Generating Contextual Images for Long-Form Text

Avijit Mitra[1], **Nalin Gupta**[2], **Chetan Naik**[2], **Abhinav Sethy**[2], **Kinsey Bice**[2], **Zeynab Raeesy**[2]
[1]University of Massachusetts Amherst, [2]Amazon
avijitmitra@umass.edu, {nalgupta, chetnaik, sethya, bickinse, raeesyzr}@amazon.com

Abstract

We investigate the problem of synthesizing relevant visual imagery from generic long-form text, leveraging Large Language Models (LLMs) and Text-to-Image Models (TIMs). Current Text-to-Image models require short prompts that describe the image content and style explicitly. Unlike image prompts, generation of images from general long-form text requires the image synthesis system to derive the visual content and style elements from the text. In this paper, we study zero-shot prompting and supervised fine-tuning approaches that use LLMs and TIMs jointly for synthesizing images. We present an empirical study on generating images for Wikipedia articles covering a broad spectrum of topic and image styles. We compare these systems using a suite of metrics, including a novel metric specifically designed to evaluate the semantic correctness of generated images. Our study offers a preliminary understanding of existing models' strengths and limitation for the task of image generation from long-form text, and sets up an evaluation framework and establishes baselines for future research.

Keywords: Large Language Model, Diffusion Model, Text-to-Image, Visual Imagery Synthesis

1. Introduction

Recent advances in Text-to-Image models (TIMs) (Rombach et al., 2022; Croitoru et al., 2022) have led to impressive results in producing high-quality images from short prompts describing both the content and style of the target image. However, prompts for generating high-quality images that reflect the intention of the image creator need to be carefully crafted (Rombach et al., 2022; Gu et al., 2023) and suffer from size, format, and style constraints. These limitations on the image generation prompt are partly the result of the small capacity of TIM text encoders, such as the CLIP (Radford et al., 2021) text encoder. Recent approaches such as GILL (Koh et al., 2023) have shown that TIMs can be effectively combined with Large Language Models (LLMs) by introducing and learning translation layers between them. This adoption of off-the-shelf state-of-the-art language models, enables TIMs to benefit from the reasoning and summarization capabilities of LLMs (Zhao et al., 2023). Indeed the work in (Koh et al., 2023) shows that TIMs with LLM integration have proven successful on tasks with a clear and concise visual component, such as visual dialog (Das et al., 2016), and visual storytelling (Huang et al., 2016).

Generating contextually relevant images for generic long-form text presents a unique challenge. Such texts often provide overarching narratives rather than detailed visual descriptors, such as style, background elements, or mood. For instance, a mention in a body of text about Renaissance architecture might be as broad as 'a historical Renaissance-era building', yet for accurate image generation, a more enriched detail like 'a sketch art of a Renaissance-era chateau with stone masonry, ornate spires, set beside a serene river' is essential. In this paper, we study different ways of combining reasoning capabilities of LLMs image generation capabilities of TIMs for contextual image generation from long-form text. We compare results from zero-shot prompting, the GILL model, and different fine-tuned models based on the GILL architecture. We use the Wikipedia-based Image Text dataset, WIT (Srinivasan et al., 2021) to conduct our experiments. Each entry in the dataset contains paragraphs of Wikipedia articles and their corresponding images. The diverse topics covered in Wikipedia articles and image styles provide a strong benchmark to test our approaches. Figure 1 shows two examples of paragraphs from the WIT dataset with the corresponding ground truth image and the image generated from one of our fine-tuned GILL models. We have provided more qualitative examples in Appendix A.

We also introduce a new metric to specifically evaluate the image content disregarding the stylistic factors. The traditional metrics used in the literature are based on CLIP and LPIPS (Zhang et al., 2018) that rely on dense embeddings storing rich information about the images, factoring in both content and style (Koh et al., 2023). However, unlike most existing TIM quality evaluation work, the visual content in our case is not explicitly described in the input text. To focus specifically on the ability of the different approaches to derive the right content in the image, we propose the BLIP-2 similarity. This metrics uses image descriptions produced by BLIP-2 (Li et al., 2023) for the generated and ground truth images and compares the descriptions in the textual domain using metrics such as BERT Score (Zhang et al., 2020), Sentence BERT similarity (Reimers and Gurevych, 2019) and ROUGE-L.

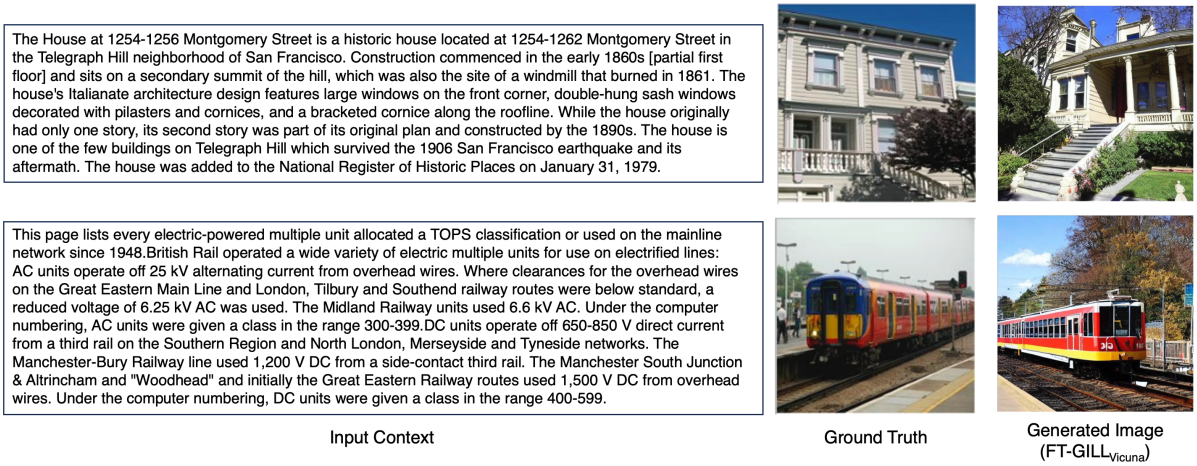


Figure 1: Examples of image generation from long-form text. The text on the left shows examples of paragraphs from the WIT dataset, with the left image being the ground truth image. The right image is synthesized from a fine-tuned GILL model with Vicuna LLM.

2. Background

Generation of contextual images for generic long-form text as studied in this paper, is based on recent research in the fields of language modeling and image generation. Recent research has shown that as the size of LLMs grows to certain critical scales they demonstrate *emergent abilities* of reasoning and abstractive summarization (Wei et al., 2022b). To benefit from these capabilities and provide better understanding of user inputs, LLMs have been utilized in building generative systems such as images from short prompts.

Generating images conditioned from short prompts has been an active area of research for the machine-learning community. Earlier work in this area focused on variants of generative adversarial networks (GAN) such as conditional GAN (Reed et al., 2016), multi-stage GAN (Zhang et al., 2017), attention GAN (Xu et al., 2018), cross-modal contrastive GAN (Zhang et al., 2021), and VQGAN (Yu et al., 2022), among others. In more contemporary research, there has been a shift towards the adoption of transformer-based decoders as a means for text-to-image generation tasks (Ding et al., 2021; Chang et al., 2022; Ramesh et al., 2021). Simultaneously, the application of diffusion models (Ho et al., 2020) for image generation has gained substantial traction, largely attributed to their capacity to generate images of enhanced quality (Nichol et al., 2022; Rombach et al., 2022; Ramesh et al., 2022). As LLMs are integrated into image and other media generation models (Aghajanyan et al., 2022; Zhu et al., 2024; Liu et al., 2023) the text input to these models can be more flexible. Specifically, the GILL model (Koh et al., 2023) we experiment with employs a diffusion model that is tied to an LLM, which contributes to an enriched semantic comprehension of long-form text input. The enhanced capabilities of the LLM lead to significant improvements in tasks such as Visual Dialog and Storytelling (Das et al., 2016; Huang et al., 2016), where the input text descriptions are more complex. An alternative to training the LLM and TIM models jointly is to build an adaptor model that takes the user input and converts into a prompt for TIM models (Hao et al., 2023; Brade et al., 2023)

3. Methods

This section presents four methods that combine the strengths of LLMs and TIMs for effective image generation from long-form narratives.

3.1. Zero-shot Prompting

Traditional TIMs possess a limited context window, making them unsuitable for directly consuming verbose long-form texts. Recognizing this limitation, we pivot to a zero-shot prompting approach using a pretrained LLM. Given a long-form text, T, the LLM is prompted to extract and generate a concise image descriptor, D_{zs}. This descriptor, tailored to fit within the TIM's context window, is subsequently processed by a pretrained TIM to yield an image I_{zs}, as represented by the sequence $T \xrightarrow{\text{LLM}} D_{zs} \xrightarrow{\text{TIM}} I_{zs}$. In our experimental results, we reference the setup where the OPT model (Zhang et al., 2022) serves as the LLM and Stable Diffusion (SD) as the TIM as SD$_{OPT}$. Conversely, when Vicuna (Touvron et al., 2023a) is employed as the LLM, the configuration is denoted as SD$_{Vicuna}$. The advantage of this method is that it offers a structured pathway to harness the capabilities of TIMs for long-form content, without the constraints of their native context window and without the need

for intensive model retraining.

3.2. GILL

Leveraging integrated models can streamline the process of transforming text to images, potentially improving cohesion between textual input and visual output. In this vein, we employ GILL (Koh et al., 2023) — a cohesive framework that synergizes an LLM and a TIM. GILL uses GILLMapper module, a lightweight Transformer conditioned on special learnt `[IMG]` tokens, to learn the mapping from the frozen LLM to the Stable Diffusion generation model. Here, a long-form text T undergoes a direct transformation into an image I_g, represented as $T \xrightarrow{\text{GILL (LLM + TIM)}} I_g$. The primary advantage of this method is its ability to produce images that inherently resonate with the textual narrative, all within a singular, unified model.

3.3. Fine-tuning GILL

While integrated models like GILL are powerful, there's potential to enhance their performance further by fine-tuning specific components, especially when adapting them to domain-specific tasks or new datasets. To test this premise, we utilize the WIT (Srinivasan et al., 2021) dataset to fine-tune the GILL model. This process aims at producing an image I_f from a long-form text T, which can be represented by the sequence $T \xrightarrow{\text{FT-GILL}_{\text{OPT}}} I_f$. Through this fine-tuning process, we aim to optimize GILL's image generation abilities, anticipating imagery that's not only accurate but also deeply contextually relevant to the source text.

3.4. Fine-Tuning GILL with Vicuna

Different LLMs come with their own set of strengths. Recognizing this, we integrated the Vicuna LLM into GILL, replacing its original OPT-based LLM. The modified GILL then aims to generate an image I_v from text T, denoted by the sequence $T \xrightarrow{\text{FT-GILL}_{\text{Vicuna}}} I_v$. By leveraging Vicuna's strengths within the GILL framework, we aim to explore whether this amalgamation can offer superior image generation results, enhancing the visual representation of long-form textual content.

4. Experiments

4.1. Dataset

Wikipedia-based Image Text (WIT) (Srinivasan et al., 2021) is a large multimodal multilingual dataset that comprises of image-text pairs with paragraphs extracted from Wikipedia articles and their corresponding images along with additional metadata such as the image caption, section title, and language. We limit our experiments to English and use the language field in the dataset to remove all non-english entries. We also remove all entries where the caption includes non-english characters. Furthermore, we manually curate a list of instance types as defined in Wikidata which would not be appropriate for image generation and filter out all articles which belong to these instances. We use the `page_url` from the WIT data and the Wikidata API [1] to get the instance type for each article. These filtered instances cover broad categories such as – humans, organization, art, landmarks, dates/numbers, locations, and events. In total, we filter out 47 instance types from the data, which we have included in Appendix B. After the non-english and instance type filtration steps, the resulting data set comprises Train (359.8K), Val (31.38K) and Test (22.55K) samples. The average number of tokens per paragraph in this filtered dataset is 169.41.

4.2. Evaluation Metrics

The goal of our evaluation is to measure the semantic and stylistic similarity of the generated images as compared to the ground truth images. To achieve this we use four semantic similarity and one stylistic similarity measures:

4.2.1. Semantic Similarity

We assess semantic similarity to ensure that the generated image conceptually aligns with the ground truth. For example, in a scenario where a Wikipedia article discusses a grocery store, if the ground truth image displays a collection of fruits and the generated image presents a view of a grocery store's produce section, we would consider them semantically similar. This judgment is based on the shared underlying theme of grocery items, despite variations in specific imagery. To achieve this, we utilize the following metrics:

1. **CLIP Similarity:** We compute the cosine similarity between the CLIP (Radford et al., 2021) ViT-L representations of the generated and reference images to capture the semantic similarity in the visual space. A higher cosine similarity indicates higher semantic similarity.

2. **BLIP-2 Similarity:** BLIP-2 leverages frozen image and text models by training a lightweight encoder between them which allows it to perform conditional text generation given an image. Therefore, we can use it to generate captions for an image. By generating captions from BLIP-2, we can focus quality evaluation

[1] https://www.wikidata.org/wiki/Wikidata:REST_API

Type	Model	CLIP Sim (↑)	LPIPS (↓)	BLIP-2 Sim_{BERT} (↑)	BLIP-2 $\text{Sim}_{\text{S-BERT}}$ (↑)	BLIP-2 $\text{Sim}_{\text{ROUGE}}$ (↑)
Reference	$\text{SD}_{caption}$	0.6477	0.7151	0.7033	0.4732	0.3462
Zero-shot	SD_{OPT}	0.5599	0.7406	0.6549	0.3364	0.2512
	SD_{Vicuna}	0.5998	0.7314	0.6669	0.3750	0.2692
	GILL	0.5674	0.7359	0.6660	0.3630	0.2624
Fine-tuned	FT-GILL_{OPT}	0.5947	0.7309	0.6798	0.3878	0.2884
	FT-GILL_{Vicuna}	**0.6054**	**0.7241**	**0.6813**	**0.3955**	**0.2925**

Table 1: Performance comparison on WIT dataset.

on the image content while disregarding stylistic factors. We propose a new metric called **BLIP-2 similarity**, which compares the similarity between the BLIP-2 generated captions for the generated and reference image. This approach allows BLIP-2 Similarity metric to act as an unbiased moderator, focusing on both the shared elements and the differences in text descriptions, such as objects and settings. This provides a different perspective from the purely visual comparisons, giving insights into the model's ability to replicate essential features of the ground truth images in the textual domain. We compute the text similarity using three metrics,

(a) **BERTScore:** F1 score using DeBERTa-XLarge-MNLI (He et al., 2023).

(b) **S-BERT Similarity:** Cosine similarity between the Sentence-BERT [2] (Reimers and Gurevych, 2019) encodings.

(c) **ROUGE:** F1 score based on ROUGE-L (Lin, 2004).

4.2.2. Stylistic Similarity

In addition to semantic alignment, ensuring stylistic congruence between the generated and ground truth images is an essential aspect of our study. For instance, if the ground truth image is rendered in black and white or exhibits a vintage style, it is essential for the generated image to reflect a similar stylistic theme. To assess this, we utilize the **Learned Perceptual Image Patch Similarity (LPIPS)** metric (Zhang et al., 2018). LPIPS calculates the perceptual similarity between images by measuring the similarity between the activations of two image patches using a pre-defined network, in our case AlexNet (Krizhevsky et al., 2012). A lower LPIPS score indicates that the two images are closer in the perceptual space and hence stylistically similar.

4.3. Experimental Setup

We compare zero-shot prompting, GILL, and fine-tuned GILL models. We use OPT-6.7B (Zhang et al., 2022), and Vicuna-7B (Zheng et al., 2023) for our zero-shot prompting experiments. Both LLMs use a maximum input length of 2048. For the TIM, we use Stable Diffusion (Rombach et al., 2022) v1.5, which has a maximum input length of 77 tokens. We experimented with three prompt prefix variations for the zero-shot experiments:

- Prompt 1: *"Summarize into one sentence that can be used as the caption of a corresponding image"*
- Prompt 2: *"From this text snippet generate the best caption to describe a relevant image"*
- Prompt 3: *"Craft a relevant image caption that represents the given text"*

We only show results for *Prompt 1* since it gave the best performance for both SD_{OPT} and SD_{Vicuna}. For the experiments with GILL, we use the model directly in image generation mode.

For the fine-tuned models FT-GILL_{OPT} and FT-GILL_{Vicuna}, we keep all pre-trained model weights frozen and only train the linear translation layers between the LLM and TIM, GILLMapper, and the `[IMG]` tokens. For both fine-tuned models we use bf16 precision, Adam (Kingma and Ba, 2015) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and a learning rate of $3e^{-4}$. The models are fine-tuned on 6 V100 GPUs using DDP with a batch size of 16 for 15 epochs. For, FT-GILL_{Vicuna}, we keep the max input length to 512 tokens due to GPU memory constraints. This required us to clip the context for 7% of the data where on average 116.91 tokens were dropped.

5. Results and Discussion

Table 1 shows our evaluation of various models on the image generation task using five metrics described in section 4.2. We begin by examining $\text{SD}_{caption}$, which represents the outcome of SD when utilizing high-quality human-generated captions associated with the gold images in Wikipedia. It is important to note that while $\text{SD}_{caption}$ serves as a valuable reference point being directly tied to the golden labels and associated images, it should not be regarded as the absolute upper bound. This is because these captions, although descriptive, are

[2] We use the all-mpnet-base-v2 model from HuggingFace.

supplementary to the image and do not have visual descriptors required for accurate image generation.

Next, we inspect the SD_{OPT} setup, where OPT is used to provide concise image description for SD to generate the image. The SD_{OPT} zero-shot setup is tested here because GILL framework integrates both OPT and SD, augmented with additional layers for image generation. The integrated GILL surpasses the standalone zero-shot SD_{OPT}, possibly due to its supplementary layers.

Our next comparison involves the SD_{OPT} and GILL models against SD_{Vicuna}, where we observe that SD_{Vicuna}, even in zero-shot setup, outperforms the integrated GILL system. However, the performance of SD_{Vicuna} might stem from Vicuna's inherent robustness as an LLM. Delving deeper, we explore if fine-tuning can help bridge the gap and help the integrated GILL model outperform pipelined models. To achieve this, we fine-tuned the GILL framework with OPT (FT-$GILL_{OPT}$) and Vicuna (FT-$GILL_{Vicuna}$) on our target dataset. While the fine-tuned FT-$GILL_{OPT}$ struggled to eclipse SD_{Vicuna} in the context of the CLIP Similarity metric, it demonstrated better performance across other evaluation metrics. Remarkably, fine-tuned FT-$GILL_{Vicuna}$ surpassed all other configurations across all the metrics.

In summary, our results highlight the effectiveness of SD when paired with a strong base LLM such as Vicuna using zero-shot prompting. Zero-shot approaches with LLMs are a promising choice for swift experimentation, especially when handling verbose input texts that demand nuanced reasoning. However, for those seeking further incremental enhancements, integrated systems like GILL, especially when supported with robust base LLMs, offer considerable promise.

6. Conclusion

In this paper, we studied zero shot and translation layer based approaches to utilize large language models and text-to-image models together for generating images from generic long-form text. In contrast, prior research has focused on improving the quality of images specified as prompts or from text with a clear visual component. We utilize a suite of metrics to evaluate generation quality and find that while zero-shot methods are competitive and easy to implement, translation layer approaches with fine-tuning perform the best. Additionally, we introduced a new BLIP-2 similarity metric, which can be used to measure image similarity in the text space. Our work also provides insights on how to adapt the GILL architecture for new tasks.

7. Ethical Considerations and Limitations

In this work we adopted off-the-shelf LLM and TIM models, thereby inheriting the accompanying caveats, such as hallucination (producing factually incorrect content) and implicit biases. We experimented with a limited set of open source LLMs and TIMs due to resource/time constraints, and using larger or stronger models can provide different results, which we intend to explore in the future. We show that using instruction tuned Vicuna-7B instead of OPT-6.7B provides gains in performance. It is also worth mentioning that our study is limited to English data, and additional work is required to test it in a multilingual setting. We compare the methods on a single dataset, WIT, which does have data from a wide variety of topics making it suitable for a preliminary study; however, there is still scope to expand the study to other types of data, such as blogs and recipes, where image generation from long-form text is applicable. Finally, as future work, we also plan to expand the evaluation to human annotations to benchmark our metrics and a have a stronger assessment of the relevance and accuracy of the system.

8. Acknowledgements

Thanks to Guoning Hu, Kellen Gillespie, Rishi Rajasekaran and Sudipta Kar for reviewing and providing feedback on our initial paper draft.

9. Bibliographical References

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. 2022. CM3: A causal masked multimodal model of the internet. *CoRR*, abs/2201.07520.

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neu-*

ral Information Processing Systems, volume 33, pages 1877–1901.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. 2022. Maskgit: Masked generative image transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11305–11315.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2022. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:10850–10869.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. Cogview: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 19822–19835.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip H. S. Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *ArXiv*, abs/2307.12980.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing prompts for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 66923–66939.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

Ting-Hao 'Kenneth' Huang, Francis Ferraro, N. Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *North American Chapter of the Association for Computational Linguistics*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2023. Generating images with multimodal language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 21487–21506.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1324. Computer Vision Foundation / IEEE Computer Society.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2022. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 833–842. Computer Vision Foundation / IEEE.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

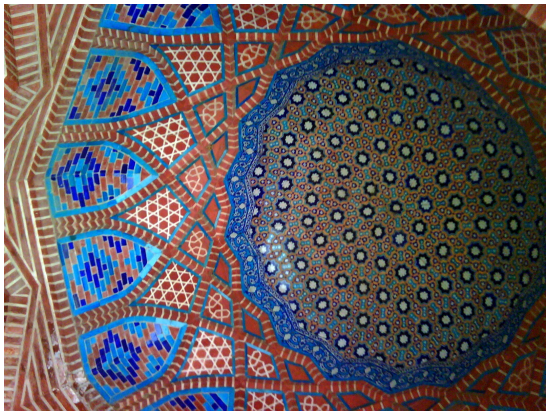
10. Language Resource References

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

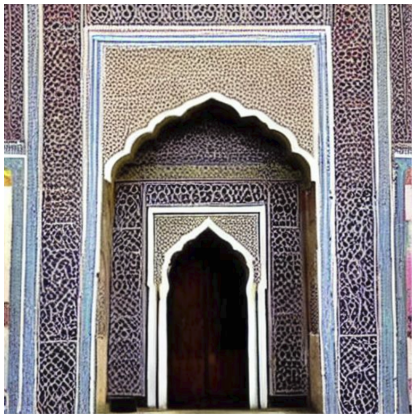
Appendix A. Qualitative Examples

In this section we provide more qualitative examples from our fine-tuned model. Each example provides a paragraph from the WIT dataset along with the ground truth and the generated image from FT-GILL$_{Vicuna}$.

Example 1: The Shah Jahan Mosque, also known as the Jamia Masjid of Thatta, is a 17th-century building that serves as the central mosque for the city of Thatta, in the Pakistani province of Sindh. The mosque is considered to have the most elaborate display of tile work in South Asia, and is also notable for its geometric brick work - a decorative element that is unusual for Mughal-period mosques. It was built during the reign of Mughal emperor Shah Jahan, who bestowed it to the city as a token of gratitude, and is heavily influenced by Central Asian architecture - a reflection of Shah Jahan's campaigns near Samarkand shortly before the mosque was designed.



Ground Truth



FT-GILL$_{Vicuna}$

Example 2: Although intended for the Mille Miglia, the 375 MM was also raced with limited success in the Carrera Panamericana, scoring fourth place in 1953 and finishing second in 1954. Other major successes in 1953 included overall wins at Spa 24 Hours, driven by Giuseppe Farina and Mike Hawthorn duo, 12 Hours of Pescara with Hawthorn and Umberto Maglioli and 12 Hours of Casablanca, won by Farina and Piero Scotti. The 375 MM with Alberto Ascari and Luigi Villoresi, was contesting the 1953 24 Hours of Le Mans alongside its 4.1-litre siblings, to no avail due to a clutch problems. In the 1000 km Nurburgring race of 1953, the 375 MM scored another victory with Giuseppe Farina, this time aided by Alberto Ascari. This race along with Spa 24 Hours counted towards the 1953 World Sportscar Championship, won for Ferrari in due honour to the 375 MM. In 1954 in Argentina, Giuseppe Farina with Umberto Maglioli won the 1000 km Buenos Aires, that was a championship race. On 760 km track of Coppa della Toscana, Piero Scotti won in the 375 MM ahead of Gordini. Later, the 375 MM competed in races in Europe, South and North Americas, winning many of them. The car did not score any more championship points as it was replaced by a bigger displacement derivative, the 375 Plus.

Ground Truth

FT-GILL$_{Vicuna}$

Example 3: Invercargill Airport is a fully secured controlled international designated aerodrome located 1.6 km west of the Central business district of Invercargill at the bottom of the South Island of New Zealand. It is the southernmost controlled airport in the Commonwealth. Formed on land reclaimed from the Waihopai/New River Estuary in 1938, the airport was prone to flooding, notably in 1984 when it was inoperable for two months. The Invercargill City Council considered moving the airport back to Dawson Farm, Myross Bush, the original site up to 1942. Instead, a large flood protection scheme was built, but during its construction heavy rain and an unusually high tidal surge flooded it again in 1987. There have been no problems since. The airport has a main secured terminal, a backup international secured terminal and 5 tarmac gates. Invercargill is the twelfth-busiest airport in New Zealand by passenger traffic.

Ground Truth

FT-GILL$_{Vicuna}$

Example 4: The 2/6th Battalion was an infantry battalion of the Australian Army that served during the Second World War. Raised in October 1939 as part of the all volunteer Second Australian Imperial Force, the battalion formed part of the 6th Division and was among the first troops raised by Australia during the war. Departing Australia in early 1940, the 2/6th were deployed to the Middle East where in January 1941, it took part in the first action of the war by Australian ground forces, the Battle of Bardia, which was followed by further actions around Tobruk. Later, the 2/6th were dispatched to take part in the Battle of Greece, although their involvement in the campaign was short before they were evacuated. Some members of the battalion also subsequently fought on Crete with a composite 17th Brigade battalion, and afterwards the battalion had to be re-formed in Palestine before being sent to Syria in 1941, 42, where they formed part of the Allied occupation force that was established there in the aftermath of the Syria, Lebanon campaign. In mid-1942, the battalion was withdrawn from the Middle East to help face the threat posed by the Japanese in the Pacific. A period of garrison duty was undertaken in Ceylon between March and July 1942, before they arrived back in Australia in August 1942. Following this, the 2/6th deployed to New Guinea in January 1943, fighting around Wau and then advancing towards Salamaua during the Salamaua, Lae campaign. They were withdrawn to the Atherton Tablelands for rest in September 1943 and subsequently did not see action again until later in the war, when they were committed to the Aitape, Wewak campaign in late 1944. The 2/6th remained in New Guinea until the end of the war, and was disbanded in February 1946, after returning to Puckapunyal the previous December.

Ground Truth

FT-GILL$_{Vicuna}$

Example 5: Both mining and logging create similar secondary deforestation through road construction. Specifically, logging companies construct new roads into previously inaccessible forest areas which facilitates the conversion of logged forests by into agricultural land. This has led to the immigration of landless farmers, in particular from eastern savanna regions, to enter primary forest areas through logging roads. Incoming farmers cause extensive land degradation in converting the natural forest into farmlands. Further, it has been suggested that increases in returns can lead to substantial increase in farm sizes and shortening of the fallow period, which in turn eventually leads to large-scale and severe natural forest area destruction.

Ground Truth

FT-GILL$_{Vicuna}$

Appendix B. List of Instances Filtered from WIT

1. human
2. aspect of history
3. meta-organization
4. regional organization
5. political organization
6. political territorial entity
7. confederation
8. economic union
9. international organization
10. intergovernmental organization
11. film
12. film series
13. television series
14. novel
15. novel series
16. literary work
17. painting
18. painting series
19. poem
20. office building
21. tourist attraction
22. skyscraper
23. observation tower
24. landmark
25. mausoleum
26. tomb
27. architectural structure
28. statue
29. architectural structure
30. mansion
31. presidential palace
32. sculpture
33. archaeological artifact
34. calendar year
35. leap year
36. even number
37. natural number
38. odd number
39. composite number
40. island
41. country
42. city
43. metropolis
44. big city
45. terrorist attack
46. suicide attack
47. demonstration