GELDA: A generative language annotation framework to reveal visual biases in image generators

Krish Kabra
Rice University
kk80@rice.edu

Kathleen M. Lewis
Massachusetts Institute of Technology
kmlewis@mit.edu

Guha Balakrishnan
Rice University
guha@rice.edu

Abstract

In this work, we propose GELDA, an automatic framework that leverages large language models (LLMs) and vision-language models (VLMs) to reveal visual biases in image generators. GELDA takes a user-defined caption describing the generated images (e.g., "a photo of a face," "a photo of a living room") and uses an LLM to hierarchically generate domain-specific attributes. GELDA then uses the LLM to select which VLM from a pre-defined set is most appropriate to annotate each attribute. To demonstrate GELDA's capabilities, we present results revealing biases of both text-to-image diffusion models (Stable Diffusion XL) and generative adversarial networks (StyleGAN2). While GELDA is not intended to completely replace human annotators, especially for sensitive attribute annotations, it can serve as a complementary tool to help humans analyze image generation models in a cheap, low-effort, and flexible manner. GELDA is available at https://github.com/krishk97/gelda.

1. Introduction

Image generation models are known to learn spurious correlations that exacerbate bias [2, 15, 17, 22]. Given the rapid advancement and deployment of such models, tools to evaluate model bias are essential for responsible usage. Traditionally, studies assessing image generation bias examine a select number of attributes compiled by researchers, which humans subsequently annotate for a sampled set of generated images. While the second step (annotation) is clearly moving rapidly towards automation with the various advances in vision foundation models [1, 3, 7, 11, 13, 16, 21, 22, 26], the first step (attribute selection) remains largely human-centered. This raises a subtle issue: the evaluation process is limited to the attributes decided upon by researchers, which can leave unforeseen bias *blind spots*. For example, while several important studies have investigated biases of text-to-image models with respect to protected attributes (e.g. age, race, gender) [2, 17, 22], we ask what other biases exist in these models beyond demographic attributes? Moreover, what biases exist for non-human image generations, such as generations of living rooms or animals?

To overcome this issue of constrained bias analysis, we propose a method, called GELDA (for **GE**nerative **L**anguage-based **D**ataset **A**nnotation), that can automatically propose and annotate a diverse set of domain-specific attributes for images sampled from a generative model, thereby providing more general insights into their attribute distributions and potential biases. The key insight behind GELDA is that generative large language models (LLMs) like GPT [5, 18] capture a significant amount of world knowledge [19] and can serve as priors [25] for linking domains to their related attributes. In addition, recent work has demonstrated the effectiveness of using LLMs to select downstream models for given tasks [8]. Therefore, we posit that LLMs may be used to automatically curate a rich set of relevant, domain-specific attributes and select vision models suited to the "type" of each attribute (for example, attributes related to objects are suited for object detectors, whereas holistic image attributes, like "color scheme" or "style", are suited for image-text matching models).

Provided a user-specified domain, GELDA queries an LLM (GPT in our experiments) for semantic categories (e.g., living room furniture and color scheme) and attributes per category (e.g., couch and coffee table for the furniture category) that can visually distinguish images from that domain. Second, we use vision-language models (VLMs) to annotate the generated attributes for the images conditioned on the attribute labels. We use a zero-shot captioning model (BLIP [13]) to annotate attributes related to image-level concepts (e.g., background setting, style), and a text-guided object grounding algorithm (OWLv2 [16]) to annotate attributes related to object-level concepts (e.g., object and part detection). GELDA is automatic with the exception of a few low-cost user inputs (e.g., domain caption, number of desired categories/attributes).

We demonstrate GELDA's capabilities on synthetic im-

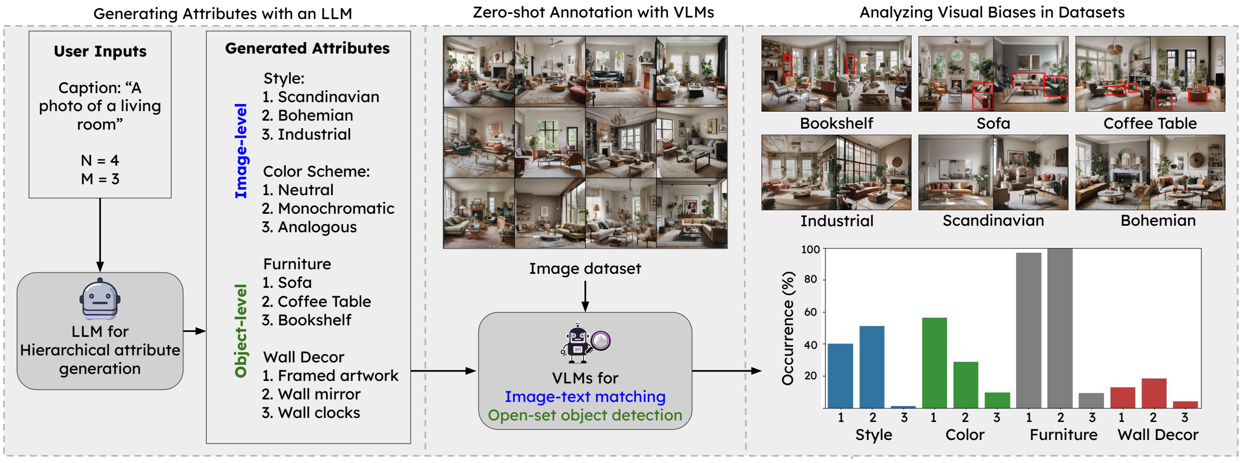

Figure 1. **Overview of GELDA.** Given a user-specified domain in the form of a caption, GELDA first queries an LLM to generate a set of visual attributes to annotate images specific to that domain that. The querying method is hierarchical, in that GELDA prompts the LLM to first generate N attribute categories, then generate M labels per attribute category, and finally describe whether each attribute is object-level or image-level. In the second stage, GELDA uses pre-trained VLMs to automatically annotate the generated attributes for sampled images. We use the LLM to assign all image-level attributes to a VLM tuned for image-text matching, and all object-level attributes to a VLM for open-vocabulary object detection. Once GELDA has identified and annotated visual attributes, we can then analyze the visual biases of the image generator.

age data produced by both text-to-image models (Stable Diffusion [20]) and generative adversarial networks (StyleGAN2 [10]). GELDA reveals that living rooms generated by Stable Diffusion almost always have neutral or monochromatic color schemes and contain coffee tables, sofas, area rugs, and throw pillows. Moreover, GELDA confirms that StyleGAN2 amplifies biases from its training set for both human faces (FFHQ [9]) and dogs (AFHQ [6]). Finally, we present some of GELDA's limitations and draw conclusions regarding its safe. While there is no substitute for human ground truth, an annotation method that trades off accuracy for flexibility and automation would enable practitioners to quickly and effortlessly gather insights about their generative models.

2. Methods

Our goal is to take a user-specified domain along with a set of images S from that domain, and automatically produce attribute annotations for each image in S from a variety of in-domain categories. Using these attributes, we can then perform bias analyses of S. There are two key challenges to this task: (1) automatically obtaining a list of relevant categories and attributes for the specified domain, and (2) automatically choosing the appropriate model for evaluating each image-attribute pair. We propose a framework (see Fig. 1) that addresses both of these challenges.

Our insight for the first challenge is that large language models (LLMs) are adept at linking concepts to one another [19, 25]. We therefore query an LLM for a list of domain categories along with their associated attributes with careful prompting. To address the second challenge, we observe that vision-language models (VLMs) offer a powerful means of performing such evaluations like zero-shot image classification [21] and object grounding [16] from text input alone. The key challenge is determining which VLM to use for a given attribute. Certain image-level attributes like style or color scheme are better suited for image-text matching (ITM) models, whereas determining the presence of an object like a couch is better suited for open-vocabulary object detectors (OVODs). We again use the LLM, this time to provide a decision into the attribute type, and automatically choose the appropriate VLM based on a pre-specified list of VLMs for each attribute type. We describe our method further in the following sections.

2.1. Attribute generation with an LLM

We use an LLM to generate attributes in a hierarchical fashion by querying the LLM for categories, followed by querying attribute examples per category. We use this hierarchical form for several reasons. First, we empirically find that querying the LLM directly for attributes results in poor coverage of visual concepts. Second, breaking up the prediction as a "chain" is known to be a successful strategy for controlling LLMs towards more human-like reasoning [24]. Third, this approach allows the user control over the number of categories and attributes per category that they desire. First, the user provides a prompt query $Q1$ of the form:

> $Q1$: "What are N attribute categories that can be used to visually distinguish images described by

> the caption `caption`?”,

where N is a number chosen by the user and `caption` is a word or phrase describing the data domain (e.g., “birds” or “a headshot photo of a person”). Second, for each of the categories {`category1, ..., categoryN`} returned by $Q1$, we obtain attribute labels with query $Q2$:

> $Q2$: “What are M different examples of the category `category` that can be used to distinguish images described by the caption `caption`?”,

where M is again chosen by the user. Lastly, we determine whether each of the N attribute categories relates to image-level or object-level concepts with query $Q3$:

> $Q3$: “Are {`att1, ..., attM`} examples of objects or items? Answer with a yes or no. Explain your answer.”,

where {`att1, att2, ...attM`} is the list of M generated attributes for a category. We require a binary yes or no answer in order to automatically filter the response into one of the two appropriate downstream models. Requiring an explanation pushes the model to provide more accurate answers, as demonstrated in prior work [24].

Dealing with stochasticity: Auto-regressive LLMs are stochastic in that they can produce different outputs given the same prompt. While stochasticity helps capture the full output distribution, determinism is helpful for reproducibility. To obtain high-quality attribute labels that are mostly consistent across experiments, we perform the queries in the previous section several times per prompt, and pick the N and M most frequently labeled categories and attributes.

2.2. Zero-shot annotation with VLMs

We assume access to pretrained VLMs that take input images and text captions and can perform annotation. In our experiments, we use two VLMs – one for image-text matching (ITM) and one for open-vocabulary object detection (OVOD). To convert LLM-generated attributes into input captions for the VLMs, we prompt the LLM to modify the initial user caption by incorporating the specified attributes.

OVOD models output bounding boxes and detection scores, allowing us to label an attribute if its detection score is simply above a threshold α. Output values of current ITM models are less predictable because they are trained with a hard negative mining strategy [12], making it difficult to set a constant threshold. Instead, we compute ITM scores for the M attribute text captions and a generic “base” reference caption describing the domain (same as the one used in query $Q1$, see Sec. 2.1). Finally, we select the highest-scoring caption among the M attributes, and label that attribute as present if it is greater than the base caption score. This process essentially performs multiclass classification.

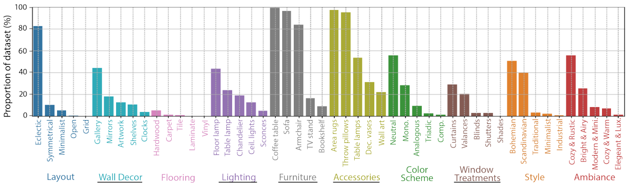

Figure 2. **Distribution of annotated attributes returned by GELDA for the SD living rooms.** Bars are grouped by attribute categories in different colors, and attribute names have been shortened for brevity. Category corresponds with attributes determined to be object-level. Certain attributes are prominent in the generated images, such as coffee tables and sofas for furniture, throw pillows and area rugs for accessories, neutral and monochromatic hues for color schemes, and Bohemian and Scandinavian styles.

3. Experiments and Results

We demonstrate using GELDA to evaluate biases in two popular image generation models. We use the public Stable Diffusion (SD) XL model [20] to generate 1,024 synthetic images using the caption “a photo of a living room”, and we use the public StyleGAN2 (SG2) models [10] trained on the FFHQ [9] and AFHQ [6] datasets to generate 10,000 images each of human faces and dogs (with truncation $\psi = 0.7$ [4]). To enable GELDA, we use the following publicly available models: GPT-3.5 for chat completion[1], BLIP (ViT-L/14) [13] for ITM, and OWLv2 (ViT-L/14) [16] for OVOD using a threshold of $\alpha = 0.3$. We heuristically find $N = 10$ categories and $M = 5$ attributes to provide good concept coverage, and, therefore, use these values for attribute generation.

We plot a histogram of generated attributes for SD living rooms in Fig. 2. Several categories have uneven attribute distributions. For example, over 90% of generated living rooms contain a coffee table, sofa, area rug, or throw pillows. Furthermore, less than 10% contain wall sconces, bookshelves, blinds, shutters, or shades. The majority of living rooms also have an “eclectic” layout, a “neutral” color scheme,” a “Bohemian” or “Scandinavian” style, and a “cozy and rustic” ambiance. We observe that BLIP struggles to annotate generated flooring attributes, with the majority of images receiving a higher score for the base caption.

Next, we analyze differences in attribute distributions between StyleGAN2 generators and their training distributions (FFHQ and AFHQ-Dogs datasets). We show the differences in attribute frequencies computed by GELDA in Fig. 3. The analysis demonstrates SG2 amplifies bias –for both SG2 Faces and SG2 Dogs, the majority attribute per category in the training dataset almost always has an exacerbated majority in the generated dataset. This is shown in the plot as a negative difference (i.e. higher frequency

[1] Model version: `gpt-3.5-turbo-1106`

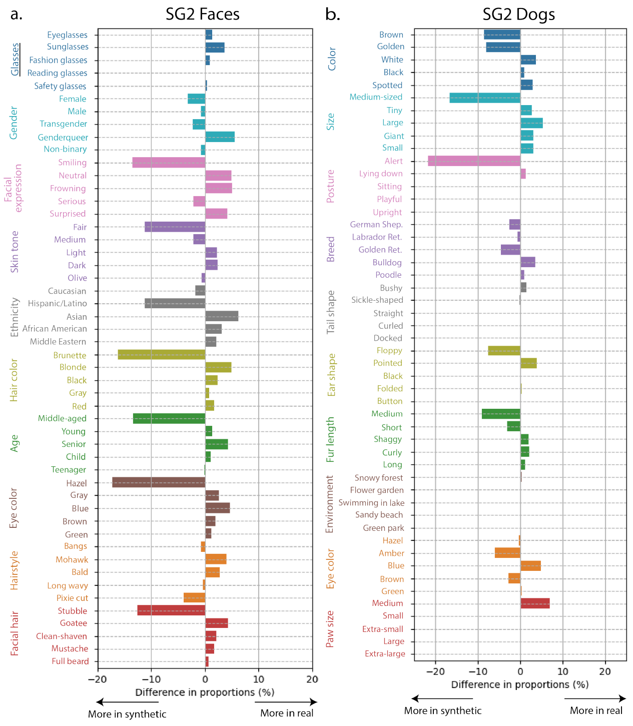

Figure 3. **Comparisons of attribute bias of synthetic StyleGAN2 (SG2) image generators with respect to their training distributions.** (a.) SG2 Faces vs. FFHQ. (b.) SG2 Dogs vs. AFHQ. The attributes are ordered from top to bottom in each category by descending frequency in the training dataset. SG2 amplifies bias – the most popular attributes in the training dataset for each category have an even greater majority in the generated dataset, as seen by large negative differences. <u>Category</u> corresponds with attributes determined to be object-level.

in the generated dataset) for several of the first attributes in each category (the attributes are sorted in order of descending frequency in the training dataset). For example, in SG2 Faces, over 10% more images contain a smiling facial expression, fair skin tone, brunette hair color, middle-aged appearance, hazel eye color, and stubble facial hair in comparison to its corresponding training dataset. For SG2 Dogs, over 20% more images contain a dog with an "alert" posture and over 10% more contain a medium-sized dog in comparison to its training dataset.

4. Discussion

We propose GELDA, the first automated framework leveraging the power of large language and vision-language models to suggest and annotate attributes for bias evaluation of image generation models. The evaluation of image generation algorithms, particularly large text-to-image models, is of great interest to the vision community. Given that a model like Stable Diffusion can generate any image distribution describable by text, it is desirable to also develop analysis algorithms like GELDA that are equally flexible. Results demonstrate that Stable Diffusion can skew color schemes, accessories, and furniture when generating "a photo of a living room." Such insight can help practitioners engineer their prompts to steer away from unwanted biased attributes. Results also demonstrate that GELDA can measure bias amplifications of a generator with respect to its training distribution, such as with StyleGAN2-produced faces and dogs.

GELDA has several limitations. First, it is only as good as its constituent LLM and VLMs, which have their own systematic errors and biases. While VLMs have improved tremendously in the past several years, they are still far from perfect on high-level semantics beyond object recognition [23, 27]. In addition, GPT can fail to recall a number of important attributes. The combination of these errors indicates that a method like GELDA cannot simply replace humans in an annotation pipeline in terms of attribute coverage or annotation accuracy. Instead, GELDA will be most useful as a fast, flexible, and automated tool to perform coarse dataset analysis, complementing existing annotations. Second, our current implementation selects one image-level attribute per category for an image (multiclass classification), though an image can contain multiple attributes together (e.g. living rooms can have both monochromatic and neutral color schemes). Third, we evaluated GELDA on image generations with "contained" domains focusing on one type of scene/object. Image generations of complex natural scenes like MS-COCO [14] would pose challenges in attribute generation (a compact prompt cannot describe arbitrary natural scenes) and image-level attribute annotations, although object-level annotations should be relatively unharmed.

4.1. Ethics and responsible use

GELDA inherits the biases of its LLM/VLM models. Biases of the LLM will mainly result in missed attribute categories which, while undesirable, are not as problematic as VLM biases. VLM biases can result in incorrect annotations, thereby skewing bias analyses. These inaccuracies may be particularly harmful when dealing with human-centered datasets like faces for which these models are not tuned for. A user should therefore always exercise caution and visually inspect image annotation results to confirm reasonable labels and understand the limitations of the VLMs. We recommend using GELDA not as a replacement to human perceptual ground truth, but as an efficient, flexible, and low-cost method to complement human annotation in bias benchmarking.

References

[1] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4502–4511, 2019. 1

[2] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023. 1

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 2, 3

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[8] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 1

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 3

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1

[12] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 3

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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, 2014. 4

[15] Vongani H Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A Efros, Angjoo Kanazawa, and Devin Guillory. Studying bias in gans through the lens of race. In *European Conference on Computer Vision*, pages 344–360. Springer, 2022. 1

[16] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 3

[17] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808, 2023. 1

[18] OpenAI. Gpt-4 technical report, 2023. 1

[19] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473. Association for Computational Linguistics, 2019. 1, 2

[20] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[22] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*, 2023. 1

[23] Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[24] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2, 3

[25] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. 1, 2

[26] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

[27] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 4