Using Large Text To Image Models with Structured Prompts for Skin Disease Identification: A Case Study

Sajith Rajapaksa[*1,3], Jean Marie Uwabeza Vianney[1], Renell Castro[1],
Farzad Khalvati[2,3,4], and Shubhra Aich[1,5]

[1]Yuko AI Inc., Toronto, ON, Canada [2] University of Toronto, ON, Canada
[3]SickKids Hospital, Toronto, ON, Canada [4] Vector Institute, Toronto, ON, Canada
[5]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

This paper investigates the potential usage of large text-to-image (LTI) models for the automated diagnosis of a few skin conditions with rarity or a severe lack of annotated datasets. As the input to the LTI model, we provide the targeted instantiation of a generic but succinct prompt structure designed upon careful observations of the conditional narratives from the standard medical textbooks. In this regard, we pave the path to utilizing accessible textbook descriptions for automated diagnosis of conditions with data scarcity through the lens of LTI models. Experiments show the efficacy of the proposed framework, including much better localization of the infected regions. Moreover, it has the immense possibility for generalization across the medical sub-domains to mitigate the data scarcity issue, and debias automated diagnostics from the all-pervasive racial biases.

1. Introduction

The limited availability of large annotated datasets is a serious impediment to progress in AI for medical imaging. In recent years, large pre-trained text-to-image (LTI) generative models, such as DALL-E2 [8], Imagen [10] have been used to greatly expedite the development of AI applications by eradicating the bottleneck of large-scale annotated datasets [5]. These models are trained on a massive amount of data using contextual unsupervised or self-supervised learning techniques. Such contextual training helps the models to learn the patterns and structures present in large amounts of text and image data and capture the contextual relationships between words and their meanings.

Although these LTI models are exposed to diverse datasets while training, the extent of the nature of data used for training is unknown. Moreover, how much diversity is captured in the pre-trained weights is also an open question. Consequently, it is not clear whether these models encountered or assimilated the information related to skin conditions sufficiently enough to be useful for automated diagnosis. Recent work shows the potential of realistic image generation to fool the classifier trained on real images [1]. However, in case of little or no real samples, the possibility of developing medical AI applications only with these LTI-generated samples is still uncharted territory.

Therefore, in this paper, we provide **an initial case study regarding the usage of LTI models for a few skin conditions for which the annotated dataset required for AI-assisted automation is scarce**. In particular, we employ DALL-E mini [2] to generate representative images for the four skin conditions – Atopic dermatitis, Urticaria Hives, Scabies, and Warts. The inputs to models like DALL-E and its variants are the text prompts describing the properties of the image to be generated. Providing effective inputs to these models is an art on its own, also known as prompt engineering [7]. Recently, prompt engineering is receiving a lot of attention from the research community to devise better prompts for more targeted image generation and refinement [7, 15]. However, in this paper, for our very initial case study, we stick to the simple (manual) formats equipped with disease names, skin tones, and symptoms and characteristics taken from the standard medical textbooks to assess the feasibility of these large models for the diagnosis of conditions. In this regard, we also investigate **the research question of utilizing textbook narrations, which are easy to obtain, for automated diagnosis of comparatively rare conditions through the lens of LTI models.**

---

[*]Corresponding Author: sajith.rajapaksa@mail.utoronto.ca

We show that the models trained with LTI images improve the classification accuracy of skin conditions over the model trained only on a scarce real dataset. Moreover, further visualization of the activation maps on real data indicates much better localization of the infected areas for the model trained with LTI-generated images compared to baseline (Figure 4). We believe such a substantial improvement in localization would help accelerate medical imaging research for which segmentation and localization play a big role in general. To the best of our knowledge, this is the first work regarding a strategic framework to utilize LTI models equipped with textbook-based structured prompts for improved medical image analysis where the need for training data is an issue that is commonplace in this domain.

Overall, our contributions are as follows:

- Our initial case study is the first known attempt to analyze the impact of LTI models for (comparatively rare) skin conditions through the lens of a strategic framework with a tentative guideline for simple prompt engineering based on medical textbooks.

- We show that the deep learning model trained on disease classification only with LTI generated images exhibits much better localization of the disease ROI in terms of class activation maps, thus demonstrating the efficacy of the proposed framework.

- Although the study in this paper is based on skin conditions, our framework is domain-agnostic to be applied to any sub-domain in medical image analysis dealing with the scarcity of targeted training data.

- Our framework with tentative prompt engineering guidelines can be easily extended by the community to significantly resolve the issue of racial bias (and potentially other prevalent ones) in medical datasets.

2. Methods

Our framework is depicted in Figure 1. It comprises two main components – (1) The choice of a set of structured text prompts used as inputs for data generation and (2) training a classification model using the synthetic images generated by the prompts.

2.1. Image Generation

Our objective is to utilize the textbook descriptions of the skin conditions to generate images for training the AI models. To do so, we employ an LTI model that generates images based on the text prompts provided as inputs. However, the textbook descriptions are usually quite long and attempt to describe the whole condition in a generalized manner. Although such descriptions are good for human-level communication, we find it somewhat difficult to generate targeted images simply using these long narratives.

Consequently, a succinct structure of the text prompts comprising placeholders for only the target set of keywords is required. The predefined prompt structure makes it generalizable across medical conditions at least in a particular sub-domain like skin. The placeholders in the generic structure can be replaced with the words sufficient and necessary to properly identify each condition.

To develop such a generic but laconic structure, we start by analyzing the textbook definitions of skin conditions. These definitions serve as the basis for the image generation process and help us understand the key features and characteristics of the conditions. We find that almost all these definitions contain four global features:

- Physical location of the occurrence of disease in the body.

- Skin tone of the affected individual following the Fitzpatrick scale [9].

- Visual cues, such as color, shape, appearance, etc.

- Sensation in/around the affected region. Note that sometimes the sensation is responsible for the partial change in appearance depending upon the actions taken by the patients to deal with that. Thus, despite being a completely physical feeling, including this information helps in better image generation.

Following this observation, our generic prompt structure takes the form of

```
<VISUAL_CUES> + <SENSATION> +
<PHYSICAL_LOCATION> + <SKIN_TONE>
```

An instantiation of our prompt based on the varying `<SKIN_TONE>` is shown in Figure 2.

Finally, to generate images with this structure, we instantiate the placeholder tags with concrete descriptions and feed the instantiations into an LTI model with different random seeds. We employ DALL-E mini [2] as our LTI candidate since it was fully open-sourced at the time of writing this paper.
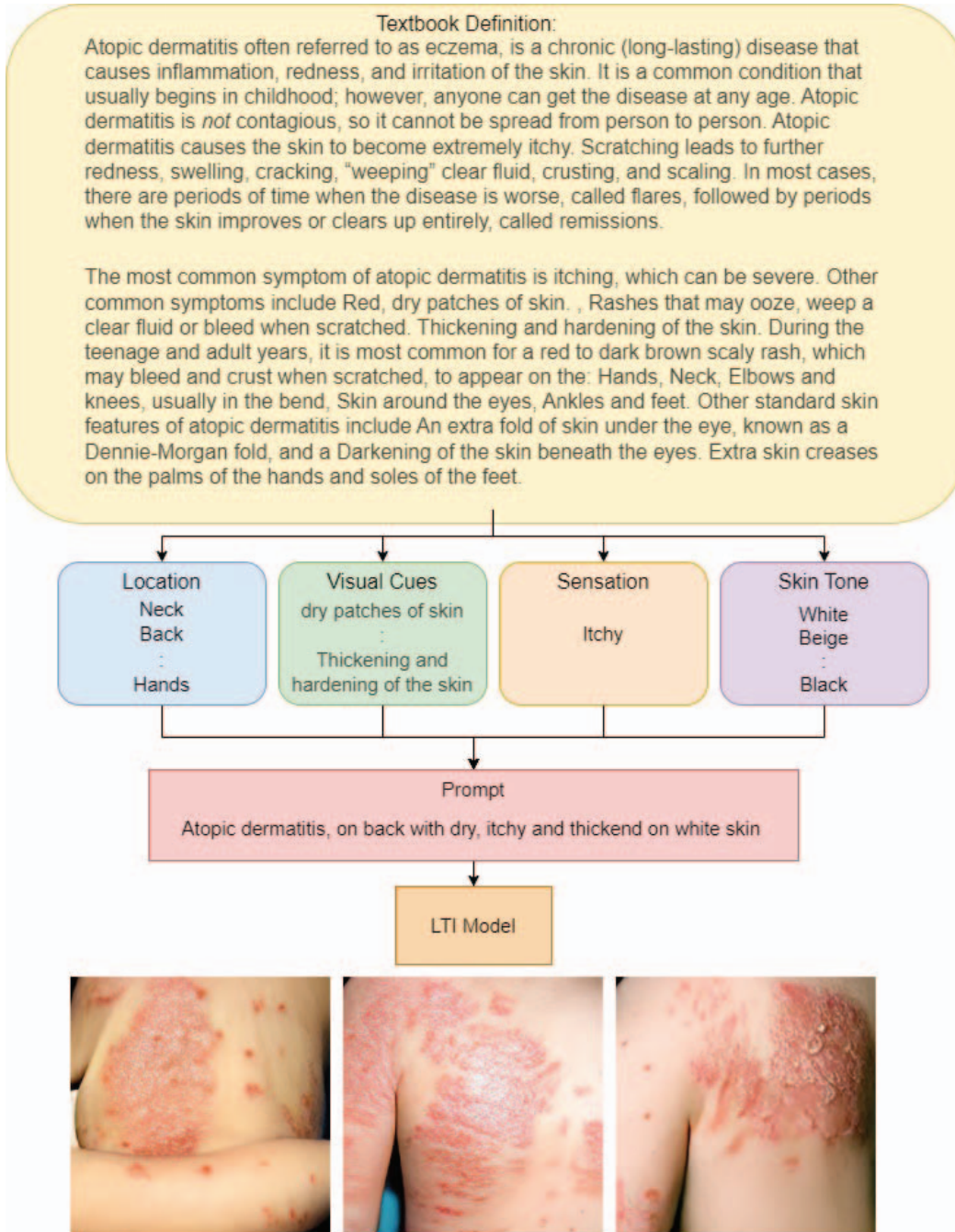
Figure 1: The proposed framework for generating images tailored to particular skin conditions based on textbook descriptions. First, the long textbook narrative is parsed into a few keywords following our generic but succinct prompt structure. Next, this prompt instantiation is fed into the LTI model to generate targeted images for further training.

2.2. Classification Model

Using the images generated with the set of instantiated prompts, we train a standard image classification model (ResNet50 [4] pretrained on ImageNet) for disease classifi-cation. This classifier is then evaluated on the real dataset both quantitatively and qualitatively. Details are provided in the Experiments section.

Figure 2: Samples generated by varying the <SKIN_TONE> parameter (left) of our prompt. The rest of the prompt instantiation is on the top of each column.

3. Experiments

Dataset: We generate images for skin conditions with comparatively rare data availability but for which there are at least some real samples available for evaluation. Such availability of a small amount of real data helps to evaluate the impact of the generated images for diagnosis quantitatively. Thus, four skin conditions are selected for our experiments – (1) Atopic dermatitis, (2) Scabies, (3) Urticaria hives, and (4)

Warts. For each of these four categories, $\sim 1K$ images are synthesized with varying physical locations and skin tones for training. A comparison between the real and synthetic images is shown in Figure 3. Additional synthetic examples are shown in Figure 5.

The real dataset used to test our framework is the publicly available Dermnet [3] containing 115, 55, 66, 131 samples for Atopic dermatitis, Scabies, Urticaria hives, and Warts,
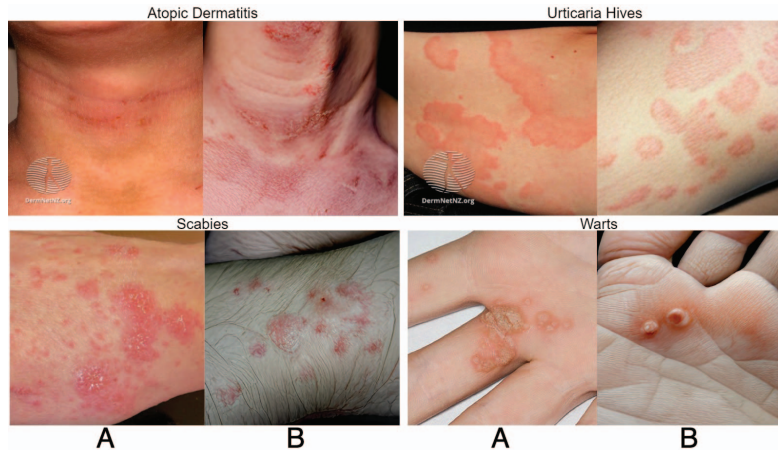
Figure 3: Sample comparison between real [3] (A) and generated images (B) for each of the four skin conditions.

Table 1: Comparison between synthetic and real fine-tuning.

	ImageNet + LTI	ImageNet + Finetune	ImageNet + LTI + Finetune
Accuracy (%)	42.0	56.0 ±0.6	63.0 ±2.8

The deviations (±) are reported over the average of 10 runs.

respectively. These samples are randomly split into 10% for finetuning and 90% for evaluation. Note that the subset for finetuning is used to show the potential improvement over the synthetic training with a small amount of real data.

We could not find any real samples with `non-white` skin tones for evaluation in this paper, whereas our synthetic training set is balanced in this regard, thanks to our generic prompt structure. Thus, our evaluation (not training) has an unavoidable racial bias all-pervasive in the medical domain that we wish to eradicate in future.

Training and Implementation: We train the ResNet50 [4] architecture for classification in PyTorch for 50 epochs with a learning rate of $1e^{-4}$ using Adam optimizer [6] and cross-entropy loss. The logos in Dermnet images were removed with standard image preprocessing techniques [14, 13]. This is to ensure consistency and eradicate any potential confounding factors that could affect the performance of the classifier.

Evaluation: All the evaluations in this paper are performed on the real dataset. For quantitative evaluation, first, we assess the model's classification accuracy trained only on our synthetic data to get a sense of the synthetic to real generalization ability. Next, we re-evaluate the synthetically trained model after finetuning it on a few real samples – ∼ 10 from each category. This is to see the potential improvement that a tiny real dataset brings to the table alongside the synthetic dataset. Moreover, we also visualize

the class activation maps [11] to analyze the saliency of the learned model and qualitatively evaluate the model's ability to localization of conditions.

Results: The LTI-generated images (Figure 3 and 5) capture the visual characteristics distinctive to individual skin conditions more or less well. Training the disease classification model only with these generated images achieves 42% accuracy on the 4 class classifications (Table 1). From the normalized confusion matrix of this synthetic-only results shown in Table 2, it is clear that the comparatively lower accuracy is mostly attributed to the `Warts` class being misclassified into `Atopic dermatitis`. This is because of the significant visual similarity of these two classes in the real dataset [3] used for evaluation. Despite the modest classification accuracy, the quality of disease localization shows significant improvement compared to baseline finetuning (see Figure 4).

Moreover, we finetune the logit layer of the base model (ImageNet pre-trained) and synthetic one on a small random subset of the real samples (10 per class). The results over ten random runs are reported in Table 1. The synthetically trained model improves the baseline model accuracy by 7.0% (63.0 *vs.* 56.0) on average.

Also, sample comparisons of the class activation maps [11] are shown in Figure 4 among the synthetic model, the baseline, and the synthetic with real finetuning. In this figure, the quality of saliency detection after the synthetic training looks much better than the one without this step. Thus, the efficacy of the synthetic examples is evident from the nu-
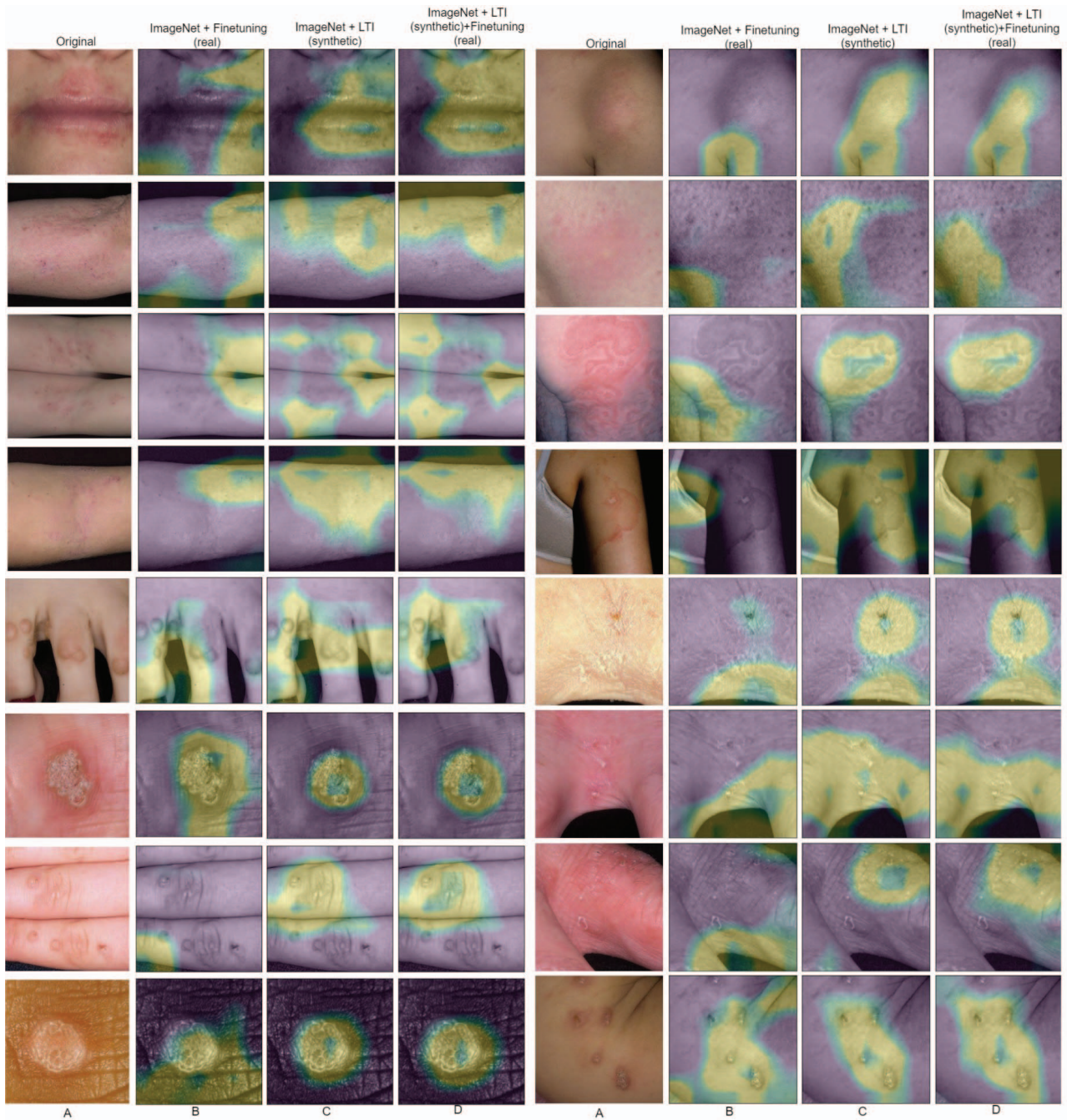
Figure 4: Comparison of the class activation maps for different protocols – (A) Images; (B) ImageNet + Finetuning (real); (C) ImageNet + LTI (synthetic); (D) ImageNet + LTI (synthetic) + Finetuning (real). The activation with the synthetic training (C and D) is qualitatively more accurate than the one without it (B).

merical and qualitative comparisons. We hypothesize that the accuracy of the synthetic-only model can be improved further by combining our structured prompt generation with non-redundant sampling strategies. Please consult the supplementary material for additional visualizations.

4. Conclusion and Future Work

In this paper, as an initial case study, we demonstrate the potential of the LTI models to be promising for skin disease detection, for which there is a serious lack of training data. At this point, we only employ the images generated with

Figure 5: Examples of the generated samples with our proposed structured prompts.

structured text prompts equipped with textbook descriptions for training and disease names or tags just for labelling. In this regard, our first attempt is somewhat unimodal, using only the generated images. However, the text prompts enriched with the laconic description following the textbooks provide an additional avenue to explore multimodal learning for improved performance and explainability. As part of future work, we will extend the methods to more diseases alongside the multimodal learning mentioned above. Moreover, the parsing of the textbook narrative into the prompt is

Table 2: Normalized confusion matrix for the synthetic-only model.

Ground truth	Prediction			
	Atopic dermatitis	Urticaria hives	Scabies	Warts
Atopic dermatitis	**0.64**	0.05	0.27	0.04
Urticaria hives	0.37	**0.47**	0.14	0.02
Scabies	0.36	0.04	**0.48**	0.12
Warts	**0.53**	0.01	0.32	0.14

performed manually in this paper. Automation of this process guided by the generic prompt structure will streamline application development. In addition, the recent advances in finetuning the diffusion models [1] for better visual data generation can also be explored as a future direction.

Most importantly, skin conditions appear differently based on skin tones and race. This issue is arguably among the most difficult ones to resolve with real datasets in the foreseeable future. The explicit and lucid nature of the text prompts (e.g. one particular tag for skin tone) used for data generation in this paper seems to be a promising way to mitigate such all-pervasive racial biases in the AI-assisted medical imaging domain. Moreover, such racial debiasing would also help us to create hyperlocal automated assistance programs for underserved communities in remote areas in alignment with our core organizational principles.

From an algorithmic perspective, our plan involves aligning the statistical moments [12] in a continual learning setup. The running estimate of the moments from real data streams will be synchronized with synthetic data. However, it would be unfair to evaluate the effectiveness of these methods based on the current data, given the skewed representation within the real images.

Finally, regarding generalization, although we are focusing on skin diseases for organizational purposes, we believe this study with a tentative guideline for prompt engineering will encourage the research community to utilize similar frameworks for other medical conditions in general as well.

References

[1] Pierre Chambon, Christian Bluethgen, Curtis P. Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. 2022.

[2] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phuc Le Khac, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 7 2021.

[3] Shubham Goel. Dermnet, Jun 2020.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Prabhu Kaliamoorthi, Aditya Siddhant, Edward Li, and Melvin Johnson. Distilling large language models into tiny and effective students using pqrnn. *arXiv preprint arXiv:2101.08890*, 2021.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[7] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2022.

[8] 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.

[9] Wendy E Roberts. Skin type classification systems old and new. *Dermatologic clinics*, 27(4):529–533, 2009.

[10] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[11] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Gradcam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[12] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.

[13] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[14] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018.

[15] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.