Appendix: Generating Images with Multimodal Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

We detail current limitations of GILL, and suggest possible directions to alleviate this in future work. We also describe the broader impact of our work, including possible applications, risks, and intended uses. Finally, we provide more quantitative and qualitative evaluations, including results on deciding whether to retrieve or generate, results on the effect of increasing context on VisDial, text-to-image generation results on MS-COCO, and present more qualitative samples from GILL.

A Limitations

GILL relies on an LLM backbone for many of its capabilities. As such, it also inherits many of the limitations that are typical of LLMs. One limitation is the potential for hallucinations [2], where the model generates content that is false or not relevant to the input data. Another limitation of the model in generating text is in repetitions and neural text degeneration [12], where the model generates the same content multiple times. We also observed that the OPT-6.7B model also does not always consistently generate coherent dialogue text.

These limitations may be addressed by techniques that address hallucinations and degenerations in text-only LLMs, or by using improved LLMs that are less prone to these issues. In GILL, we used a 6.7B model. In the future, it will be valuable to scale up the approach with even larger LMs, or those trained with improved objectives [25], instruction finetuning [26] or human feedback [19]. Depending on downstream applications, using models trained explicitly on dialogue data [7] may also be helpful for dialogue capabilities (e.g., deploying multimodal chatbots).

With regards to the visual models, another limitation of our approach is in its limited visual processing. At the moment, we use only $k = 4$ visual vectors to represent each input image (due to computational constraints), which may not capture all the relevant visual information needed for downstream tasks. These vectors are produced by a frozen pre-trained visual encoder, and so the visual information in the vectors is heavily constrained by the pre-training task. As a result, the model may not always process images correctly or in enough detail to produce accurate or high-quality results. However, this limitation can potentially be addressed in the future by scaling up the visual model, using models with varied pre-training objectives that encode more visual information while still being mappable to the hidden space of the LLM, or using more sophisticated visual mappings [1, 15] that can capture a richer set of visual features. Similarly, we observed during inference that our model sometimes does not generate relevant images for certain types of prompts. We attribute this to our finetuning dataset being CC3M, which is relatively small compared to modern large scale image-text datasets [24]. It is likely that training GILLMapper on an even larger corpus of text data will improve its alignment to the image generation backbone.

One of the advantages of our model is that it is modular, and can benefit from stronger visual and language models released in the future. It is likely that it will also benefit from stronger text-to-image generation backbones, or through finetuning the generation backbone rather than just the GILLMapper module. We leave such scaling explorations for future work.

B Broader Impact

AI Assistants Recent advances in dialogue based chatbots have sparked interest in using LLMs for interactive conversational applications. GILL is a multimodal language model capable of processing image and text inputs, and producing image and text outputs. These capabilities may enable a wider range of applications. For example, AI assistants which can produce image and text outputs would be able to answer a wider range of queries, providing visual content when necessary to illustrate certain points. Concrete applications may include creative endeavors (e.g., iteratively refining a generated image with instructions), answering questions that benefit from visual outputs (e.g., describing food items), and more. Scaling GILL and refining it with methods such as reinforcement learning from human feedback (RLHF) [14] are promising directions to improve the capabilities of multimodal AI assistant systems.

Disinformation and Harms Aside from the technical limitations detailed in Sec. A, there are broader societal issues that should be considered with the development of generative models of text and images. LLMs have the potential to generate plausible sounding (but false) text [10, 2], propagating disinformation at scale. As GILL uses an LLM backbone, it is also susceptible to these potential issues. Furthermore, as multimodal generative models which can also produce image content, models such as GILL also introduce potential issues with producing even more convincing disinformation through interleaving text with realistic generated images. As GILL makes use of an image generation backbone, it is also susceptible to the risks that typical text-to-image generation models introduce, such as generating false images of real people. These harms may possibly be mitigated by introducing watermarking into generated images [17, 28], or by deploying systems to detect generated images [5].

Bias and Safety GILL makes use of pretrained LLMs and multimodal models (such as CLIP [20] and Stable Diffusion [22]), which are trained on large, noisy, Internet-scraped data (such as LAION-400M [24]). Due to their curation process, these datasets often contain undesired biases, malignant stereotypes (see [3] for a comprehensive discussion on large scaled multimodal datasets). One advantage of GILL is that it is efficient to train and completely *modular*, allowing its components (i.e., the LLM, visual encoder, or image generator) to be swapped out for other pretrained models (for example, models which have been further calibrated to reduce unintended biases).

Intended Uses GILL is a research prototype which showcases possible capabilities of multimodal language models which can both process and produce image and text outputs. Due to the limitations described above, GILL is not in its current state intended for deployment in practical applications, especially in high risk or sensitive domains without further analysis. At its current model scale (a 6.7B parameter LLM), GILL also lacks many of the abilities of larger language models [4], and applications would likely benefit from increased scaling of the LLM and visual models.

C Deciding to Generate or Retrieve

As detailed in Sec. 3.3 of the main paper, we evaluate several models on the annotated PartiPrompts [27] dataset. Each prompt is annotated with one of two labels: "ret" or "gen", indicating whether image retrieval or image generation produces a more appropriate image for the corresponding prompt. For example, the prompt *"a portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket, flying over the city of Mars."* is labeled as "gen", as there are (understandably) no appropriate images in the CC3M retrieval set, and generation produces a more relevant output. In contrast, *"the geyser Old Faithful"* is labeled as "ret," as there are very relevant candidate images available for this prompt. We evaluate several models for making this decision on the validation set (Tab. 1), evaluating using F1 score given the class imbalance of the dataset (201 "gen", 110 "ret" in the validation set labels):

1. **Baselines:** We measure the F1 score of several baseline methods, which provide a lower bound for how well data-driven approaches can do. We find that always retrieving an image,

Table 1: Results on PartiPrompts for classifying retrieval or generation.

Method	**F1**
Always retrieve	0.267
Always generate	0.389
Random	0.451
Heuristic	0.261 – 0.559
Linear classifier	0.393 – 0.552
Human performance	0.851

always generating an image, or simply deciding randomly (with a prior proportional to class frequencies) achieve F1 scores of 0.267, 0.389, and 0.451 respectively.

2. **Heuristic:** We also consider a simple heuristic which considers the maximum cosine similarity of the retrieval embedding against the entire image candidate set (i.e., the training set of CC3M). We run a grid search from 0 to 1 for possible threshold values. Whenever the maximum cosine similarity is above a threshold, we return "ret" and "gen" otherwise. This achieves an F1 of 0.261 – 0.559, depending on the threshold used (a threshold of 0.5 gives F1 of 0.261).

3. **Linear classifier:** Lastly, we train a linear classifier that takes as input the outputs of the LLM for the `[IMG]` tokens and the maximum cosine similarity. This classifier is trained with the binary cross-entropy loss over the training set of PartiPrompts annotations. This linear classifier achieves an F1 score of between 0.393 – 0.552, depending on the probability threshold used (a threshold of 0.5 gives an F1 score of 0.547).

We use the linear classifier in our final model, as it requires less hyperparameter tuning compared to the heuristic baseline, and performs comparably on quantitative metrics. During generation of qualitative samples (Fig. 1 and Fig. 5 in the main paper), we observed that the linear classifier generally performed well for many prompts, and decided correctly whether to retrieve or generate.

D Qualitative Results

We present further qualitative samples in Fig. 1. We find that GILL is able to process complex text prompts more effectively than Stable Diffusion for many examples in PartiPrompts [27]. On VisDial [8] dialogue inputs, GILL is able to generate more relevant outputs (as measured against groundtruth images). We attribute these improved results to the stronger text representations of the LLM, and the effectiveness of our GILLMapper network.

E Other Evaluations

E.1 Increasing Context on VisDial

GILL leverages an LLM backbone, which allows it to inherit some of the LLM's capabilities, such as improved sensitivity to long input contexts. In the main paper, we showed that GILL can better condition on longer image and text inputs to generate more relevant images for VIST [13]. We run a similar experiment on Visual Dialogue [8], varying the number of dialogue rounds provided as input context to GILL and Stable Diffusion (SD) [22].

The results are presented in Fig. 2. We find that when longer text context is provided to both models, the performance of generating relevant images steadily improves. Interestingly, SD performance plateaus after 6 rounds of dialogue, while GILL continues to improve, outperforming SD when 7 or more rounds of dialogue are provided. These results showcase the improved sensitivity of our model to conditioning on long, dialogue-like text. Despite both approaches using the same image generation backbone, GILL is able to better make use of longer dialogue-text inputs (despite being only finetuned on image caption data).

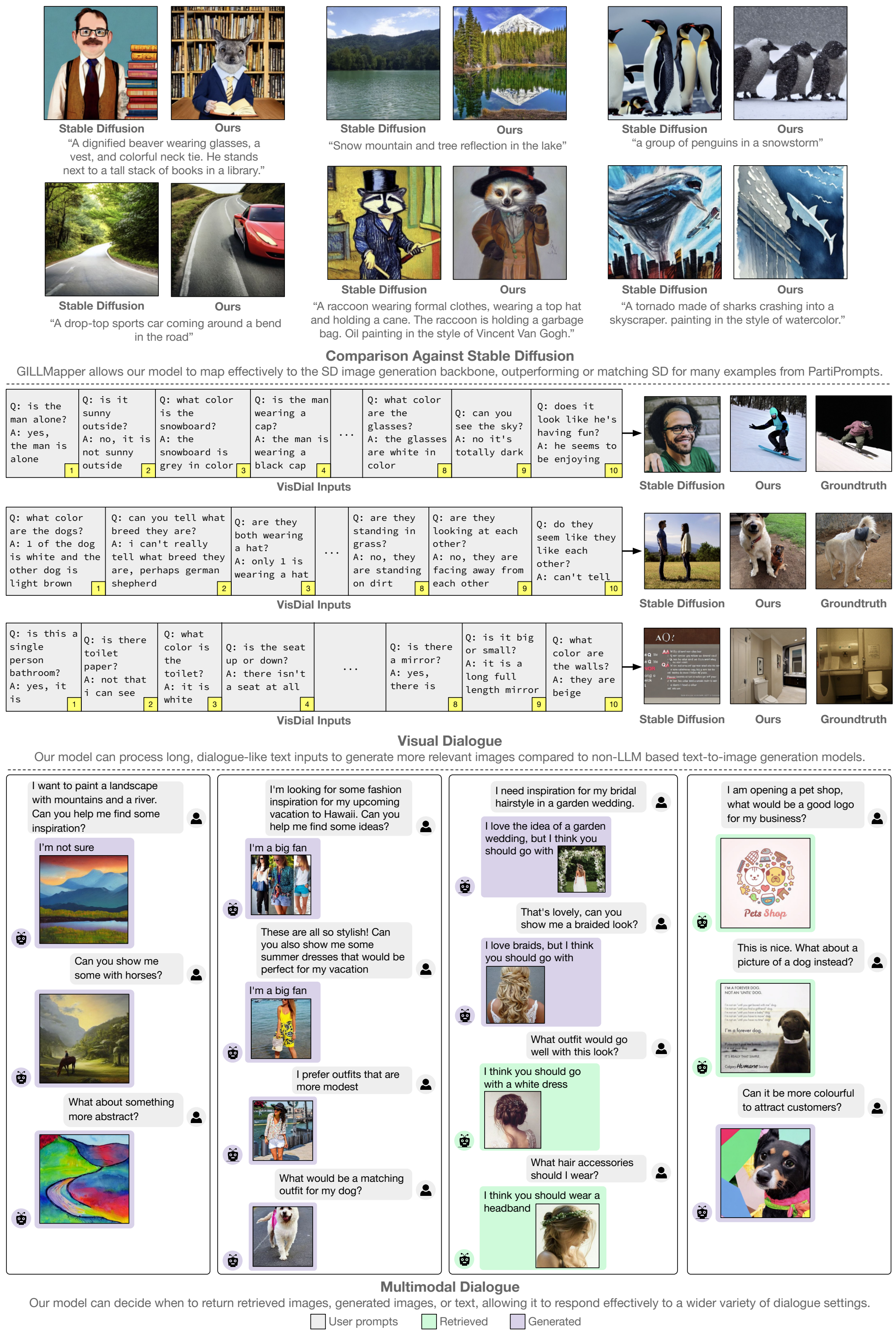

Figure 1: Further qualitative samples from GILL. It is more sensitive to text inputs due to its LLM backbone, and better at processing complex text prompts.

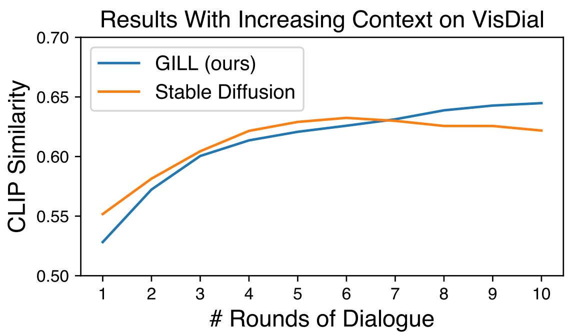

Figure 2: Performance of our model and Stable Diffusion [22] with increasing context for generating VisDial [8] images. Our model is able to better process long dialogue-like text descriptions.

Table 2: Zero-shot FID [11] on the MS-COCO [16] (2014) validation set. 30,000 random samples are used to evaluate all models.

Model	**FID** (↓)
GLIDE [18]	12.24
Make-A-Scene [9]	11.84
DALL-E 2 [21]	10.39
LAFITE2 [29]	8.42
Imagen [23]	7.27
Parti [27]	7.23
Re-Imagen [6]	6.88
SD [22] v1.5	9.22
GILL (ours)	12.2

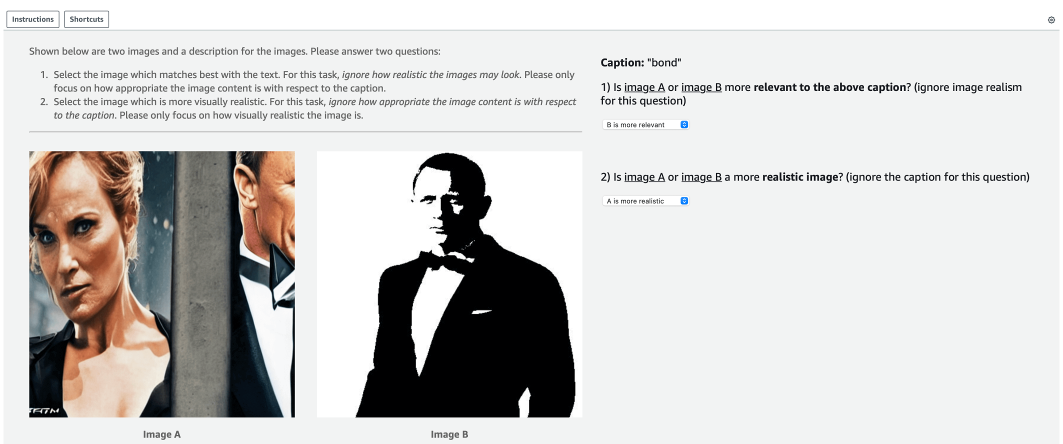

Figure 3: User interface shown to human annotators for annotating PartiPrompts [27] examples.

E.2 Image Generation

In addition to our evaluations on VIST [13] and VisDial [8], we also run evaluations on our model's ability to generate images from MS-COCO [16] captions (Tab. 2). We generate images using 30,000 randomly sampled captions from the MS-COCO (2014) validation set, which is the standard evaluation of text-to-image generation models. We report zero-shot FID scores [11] of our model, Stable Diffusion [22] v1.5 (which we use as our backbone image generator), and several other approaches in Tab. 2. For our generation results and SD results, we use a classifier-free guidance scaling factor of 3.0 and 250 DDIM inference steps. On MS-COCO, our approach achieves a worse FID score than SD (9.22 to 12.2). This is likely because this task does not benefit as much from the LLM backbone, which has not been trained on as many image captions as SD (which exclusively trains on caption-like data). These numbers will likely improve further by finetuning GILL on even more text data (including image captions), which will allow our model to align more closely to the input space of the SD image generator.

F Human Annotation on PartiPrompts

In Sec. 3.3 of the main paper, we described the process of annotating PartiPrompts [27] with per-example labels to retrieve or generate. The interface shown to human annotators is shown in Fig. 3. Annotators are tasked to determine which of two anonymized images are (1) more relevant to the provided prompt, and (2) more realistic. We randomize the order of the two images as well (i.e., the output of the retrieval model shows up 50% of the time as Image A).

We show each example to 5 independent human annotators. For determining whether to label a particular example as "ret" or "gen", we take the majority vote of the 5 annotators on the image relevance question ("Is image A or image B more relevant to the above caption?"), and only keep the examples with an inter-annotator agreement of at least 4/5. This results in approximately 900 examples remaining (out of the 1,632 examples in PartiPrompts). Our annotations will be publicly released to facilitate future evaluations on this task.

We conducted evaluations on the Amazon Mechanical Turk platform with human annotators located in the US and Canada. Annotators were paid at an estimated hourly rate of 15 USD per hour. In total, we spent approximately 326 USD to collect these annotations.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

[2] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[3] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[4] 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. *NeurIPS*, 2020.

[5] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Alexander Binder, and Ngai-Man Cheung. Discovering transferable forensic features for cnn-generated images detection. In *ECCV*, 2022.

[6] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017.

[9] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, 2022.

[10] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *EMNLP*, 2020.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

[12] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *ICLR*, 2020.

[13] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *NAACL-HLT*, 2016.

[14] Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*, 2022. https://huggingface.co/blog/rlhf.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023.

[16] 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. *ECCV*, 2014.

[17] Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar. Distortion agnostic deep watermarking. In *CVPR*, 2020.

[18] Alex 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. *ICML*, 2022.

[19] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[20] 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 *ICML*, 2021.

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[23] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

[24] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[25] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022.

[26] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *ICLR*, 2022.

[27] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022.

[28] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

[29] Yufan Zhou, Chunyuan Li, Changyou Chen, Jianfeng Gao, and Jinhui Xu. Lafite2: Few-shot text-to-image generation. *arXiv preprint arXiv:2210.14124*, 2022.