusion [18]
Attributed Objects	4.4	46.7 $\pm$0.0
Generic Objects	31.1	15.6 $\pm$0.0
Relationships	24.4	33.3 $\pm$1.1
Scene	15.6	48.9 $\pm$3.3

Table 1. **Evaluation of OVSG Image Generation Quality.** The table is the result of a CLIP-based VQA quiz [14, 21] probing the accuracy of the images generated with OVSGs.

7. Conclusion

In this paper, we present an argument for the importance of OVSGs in the landscape of current image generation systems. We demonstrate two methods to rapidly adapt existing open and closed vocabulary systems so that they can easily be paired with OVSGs. We build an evaluation system that demonstrates how current image generation systems fail when coupled with OVSGs through a detailed qualitative and quantitative analysis. The potential benefits of OVSGs are not fully realized with current image generators, indicating that there is room for growth to build OVSG-native generators that would help resolve observed failures during the evaluation.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, Mar. 2021. Association for Computer Machinery – ACM.

[2] Kathy Charmaz. Constructing grounded theory : A practical guide through qualitative analysis., 2006.

[3] Ciro de Mauro, Michelangelo Diligenti, Marco Gori, and Marco Maggini. Similarity learning for graph-based image representations. *Pattern Recognition Letters*, 24(8):1115 – 1122, 2003. Graph-based Representations in Pattern Recognition.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. cite arxiv:1810.04805Comment: 13 pages.

[6] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *ICLR*, 2023.

[7] Daniel De Freitas, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977, 2020.

[8] Rachel Gordon. Large language models are biased. can logic help save them?, 2023.

[9] Louis Hickman, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srinivasan. Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146, 2022.

[10] Alex Hughes. Chatgpt: Everything you need to know about openai's gpt-4 tool, 2023.

[11] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. *CVPR*, 2018.

[12] Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *ArXiv*, abs/2112.10508, 2021.

[13] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 2022.

[14] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *NeurIPS Datasets and Benchmarks*, 2021.

[15] Franco P. Preparata and Sylvian R. Ray. An approach to artificial nonsymbolic cognition. *Information Sciences*, 4(1):65 – 86, 1972.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[17] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[19] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-

to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022.

[20] Linda G. Shapiro, John D. Moriarty, Robert M. Haralick, and Prasanna G. Mulgaonkar. Matching three-dimensional objects using a relational paradigm. *Pattern Recognition*, 17(4):385 – 405, 1984.

[21] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022.

[22] Sjoerd van Steenkiste, Klaus Greff, and Jürgen Schmidhuber. A perspective on objects and systematic generalization in model-based RL. *CoRR*, abs/1906.01035, 2019.

[23] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey, 2023.

[24] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, November 2016. Association for Computational Linguistics.

Appendix

A. Model Details

Several pre-trained models were used in the course of this paper. Below are links referring to the origin of each model:

- **sg2im:** The model used in sg2im [11] can be downloaded here. The outputs of sg2im are unaffected by choice of random seed so none was used.

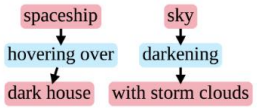
- **Stable Diffusion:** The model card for the version of Stable Diffusion [18] used in this paper can be found here. Random seeds 0,1 and 2 were used when running model over 3 different evaluation rounds.

- **BERT:** The BERT [5] model used in this paper was based upon Sentence Transformer and the model card can be found here.

- **CLIP:** The model card for the version of CLIP [16] used in this paper can be found here.

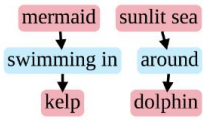
B. Human Evaluation Data

The human evaluation was conducted by the first author of this paper using the qualitative codes outlined in Section 4.1 of the paper. For each generated image example, a number of qualitative codes (such as ”MO” for ”Missing Object”) are assigned and a simple detail about each is given right after (e.g. ”MO: cat”).



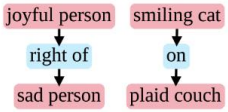
MO: house, storm clouds
MA: dark
MR: darkening

MO: house
MA: dark



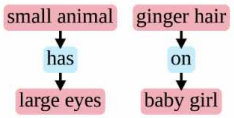
MO: kelp, dolphin
AO: mermaid
MR: swimming

BO: mermaid/dolphin
MO: kelp, sea
MA: sunlit



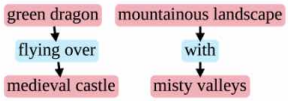
MO: person
AO: couch, cat
MR: right of

MO: person
MR: right of
MA: sad
WA: plaid on shirt



MO: animal
MA: ginger, large, small
MR: has

MT: <small animal, has,large eyes>



MO: house, storm clouds
MA: dark
MR: darkening

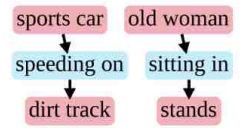
MO: house
MA: dark

MO: snowboarder, drifts, clouds
MA: billowing, dark, stormy
MR: riding over

MO: drifts
MA: billowing

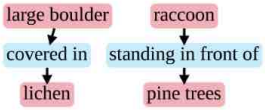
MO: rug, table
MA: striped, natural fiber
MR: laying on

MO: rug, table
WA: patterned instead of red
MA: natural fiber, striped



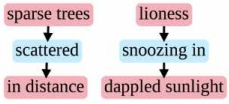
MO: stands, woman
MA: sports, od
MR: sitting

MO: stands
AO: old woman



MO: raccoon
MA:pine
MT: <large boulder, covered in, lichen>

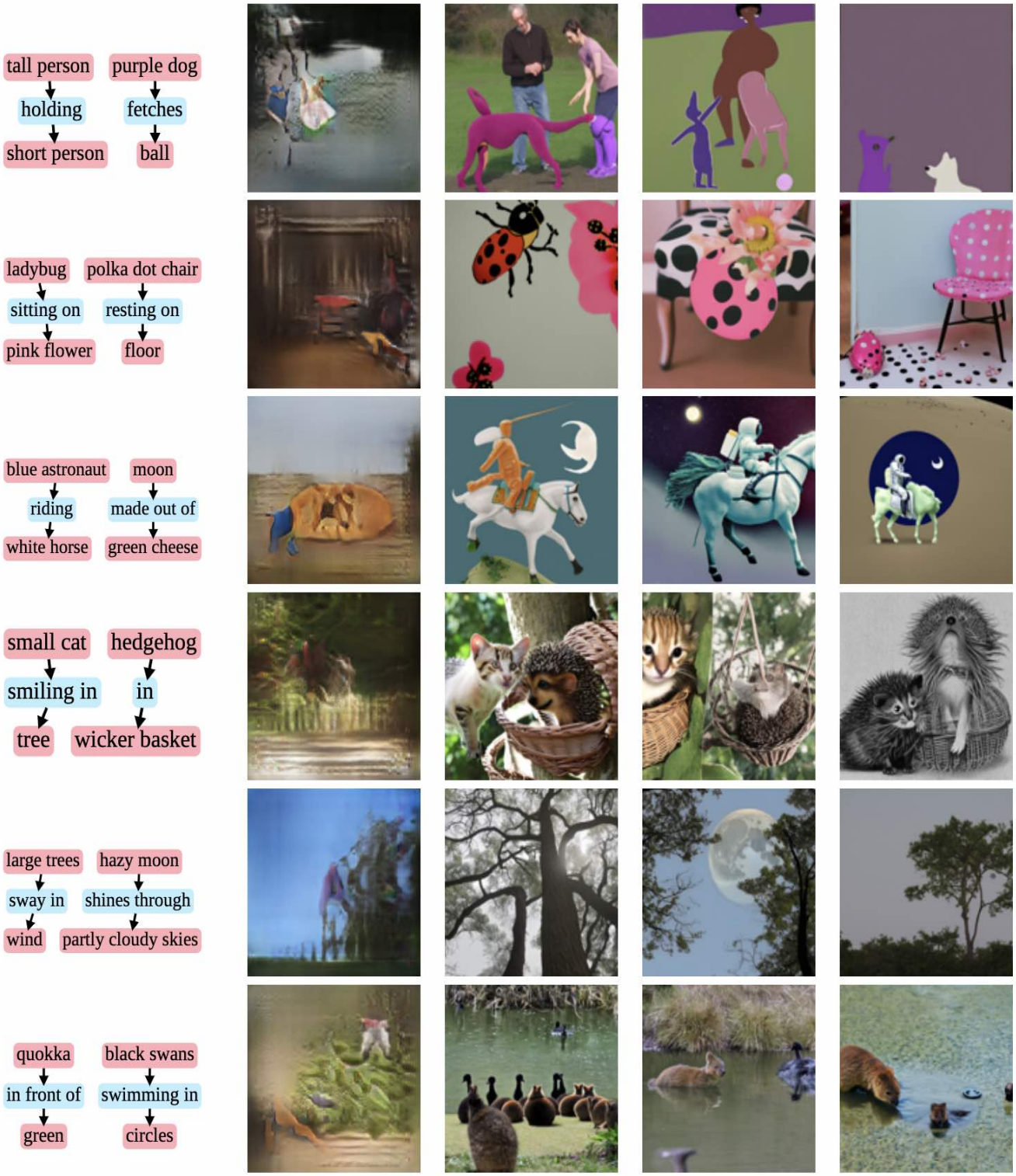
AO: boulder
WR: laying on



MO: sunlight
AO: trees, lioness
MR: snoozing, scattered

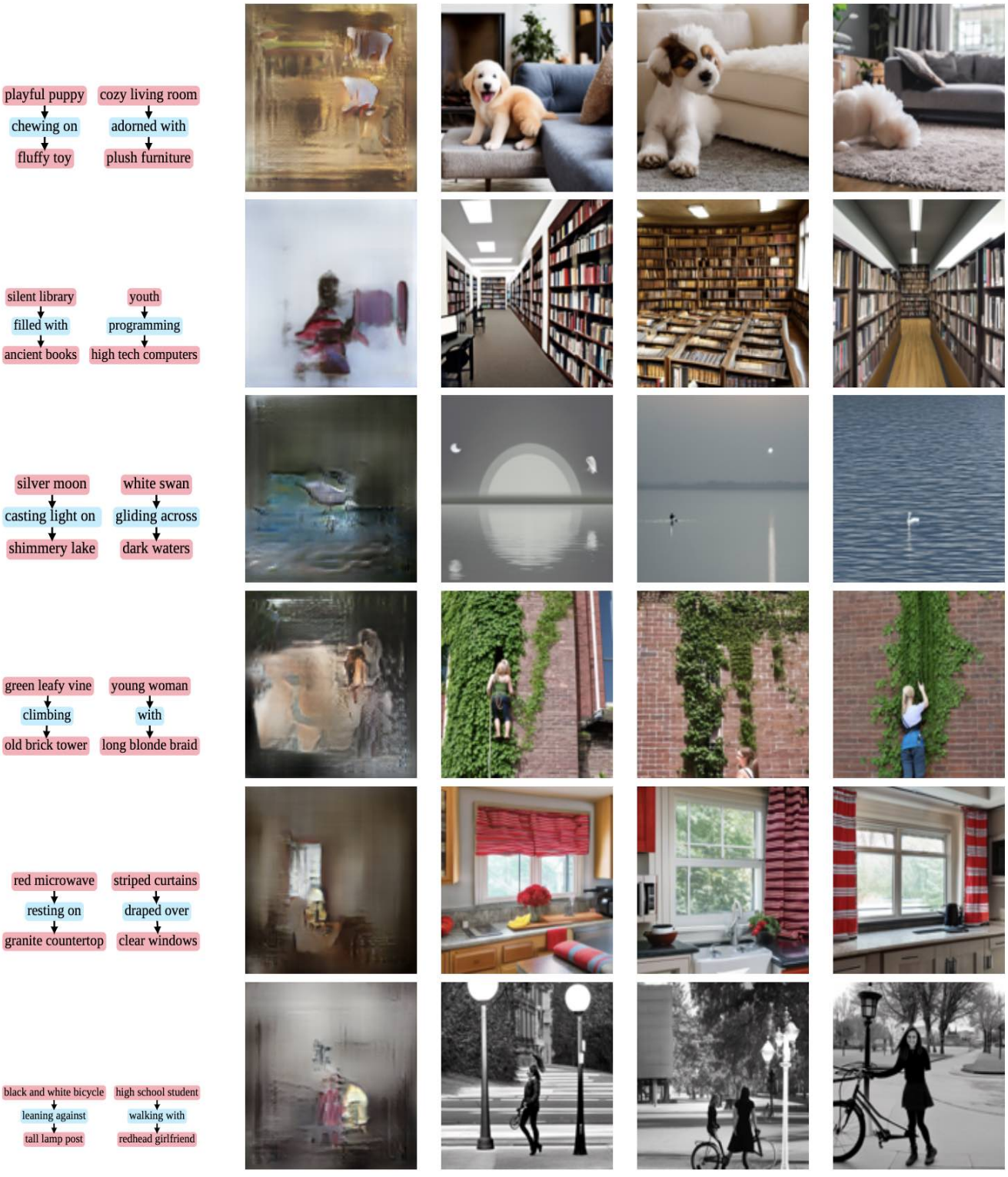
MO: sunlight
MT: <sparse trees, scattered, in distance>

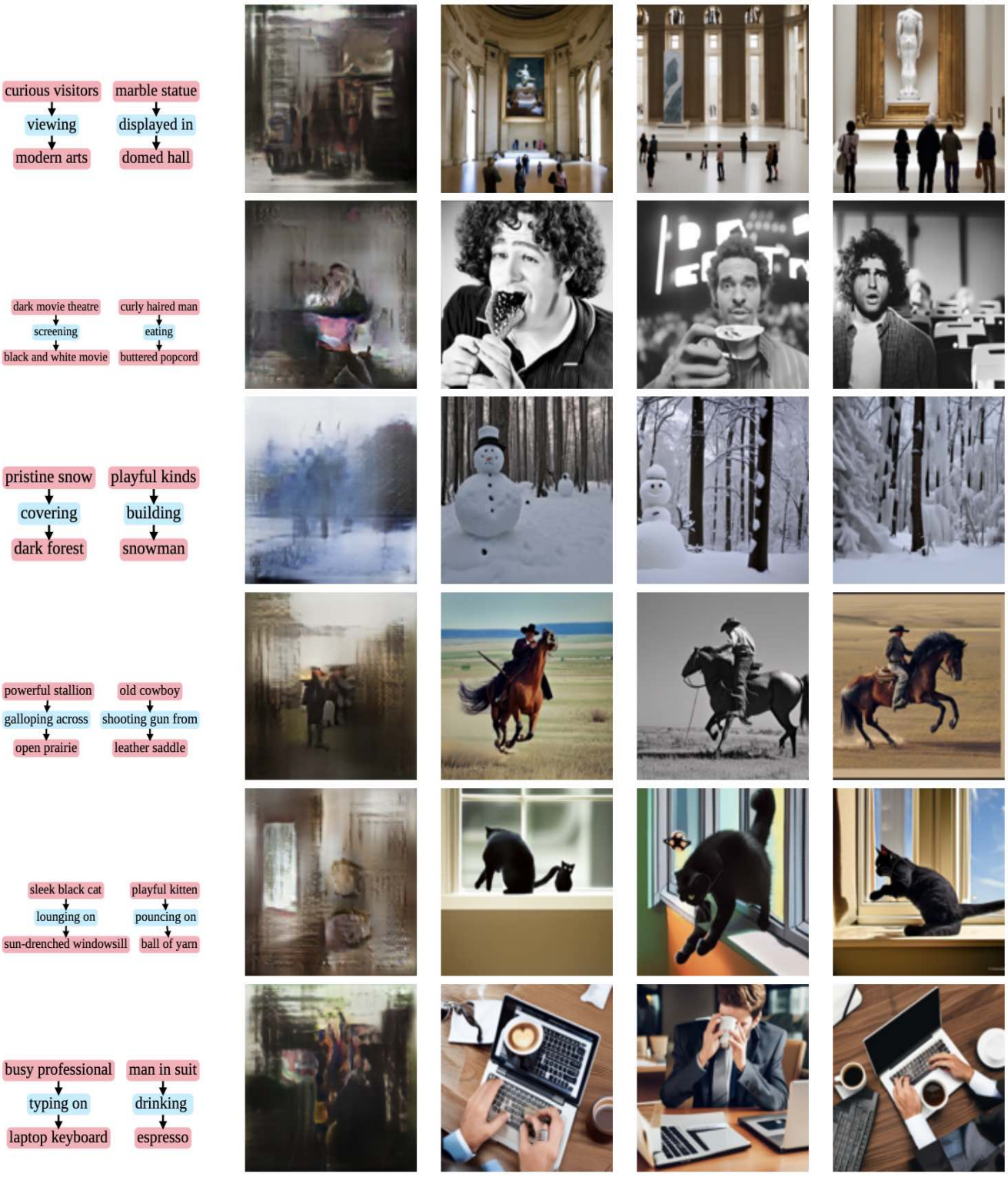
C. Automatic Evaluation Data




old woman
embraces
strong mustang
red barn
standing in
wheat field
bearded man
hikes up
steep trail
bright orange tent
sitting in
wooded campsite
yellow tulips
placed in
patterned vase
calico cat
sleeping on
coffee table
female musician
plays
clarinet
large crowd
listens in
concert hall
bright coffee cup
filled with
espresso
stainless steel appliances
standing in
kitchen
tabby cat
lounging on
wooden stool
purple couch
placed on
striped rug


butternut squash
corn
laying on
standing in
ground
verdant field
sparrows
fat black squirrel
flocking to
hanging from
bird feeder
avocado tree
tortoiseshell cat
brown mouse
lounging on
runs along
blue polka dot rug
baseboard
sparking ocean
old tree
crashing against
shading
rugged cliff
playful children
roaring lion
snow-capped mountain
hunting
overlooking
gazelle
serene valley
crowded market
delivery truck
selling
parked in front of
colorful produce
busy building


playful puppy
chewing on
fluffy toy
cozy living room
adorned with
plush furniture
silent library
filled with
ancient books
youth
programming
high tech computers
silver moon
casting light on
shimmery lake
white swan
gliding across
dark waters
green leafy vine
climbing
old brick tower
young woman
with
long blonde braid
red microwave
resting on
granite countertop
striped curtains
draped over
clear windows
black and white bicycle
leaning against
tall lamp post
high school student
walking with
redhead girlfriend



D. Automatic Evaluation JSON Quiz

The following is an embedded JSON document which contains all of the quiz questions used in the Automatic Evaluation. Each element of the quiz question array contains a scene graph, a set of 3 answers, and the question type. Type ”obj” are

questions about objects (e.g. ”man”) in the scene, type ”rel” are questions about relationships (e.g. ”holding”) and type ”scene” are questions about scene-level details (e.g. ”man holding dog”). There are 30 unique scene graphs in the quiz and there are 3 types of question per scene graph (”obj”, ”rel” and ”scene”), so the full quiz array has 90 questions (30 x 3) in total. Since each quiz question probes for specific object, relationship or scene-level detail elements, if the generator gets the question wrong it is likely that the element (one of the three types) in question is missing from the image. If the generator gets the question right, it means that it has identified that that element exists in the image.

In the ”answer” element of each quiz question, there are 3 answers. The first one is always the correct answer for all question types. In the specific case of the ”obj” question referring to objects, there are 2 possible correct answers: the first answer is the correct answer which corresponds to the ”Attributed Object” result in Table 1, the second answer is a partially correct answer (see Figure 5) which corresponds to the ”Generic Object” result in Table 1. Both are valid responses and should be tallied separately. The way to interpret this result is to say, ”Out of 30 object quiz questions, for answers the system selected 46.7% as Attributed Objects, 15.6% as Generic Objects and 37.7% as incorrect (the rest). All told, the system could identify an object 62.3% of the time.” The sample results here correspond to the Stable Diffusion result in Table 1.

JSON Quiz Data


```
{
   {"scene_graph": {"objects": ["tall person","short person", "purple dog", "ball"],
 "relationships": [[0, "holding", 1],[2, "fetches", 3]] },
  "answers": ["a tall person",
              "person",
              "snowman"],
  "type": "obj"},
  {"scene_graph": {"objects": ["tall person","short person", "purple dog", "ball"],
  "relationships": [[0, "holding", 1],[2, "fetches", 3]] },
  "answers": ["person holding person ",
              "person dancing with dog ",
              "cat holding dog"],
  "type": "scene"},
  {"scene_graph": {"objects": ["tall person","short person", "purple dog", "ball"],
  "relationships": [[0, "holding", 1],[2, "fetches", 3]] },
  "answers": ["holding",
              "dancing ",
              "eating"],
  "type": "rel"},

 {"scene_graph": {"objects": ["ladybug","pink flower", "polka dot chair", "floor"],
  "relationships": [[0, "sitting on", 1], [2, "resting on", 3]] },
  "answers": ["chair with polka dots",
              "chair",
              "couch"],
  "type": "obj"},
 {"scene_graph": {"objects": ["ladybug","pink flower", "polka dot chair", "floor"],
  "relationships": [[0, "sitting on", 1], [2, "resting on", 3]] },
  "answers": ["ladybug sitting on flower",
              "ladybug next to flower",
              "insect flying over flower"],
  "type": "scene"},
 {"scene_graph": {"objects": ["ladybug","pink flower", "polka dot chair", "floor"],
  "relationships": [[0, "sitting on", 1], [2, "resting on", 3]] },
  "answers": ["sitting on",
              "next to",
              "flying over"],
  "type": "rel"},
```

```
{"scene_graph": {"objects": ["blue astronaut","white horse", "moon", "green cheese"],
 "relationships": [[0, "riding", 1], [2, "made out of", 3]] },
 "answers": ["a white horse",
             "an animal",
             "turtle"],
 "type": "obj"},
{"scene_graph": {"objects": ["blue astronaut","white horse", "moon", "green cheese"],
 "relationships": [[0, "riding", 1], [2, "made out of", 3]] },
 "answers": ["moon made out of cheese",
             "moon eating cheese",
             "sun shining bright"],
 "type": "scene"},
 {"scene_graph": {"objects": ["blue astronaut","white horse", "moon", "green cheese"],
 "relationships": [[0, "riding", 1], [2, "made out of", 3]] },
 "answers": ["made out of",
             "eating",
             "shining"],
 "type": "rel"},

{"scene_graph": {"objects": ["small cat","tree", "hedgehog", "wicker basket"],
 "relationships": [[0, "smiling in", 1], [2, "in", 3]] },
 "answers": ["a smiling cat",
             "a cat",
             "dog"],
 "type": "obj"},
{"scene_graph": {"objects": ["small cat","tree", "hedgehog", "wicker basket"],
 "relationships": [[0, "smiling in", 1], [2, "in", 3]] },
 "answers": ["a hedgehog in a basket",
             "a hedgehogs eating a basket",
             "a cat sleeping in box"],
 "type": "scene"},
 {"scene_graph": {"objects": ["small cat","tree", "hedgehog", "wicker basket"],
 "relationships": [[0, "smiling in", 1], [2, "in", 3]] },
 "answers": ["in",
             "eating",
             "sleeping"],
 "type": "rel"},

{"scene_graph": {"objects": ["large trees","wind", "hazy moon",
                  "partly cloudy skies"],
 "relationships": [[0, "sway in", 1], [2, "shines through", 3]] },
 "answers": ["partly cloudy skies ",
             "nighttime",
             "daytime"],
 "type": "obj"},
{"scene_graph": {"objects": ["large trees","wind", "hazy moon",
                "partly cloudy skies"],
 "relationships": [[0, "sway in", 1], [2, "shines through", 3]] },
 "answers": ["the moon in the sky",
             "the moon under the sky",
             "sun in the sky"],
 "type": "scene"},
```

```
 {"scene_graph": {"objects": ["large trees","wind", "hazy moon",
                "partly cloudy skies"],
 "relationships": [[0, "sway in", 1], [2, "shines through", 3]] },
 "answers": ["in",
             "under",
             "above"],
 "type": "rel"},

{"scene_graph": {"objects": ["quokka","green", "black swans", "circles"],
 "relationships": [[0, "in front of", 1], [2, "swimming in", 3]] },
 "answers": ["black swans",
             "swans",
             "eagle"],
 "type": "obj"},
{"scene_graph": {"objects": ["quokka","green", "black swans", "circles"],
 "relationships": [[0, "in front of", 1], [2, "swimming in", 3]] },
 "answers": ["animal in front of green bush",
             "animal behind green bush",
             "human laying in front bush"],
 "type": "scene"},
 {"scene_graph": {"objects": ["quokka","green", "black swans", "circles"],
 "relationships": [[0, "in front of", 1], [2, "swimming in", 3]] },
 "answers": ["in front of",
             "behind",
             "laying in"],
 "type": "rel"},

{"scene_graph": {"objects": ["old woman","strong mustang", "red barn", "wheat field"],
 "relationships": [[0, "embraces", 1], [2, "standing in", 3]] },
 "answers": ["old woman",
             "woman",
             "man"],
 "type": "obj"},
{"scene_graph": {"objects": ["old woman","strong mustang", "red barn", "wheat field"],
 "relationships": [[0, "embraces", 1], [2, "standing in", 3]] },
 "answers": ["woman embracing horse",
             "woman brushing horse",
             "horse standing in field"],
 "type": "scene"},
 {"scene_graph": {"objects": ["old woman","strong mustang", "red barn", "wheat field"],
 "relationships": [[0, "embraces", 1], [2, "standing in", 3]] },
 "answers": ["embracing",
             "brushing",
             "standing"],
 "type": "rel"},

{"scene_graph": {"objects": ["bearded man","steep trail", "bright orange tent",
                         "wooded campsite"],
 "relationships": [[0, "hikes up", 1], [2, "sitting in", 3]] },
 "answers": ["bearded man",
             "man",
             "woman"],
 "type": "obj"},
```

```
{"scene_graph": {"objects": ["bearded man","steep trail", "bright orange tent",
                             "wooded campsite"],
 "relationships": [[0, "hikes up", 1], [2, "sitting in", 3]] },
 "answers": ["man hiking up trail",
             "man sleeping on trail",
             "dog wandering on trail"],
 "type": "scene"},
 {"scene_graph": {"objects": ["bearded man","steep trail", "bright orange tent",
                             "wooded campsite"],
 "relationships": [[0, "hikes up", 1], [2, "sitting in", 3]] },
 "answers": ["hiking",
             "sleeping",
             "wandering"],
 "type": "rel"},

{"scene_graph": {"objects": ["yellow tulips","patterned vase", "calico cat",
                               "coffee table"],
 "relationships": [[0, "placed in", 1], [2, "sleeping on", 3]] },
 "answers": ["calico cat",
             "cat",
             "lion"],
 "type": "obj"},
{"scene_graph": {"objects": ["yellow tulips","patterned vase", "calico cat",
                               "coffee table"],
 "relationships": [[0, "placed in", 1], [2, "sleeping on", 3]] },
 "answers": ["tulips sitting in vase",
             "tulips reclining in vase",
             "flowers in a pot"],
 "type": "scene"},
 {"scene_graph": {"objects": ["yellow tulips","patterned vase", "calico cat",
                             "coffee table"],
 "relationships": [[0, "placed in", 1], [2, "sleeping on", 3]] },
 "answers": ["sitting",
             "reclining in",
             "in"],
 "type": "rel"},

{"scene_graph": {"objects": ["female musician","clarinet", "large crowd", "concert hall"],
 "relationships": [[0, "plays", 1], [2, "listens in", 3]] },
 "answers": ["clarinet",
             "musical instrument",
             "television"],
 "type": "obj"},
{"scene_graph": {"objects": ["female musician","clarinet", "large crowd", "concert hall"],
 "relationships": [[0, "plays", 1], [2, "listens in", 3]] },
 "answers": ["crowd listens in hall",
             "crowd eats in hall",
             "a person sleeping on a couch"],
 "type": "scene"},
 {"scene_graph": {"objects": ["female musician","clarinet", "large crowd", "concert hall"],
 "relationships": [[0, "plays", 1], [2, "listens in", 3]] },
 "answers": ["listens",
             "eats",
```

```
            "sleeping on"],
 "type": "rel"},

{"scene_graph": {"objects": ["bright coffee cup","espresso",
                             "stainless steel appliances", "kitchen"],
 "relationships": [[0, "filled with", 1], [2, "standing in", 3]] },
 "answers": ["coffee cup",
             "glass",
             "bowl"],
 "type": "obj"},
{"scene_graph": {"objects": ["bright coffee cup","espresso",
                             "stainless steel appliances", "kitchen"],
 "relationships": [[0, "filled with", 1], [2, "standing in", 3]] },
 "answers": ["appliance standing in kitchen",
             "appliance eating in kitchen",
             "a television located in the bedroom"],
 "type": "scene"},
 {"scene_graph": {"objects": ["bright coffee cup","espresso",
                             "stainless steel appliances", "kitchen"],
 "relationships": [[0, "filled with", 1], [2, "standing in", 3]] },
 "answers": ["standing",
             "eating",
             "located in"],
 "type": "rel"},

{"scene_graph": {"objects": ["tabby cat","wooden stool", "purple couch", "striped rug"],
 "relationships": [[0, "lounging on", 1], [2, "placed on", 3]] },
 "answers": ["tabby cat",
             "cat",
             "dog"],
 "type": "obj"},
{"scene_graph": {"objects": ["tabby cat","wooden stool", "purple couch", "striped rug"],
 "relationships": [[0, "lounging on", 1], [2, "placed on", 3]] },
 "answers": ["furniture on striped rug",
             "furniture swimming on striped run",
             "a television located in the bedroom"],
 "type": "scene"},
 {"scene_graph": {"objects": ["tabby cat","wooden stool", "purple couch", "striped rug"],
 "relationships": [[0, "lounging on", 1], [2, "placed on", 3]] },
 "answers": ["on",
             "swimming",
             "located in"],
 "type": "rel"},

{"scene_graph": {"objects": ["butternut squash","ground", "corn", "verdant field"],
 "relationships": [[0, "laying on", 1], [2, "standing in", 3]] },
 "answers": ["butternut squash",
             "vegetable",
             "fruit"],
 "type": "obj"},
{"scene_graph": {"objects": ["butternut squash","ground", "corn", "verdant field"],
 "relationships": [[0, "laying on", 1], [2, "standing in", 3]] },
 "answers": ["corn standing in field",
```

```
            "corn running in field",
            "a tree laying in the woods"],
 "type": "scene"},
 {"scene_graph": {"objects": ["butternut squash","ground", "corn", "verdant field"],
 "relationships": [[0, "laying on", 1], [2, "standing in", 3]] },
 "answers": ["standing",
            "running",
            "laying in"],
 "type": "rel"},

{"scene_graph": {"objects": ["sparrows","bird feeder", "fat black squirrel",
                             "avocado tree"],
 "relationships": [[0, "flocking to", 1], [2, "hanging from", 3]] },
 "answers": ["black squirrel",
            "animal",
            "fruit"],
 "type": "obj"},
{"scene_graph": {"objects": ["sparrows","bird feeder", "fat black squirrel",
                             "avocado tree"],
 "relationships": [[0, "flocking to", 1], [2, "hanging from", 3]] },
 "answers": ["birds hovering at feeder",
            "birds swimming at feeder",
            "a tree with birds"],
 "type": "scene"},
 {"scene_graph": {"objects": ["sparrows","bird feeder", "fat black squirrel",
                             "avocado tree"],
 "relationships": [[0, "flocking to", 1], [2, "hanging from", 3]] },
 "answers": ["hovering",
            "swimming",
            "with"],
 "type": "rel"},

{"scene_graph": {"objects": ["tortoiseshell cat","blue polka dot rug",
                             "brown mouse", "baseboard"],
 "relationships": [[0, "lounging on", 1], [2, "runs along", 3]] },
 "answers": ["polka dot rug",
            "rug",
            "wooden floor"],
 "type": "obj"},
{"scene_graph": {"objects": ["tortoiseshell cat","blue polka dot rug",
                           "brown mouse", "baseboard"],
 "relationships": [[0, "lounging on", 1], [2, "runs along", 3]] },
 "answers": ["mouse runs along wall",
            "mouse swims along wall",
            "a cat with mouse"],
 "type": "scene"},
 {"scene_graph": {"objects": ["tortoiseshell cat","blue polka dot rug",
                             "brown mouse", "baseboard"],
 "relationships": [[0, "lounging on", 1], [2, "runs along", 3]] },
 "answers": ["runs",
            "swims",
            "with"],
 "type": "rel"},
```

```
{"scene_graph": {"objects": ["sparking ocean","rugged cliff", "old tree",
                             "playful children"],
 "relationships": [[0, "crashing against", 1], [2, "shading", 3]] },
 "answers": ["playful children",
             "people",
             "elf"],
 "type": "obj"},
{"scene_graph": {"objects": ["sparking ocean","rugged cliff", "old tree",
                             "playful children"],
 "relationships": [[0, "crashing against", 1], [2, "shading", 3]] },
 "answers": ["ocean crashing against cliffs",
             "ocean swimming near cliffs",
             "a whale in the ocean"],
 "type": "scene"},
 {"scene_graph": {"objects": ["sparking ocean","rugged cliff", "old tree",
                             "playful children"],
 "relationships": [[0, "crashing against", 1], [2, "shading", 3]] },
 "answers": ["crashing",
             "swimming",
             "in"],
 "type": "rel"},

{"scene_graph": {"objects": ["roaring lion","gazelle", "snow-capped mountain",
                             "serene valley"],
 "relationships": [[0, "hunting", 1], [2, "overlooking", 3]] },
 "answers": ["roaring lion",
             "cat",
             "tiger"],
 "type": "obj"},
{"scene_graph": {"objects": ["roaring lion","gazelle", "snow-capped mountain",
                             "serene valley"],
 "relationships": [[0, "hunting", 1], [2, "overlooking", 3]] },
 "answers": ["mountains near a valley",
             "mountains shouting at valley",
             "a mountain with trees"],
 "type": "scene"},
 {"scene_graph": {"objects": ["roaring lion","gazelle", "snow-capped mountain",
                             "serene valley"],
 "relationships": [[0, "hunting", 1], [2, "overlooking", 3]] },
 "answers": ["near",
             "shouting",
             "with"],
 "type": "rel"},

{"scene_graph": {"objects": ["crowded market","colorful produce", "delivery truck",
                             "busy building"],
 "relationships": [[0, "selling", 1], [2, "parked in front of", 3]] },
 "answers": ["delivery truck",
             "car",
             "boat"],
 "type": "obj"},
{"scene_graph": {"objects": ["crowded market","colorful produce", "delivery truck",
```

```
                                "busy building"],
 "relationships": [[0, "selling", 1], [2, "parked in front of", 3]] },
 "answers": ["market selling vegetables",
             "market running vegetables",
             "a man eating vegetables"],
 "type": "scene"},
 {"scene_graph":{"objects": ["crowded market","colorful produce", "delivery truck",
                                "busy building"],
 "relationships": [[0, "selling", 1], [2, "parked in front of", 3]] },
 "answers": ["selling",
             "running",
             "eating"],
 "type": "rel"},

{"scene_graph": {"objects": ["playful puppy","fluffy toy", "cozy living room",
                                "plush furniture"],
 "relationships": [[0, "chewing on", 1], [2, "adorned with", 3]] },
 "answers": ["living room",
             "room",
             "bathroom"],
 "type": "obj"},
{"scene_graph": {"objects": ["playful puppy","fluffy toy", "cozy living room",
                                "plush furniture"],
 "relationships": [[0, "chewing on", 1], [2, "adorned with", 3]] },
 "answers": ["dog chewing on toy",
             "dog running with toy",
             "dog swimming in water"],
 "type": "scene"},
 {"scene_graph": {"objects": ["playful puppy","fluffy toy", "cozy living room",
                                   "plush furniture"],
 "relationships": [[0, "chewing on", 1], [2, "adorned with", 3]] },
 "answers": ["chewing",
             "running",
             "swimming"],
 "type": "rel"},

{"scene_graph": {"objects": ["silent library","ancient books", "youth",
                                "high tech computers"],
 "relationships": [[0, "filled with", 1], [2, "programming", 3]] },
 "answers": ["library",
             "room",
             "bathroom"],
 "type": "obj"},
{"scene_graph": {"objects": ["silent library","ancient books", "youth",
                                "high tech computers"],
 "relationships": [[0, "filled with", 1], [2, "programming", 3]] },
 "answers": ["person programming computer",
             "person eating at computer",
             "person throwing computer"],
 "type": "scene"},
 {"scene_graph": {"objects": ["silent library","ancient books", "youth",
                                   "high tech computers"],
 "relationships": [[0, "filled with", 1], [2, "programming", 3]] },
```

```
 "answers": ["programming",
             "eating",
             "throwing"],
 "type": "rel"},

{"scene_graph": {"objects": ["silver moon","shimmery lake", "white swan", "dark waters"],
 "relationships": [[0, "casting light on ", 1], [2, "gliding across", 3]] },
 "answers": ["white swan",
             "swan",
             "gnome"],
 "type": "obj"},
{"scene_graph": {"objects": ["silver moon","shimmery lake", "white swan", "dark waters"],
 "relationships": [[0, "casting light on ", 1], [2, "gliding across", 3]] },
 "answers": ["moon shining on lake",
             "moon sleeps on lake",
             "moon climbs on lake"],
 "type": "scene"},
 {"scene_graph": {"objects": ["silver moon","shimmery lake", "white swan", "dark waters"],
 "relationships": [[0, "casting light on ", 1], [2, "gliding across", 3]] },
 "answers": ["shining",
             "sleeps",
             "climbs"],
 "type": "rel"},

{"scene_graph": {"objects": ["green leafy vine","old brick tower", "young woman",
                              "long blonde braid"],
 "relationships": [[0, "climbing", 1], [2, "with", 3]] },
 "answers": ["brick tower",
             "tower",
             "cottage"],
 "type": "obj"},
{"scene_graph": {"objects": ["green leafy vine","old brick tower", "young woman",
                              "long blonde braid"],
 "relationships": [[0, "climbing", 1], [2, "with", 3]] },
 "answers": ["vine climbing brick building",
             "vine sleeping on brick building",
             "moon on brick building"],
 "type": "scene"},
 {"scene_graph": {"objects": ["green leafy vine","old brick tower", "young woman",
                              "long blonde braid"],
 "relationships": [[0, "climbing", 1], [2, "with", 3]] },
 "answers": ["climbing",
             "sleeping",
             "on"],
 "type": "rel"},

{"scene_graph": {"objects": ["red microwave","granite countertop",
                              "striped curtains", "clear windows"],
 "relationships": [[0, "resting on", 1], [2, "draped over", 3]] },
 "answers": ["striped curtains",
             "curtains",
             "window blinds"],
 "type": "obj"},
```

```
{"scene_graph": {"objects": ["red microwave","granite countertop",
                             "striped curtains", "clear windows"],
 "relationships": [[0, "resting on", 1], [2, "draped over", 3]] },
 "answers": ["microwave on counter",
             "microwave sleeping on counter",
             "microwave eating at counter"],
 "type": "scene"},
 {"scene_graph": {"objects": ["red microwave","granite countertop",
                             "striped curtains", "clear windows"],
 "relationships": [[0, "resting on", 1], [2, "draped over", 3]] },
 "answers": ["on",
             "sleeping",
             "eating"],
 "type": "rel"},

{"scene_graph": {"objects": ["black and white bicycle","tall lamp post",
                             "high school student", "redhead girlfriend"],
 "relationships": [[0, "leaning against", 1], [2, "walking with", 3]] },
 "answers": ["tall lamp post",
             "pole",
             "parking meter"],
 "type": "obj"},
{"scene_graph": {"objects": ["black and white bicycle","tall lamp post",
                             "high school student", "redhead girlfriend"],
 "relationships": [[0, "leaning against", 1], [2, "walking with", 3]] },
 "answers": ["boy walking with redhead girl",
             "boy kissing redhead girl",
             "boy skiing with redhead girl"],
 "type": "scene"},
 {"scene_graph": {"objects": ["black and white bicycle","tall lamp post",
                             "high school student", "redhead girlfriend"],
 "relationships": [[0, "leaning against", 1], [2, "walking with", 3]] },
 "answers": ["walking with",
             "kissing",
             "skiing"],
 "type": "rel"},

{"scene_graph": {"objects": ["curious visitors","modern arts", "marble statue",
                             "domed hall"],
 "relationships": [[0, "viewing", 1], [2, "displayed in", 3]] },
 "answers": ["domed hall",
             "room",
             "bathroom"],
 "type": "obj"},
{"scene_graph": {"objects": ["curious visitors","modern arts", "marble statue",
                             "domed hall"],
 "relationships": [[0, "viewing", 1], [2, "displayed in", 3]] },
 "answers": ["people viewing paintings",
             "people painting paintings",
             "people eating people"],
 "type": "scene"},
 {"scene_graph": {"objects": ["curious visitors","modern arts", "marble statue",
                     "domed hall"],
```

```
 "relationships": [[0, "viewing", 1], [2, "displayed in", 3]] },
 "answers": ["viewing",
             "painting",
             "eating"],
 "type": "rel"},

{"scene_graph": {"objects": ["dark movie theatre","black and white movie",
                             "curly haired man", "buttered popcord"],
 "relationships": [[0, "screening", 1], [2, "eating", 3]] },
 "answers": ["curly haired man",
             "man",
             "dog"],
 "type": "obj"},
{"scene_graph": {"objects": ["dark movie theatre","black and white movie",
                             "curly haired man", "buttered popcord"],
 "relationships": [[0, "screening", 1], [2, "eating", 3]] },
 "answers": ["theatre showing movie",
             "theatre eating movie",
             "movie sleeping in theatre"],
 "type": "scene"},
 {"scene_graph": {"objects": ["dark movie theatre","black and white movie",
                                  "curly haired man", "buttered popcord"],
 "relationships": [[0, "screening", 1], [2, "eating", 3]] },
 "answers": ["showing",
             "eating",
             "sleeping"],
 "type": "rel"},

{"scene_graph": {"objects": ["pristine snow","dark forest", "playful kinds", "snowman"],
 "relationships": [[0, "covering", 1], [2, "building", 3]] },
 "answers": ["snowman",
             "mound of snow",
             "snowball"],
 "type": "obj"},
{"scene_graph": {"objects": ["pristine snow","dark forest", "playful kinds", "snowman"],
 "relationships": [[0, "covering", 1], [2, "building", 3]] },
 "answers": ["snow covering forest",
             "snow swimming in forest",
             "forest covered in water"],
 "type": "scene"},
 {"scene_graph": {"objects": ["pristine snow","dark forest", "playful kinds", "snowman"],
 "relationships": [[0, "covering", 1], [2, "building", 3]] },
 "answers": ["covering",
             "swimming",
             "covered"],
 "type": "rel"},

{"scene_graph": {"objects": ["powerful stallion","open prairie", "old cowboy",
                             "leather saddle"],
 "relationships": [[0, "galloping across", 1], [2, "shooting gun from", 3]] },
 "answers": ["cowboy",
             "boy",
             "woman"],
```

```
 "type": "obj"},
{"scene_graph": {"objects": ["powerful stallion","open prairie", "old cowboy",
                             "leather saddle"],
 "relationships": [[0, "galloping across", 1], [2, "shooting gun from", 3]] },
 "answers": ["horse running on prairie",
             "horse eating on prairie",
             "horse standing on prairie"],
 "type": "scene"},
 {"scene_graph": {"objects": ["powerful stallion","open prairie", "old cowboy",
                                "leather saddle"],
 "relationships": [[0, "galloping across", 1], [2, "shooting gun from", 3]] },
 "answers": ["running",
             "eating",
             "standing"],
 "type": "rel"},

{"scene_graph": {"objects": ["sleek black cat","sun-drenched windowsill",
                             "playful kitten", "ball of yarn"],
 "relationships": [[0, "lounging on", 1], [2, "pouncing on", 3]] },
 "answers": ["black cat",
             "cat",
             "dog"],
 "type": "obj"},
{"scene_graph": {"objects": ["sleek black cat","sun-drenched windowsill",
                             "playful kitten", "ball of yarn"],
 "relationships": [[0, "lounging on", 1], [2, "pouncing on", 3]] },
 "answers": ["kitten pouncing on ball of yarn",
             "kitten running with ball of yarn",
             "cat eating food"],
 "type": "scene"},
 {"scene_graph": {"objects": ["sleek black cat","sun-drenched windowsill",
                             "playful kitten", "ball of yarn"],
 "relationships": [[0, "lounging on", 1], [2, "pouncing on", 3]] },
 "answers": ["pouncing",
             "running",
             "eating"],
 "type": "rel"},

{"scene_graph": {"objects": ["busy professional","laptop keyboard",
                             "man in suit", "espresso"],
 "relationships": [[0, "typing on", 1], [2, "drinking ", 3]] },
 "answers": ["man in suit",
             "man",
             "child"],
 "type": "obj"},
{"scene_graph": {"objects": ["busy professional","laptop keyboard",
                             "man in suit", "espresso"],
 "relationships": [[0, "typing on", 1], [2, "drinking ", 3]] },
 "answers": ["person using laptop",
             "person throwing laptop",
             "persion running with laptop"],
 "type": "scene"},
 {"scene_graph": {"objects": ["busy professional","laptop keyboard",
```

```
                                    "man in suit", "espresso"],
  "relationships": [[0, "typing on", 1], [2, "drinking ", 3]] },
  "answers": ["using",
              "throwing",
              "running"],
  "type": "rel"}
}
```