

# LLM-Generated Feedback for Generated Images

Sumit Nawathe, Aidan Melvin, Anthony Ma, Daniel Chung

CMSC 421 Artificial Intelligence  
University of Maryland College Park

## Abstract

We propose a method for improving artificial image generation using automated feedback. Current diffusion models don't always generate images that match the description. We wish to leverage the expansive prior-knowledge of LLMs to provide feedback to improve generation. To do this, we generated a dataset of captions, images, and feedbacks, and fine-tuned both an LLM and VLM to perform the feedback generation tasks. After fine-tuning, both the LLM and VLM were able to produce mostly correct feedback, which resulted in improved image generation. Our code can be found at: <https://github.com/aidmelvin/cmse421groupproject>.

## Introduction

Image generation models have made significant advances in recent years. Models such as Stable Diffusion have made a significant impact on the art industry and social media due to their ability to understand textual prompts and follow instructions to create detailed high-quality images. However, these models can be very difficult to control; techniques such as prompt engineering are necessary to achieve good results, which requires significant human involvement.

A method to improve image generation automatically through a feedback mechanism would be greatly beneficial to such models. Such methods have been proposed, but usually rely on extensively-trained vision-language models (VLMs).

Our approach is to fine-tune a pretrained large language model (LLM) to generate text feedback for diffusion-generated images. This novel method would leverage LLM's extensive prior knowledge and adaptability to downstream tasks. We aim to show that a fine-tuned LLM can achieve comparable performance to VLMs in this task.

We take a multi-step approach. Using a dataset of Pokemon images and captions, we construct a dataset of feedbacks for generated images. We fine-tune a pretrained LLM and pretrained VLM on a training dataset of these feedbacks and qualitatively assess their performance on a test dataset.

At a small scale, if successful, our work could be used to automate prompt engineering and improve the quality of image generation models without requiring additional effort from the user. On a larger scale, our work illustrates

the cross-domain power and expansive prior knowledge of LLMs, which could be further extended to other types of language and vision tasks.

## Outline

After discussing related works, we describe our LLM-based architecture as well as the VLM baseline we are comparing it against. We describe the dataset creation and training procedures, and then perform a qualitative evaluation of experimental test results. We conclude with a discussion of the consequences and limitations of our study.

## Related Works

In (Li et al. 2019), the authors proposed a novel controllable text-to-image generative model called ControlGAN. Their goal was to create a controllable model where only specific parts of an image are modified to the user's preference, since most current text-to-image models are uncontrollable and usually change the whole image when the prompt is changed. ControlGAN has three main components: channel-wise attention, word-level discriminator, and perceptual loss. Channel-wise attention exploits the connection between words and channels. Word-level discriminators identify the visual-change portions of feedback. Perceptual loss preserves the content of the image apart from desired modifications. Li et al., showed that ControlGAN can manipulate specific aspects of an image while maintaining the generation of the original content. Their results were strong, and suggests that ControlGAN was effective and superior to other models. This paper is relevant to our work because it shows the importance of maintaining the generation of the original content. The text-to-image generation model uses natural language to generate the images, and this is in direct parallel with how our project incorporates feedback to generate new images. However, our work fine-tunes pretrained models; we do not develop a new architecture.

In (Betker et al. 2023), the authors noticed that it is difficult to control image output due to low-quality language in training. Existing text-to-image models have trouble following detailed prompts, and the researchers suspect it is because so many datasets have poorly captioned images. To solve this and demonstrate the benefits of improved captioning for artificial image generation, the researchers trained an image captioner and used it to re-caption a dataset. Training

on this new synthetic dataset, they found significant quality improvements in text-to-image models. Finally, using these findings, the researchers created DALL-E 3. This is relevant to our work because it shows the importance of good captions when generating images. With an effective system to automatically generate feedback on generated images, image generation could see significant improvements in training.

(Hao et al. 2022) describes how the generation of quality images requires well-designed prompts. Currently, text-to-image models operate using text encoders that have a small capacity, and user inputs often are insufficient in creating quality pictures. Thus, they proposed a prompt adaptation framework that automates the prompt engineering process through reinforcement learning. Hao et al. began by fine-tuning a pre-trained language model on a set of prompts that were manually generated. Then, they employed diverse beam search to perform reinforcement learning. For image generation, they utilized open-source Stable Diffusion models. Hao et al. defined their reward function using a combination of aesthetic and relevance scores based on the newly generated images. They discovered that using reinforcement learning to generate optimized prompts yielded a greater gain in reward compared to their supervised fine-tuning baseline. They also saw that their model outperformed the baseline on out-of-domain data. This paper demonstrates the existing challenges in improving text-to-image generation and the need for improved prompt engineering methods. Our work attempts to similarly solve this challenge through supervised learning methods by fine-tuning LLMs using a human-made feedback dataset.

## Approach

Our goal is to train a model to provide textual feedback to improve image generation by a diffusion model. Starting with a dataset of pokemon images and captions, we use a Stable Diffusion model to generate images from the captions. For the feedback-generation model, we fine-tune OPT-2.7b, trained on instruction-formatted prompts with embeddings of the images using an off-the-shelf ViT. As a benchmark, we also slightly fine-tuned a BLIP2 on a similar dataset. Both fine-tuning processes were utilized LoRA.

## Image Generation and Dataset Creation

We utilize the Pokemon Blip Captions<sup>1</sup> dataset containing over 800 (image, text) pairs. Each image is a Pokemon and the corresponding text is a very short caption. These captions are vague and generally of dubious quality; sometimes, they describe the depicted character in a misleading fashion.

To generate images from these captions, we used Stable Diffusion v1.5<sup>2</sup>. We pass in the caption with no modifications to generate the image. The diffusion model gets some key details correct if provided in the caption. There are some strange visual artifacts, as is common with diffusion models.

---

<sup>1</sup><https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions>

<sup>2</sup><https://huggingface.co/runwayml/stable-diffusion-v1-5>

The model misinterprets some words; for example, it associates the word 'Pokemon' only with Pikachu, so any caption with that word (aside from Pikachu itself) has little connection to the original image.

To create our training dataset, we randomly selected 80 samples from the full dataset and wrote custom feedbacks for the diffusion model. We limited our feedbacks to be 1-2 short sentences. They address the main visual discrepancies between the true Pokemon image and the generated image. Table 1 provides some examples of such custom feedbacks and the potential benefit they can have on image generation.

For training the models, we use only the (caption, generated image) pairs, not the original image. This is because, in applications or testing of the model, we will not have access to the true image – that is what the diffusion model is trying to create. Thus, the feedback-generation models only have access to the original caption and the resulting image generated by the diffusion model.

## Fine-Tuning

The goal for both models is to learn the distribution of true Pokemon images indirectly through the feedback we provide, and thus be able to provide sensible feedback for any provide diffusion-generated image. Since this distribution is very specific, fine-tuning will be necessary; off-the-shelf models will be inadequate. Table 2 demonstrates feedbacks for models that are not fine-tuned.

Note that our training dataset size is relatively small for this task. This may lead to overfitting or memorization effects, as we discuss later.

## Benchmark Model: BLIP2 VLM

For a benchmark VLM model to generate similar feedbacks to those we had written, we utilized BLIP2<sup>3</sup>, a powerful image-to-text generation model by Salesforce. This model is commonly used for image captioning and visual question-answering (VQA) tasks. Our end-goal is similar to both of these tasks: the model is to create an image description based on the provided text prompt (generative model caption) and generated image.

Since by default, the text responses generated by BLIP2 are very short, we utilize minimal fine-tuning on our custom feedback dataset. We utilize 10 epochs of fine-tuning at a learning rate of 3e-4. (afterwards, the model significantly overfits). We give BLIP2 the image generated by the diffusion model, as well as the text in the following form:

Question: Below is an image generated by a diffusion model based on the caption '{original\_caption}'. What feedback would improve the model output?

Answer: {human\_written\_feedback}

## Target Architecture: OPT LLM

For our LLM target architecture, we utilize the OPT-2.7B<sup>4</sup> model by Meta. This is a transformer-based LLM that can

---

<sup>3</sup><https://huggingface.co/spaces/Salesforce/BLIP2>

<sup>4</sup><https://huggingface.co/facebook/opt-2.7b>

True Image	Original Caption	Generated Image	Custom Feedback	New Generated Image (with feedback)
	a green and white stuffed animal wearing a green hat		The character should have spikes like a cactus and be drawn against a plain white background.	
	a very cute looking pokemon with big eyes		It should look like a brown fox with a bushy tail against a white background.	

Table 1: A selection of datapoints to demonstrate the potential impact of feedback. The first two columns are from the Pokemon Blip Captions dataset, and the middle column is generated by Stable Diffusion based on the caption. When stable diffusion is applied to the full prompt of caption+feedback, the result is the rightmost column. In the first row, the caption incorrectly describes the figure as a ‘stuffed animal’, which is reflected in the first diffusion image. The feedback corrects this to be cactus-like. In the second row, the diffusion model incorrectly takes the word ‘Pokemon’ to mean Pikachu. The feedback provides an appropriate description of the figure, which the diffusion model responds to.

True Image	Original Caption	Generated Image	Not-fine-tuned OPT Feedback	Not-fine-tuned BLIP2 Feedback
	a cartoon ball with a smile on its face		I'm sorry, but I'm not able to find the link to the video. I'm not sure if it's a bug or not. I'll try again later. Thanks!	The smiley face is too small and the ball is too large.
	a picture of a cartoon character with a sword		I'm sorry, but I'm not sure what you're trying to do here. You're trying to make a simple, simple, simple, simple	The model output is not good enough.

Table 2: Examples of feedbacks generated by OPT and BLIP2 models without any fine-tuning. Prompt formats are the same as described in their respective architecture sections. Both BLIP2 feedbacks are very short; they occasionally reference visual attributes, but generally are not very helpful. The OPT model feedbacks are completely irrelevant to the prompt. In other example (not shown), the model outputs nonsensical tokens and is completely unintelligible.

be fine-tuned for text generation tasks. We chose the largest available pretrained model that we could fit on our available GPUs and train in a reasonable amount of time. Figure 1 visualizes our architecture.

The prompt for the LLM includes the original caption and diffusion generated image. For the former, we use the standard tokenizer for OPT models. For the latter, we use an off-the-shelf vision transformer model, ViT-MAE<sup>5</sup>. Passing the diffusion-generated image into this model produces 50 features, each of length 768. However, including all of these features in the LLM prompt would make the prompt too long and potentially drown out information about the context and caption; it would also dramatically increase training time. To simplify our input, we average these 50 features to create a single composite vector of length 768. Additionally, the LLM input expects integer tokens, but ViT-MAE outputs floating-point values. To adapt, we make the following simple transformation:

$$(\text{scaled\_vit\_embed}) = [1000 \times (\text{averaged\_vit\_embed})]$$

Based on the range of possible values for ViT-MAE embeddings and OPT tokens, these values are valid. While simplistic, the transformation is very roughly linear, and with fine-tuning, should be sufficient.

For the complete LLM prompt, we use instruction formatting:

Below is an image generated by a diffusion model based on the caption. Write feedback for the model to improve.

```
### Caption:  
{original_caption}  
  
### Image:  
{encoded_image}  
  
### Feedback:  
{custom_feedback}
```

This format displayed superior performance over a single-paragraph prompt. Intuitively, it allows the LLM to separate different parts of the input based on section and more easily attribute meaning.

Initially, the LLM views the ViT-encoded image as gibberish. Thus, a large amount of training is necessary; we ran our fine-tuning for 8000 epochs at a learning rate of 1e-4.

## PEFT Training

For both BLIP2 and OPT fine-tuning, we make use of LoRA, a method of parameter-efficient fine-tuning (PEFT). The method allows us to fine-tune these models by adding a small number of trainable parameters and keeping the original pre-trained weights frozen. This dramatically reduces the time and memory costs of training while retaining decent performance.

All training (for both models) was performed using the HuggingFace Trainer API using language modeling loss: cross-entropy loss for next-word prediction.

<sup>5</sup>[https://huggingface.co/docs/transformers/model\\_doc/vit\\_mae](https://huggingface.co/docs/transformers/model_doc/vit_mae)

## Evaluation

We find that the LLM model creates reasonably good text feedback based on the caption and encoded image. Qualitatively, it outperforms the lightly fine-tuned VLM benchmark. However, generating new images based on the feedback encounters additional challenges.

### Qualitative Text Analysis

To assess the performance of our method, we apply both the OPT- and BLIP2-based models on test examples. While we could use approaches to numerically judge which model performs better (such as the structural similarity index), since our sample space is small, a qualitative analysis of image components will suffice. Tables 3 and 4 provide detailed commentary that illustrate differences between model outputs.

**BLIP2 Results** The BLIP2 model was difficult to fine-tune due to its tendency of overfit quickly. As previously discussed, we only fine-tuned for a few epochs to prevent the model from solely memorizing our training dataset. Qualitatively, while the generated feedbacks are appropriate, they often do not add much to the original prompt. Many times, the feedbacks reference incorrect image features from the diffusion-generated image.

**OPT Results** While the LLM initially saw the ViT-encoded image as nonsense, it eventually managed to somewhat learn the representations. Some of the generated feedbacks referenced components of the true image that were not found in the caption nor in the generated image.

### Overfitting

As repeatedly mentioned, both of these models are prone to overfitting during training. The dataset we are using is fairly homogeneous: all of the images have similar style and content (cartoonish animals). The original image captions are of poor quality; it would be difficult even for a human to draw an reasonable image based on some captions. Therefore, our custom feedbacks are also fairly homogeneous, containing similar details (such as references to a 'white background' or 'cartoon style'.). We also only have 80 samples in our training dataset. As a result, both the BLIP2 and OPT fine-tuned models generate feedback that suffers from the effects of overfitting, including referencing phrases directly from the training dataset and giving feedback is tangentially or incorrectly related to test samples. To limit these effects, we employed early stopping procedures, but some residuals effects remain.

### Generated Images

As a final evaluation, we pass the feedbacks generated by both models back into Stable Diffusion to generate new images. The new prompts are simply the original caption concatenated with the feedback. Unfortunately, the new images still do not resemble the true Pokemon images. Even when the generated feedbacks are good (from our qualitative perspective), the diffusion model may does not generate good images. There are a few potential reasons for this:

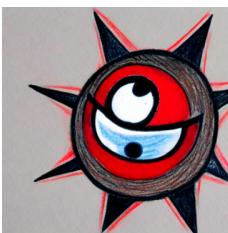
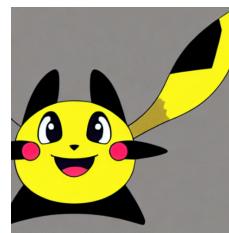
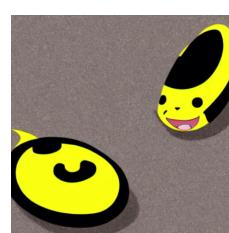
	Example 1	Example 2	Example 3
True Pokemon Image			
Pokemon Caption	a drawing of a star with a red eye	a drawing of a woman in a red cape	a yellow and black cartoon character with a big smile
Original Generated Image (caption only)			
OPT-generated Feedback	The Pokemon is a large bipedal star with a red eye. It has a white background.	The Pokemon is a large bipedal woman in a red and gray cape. It has a face consisting of dots in the shape of the letter H.	The Pokemon resembles a yellow cat with long droopy ears. It has a yellow tail with a black tail. It has a white background.
Generated Image with OPT Feedback			
BLIP2-generated Feedback	The image should be of a pokemon that has many rings and 5 floating hands, in a white background.	The image should be of a pokemon that is a red spiky cocoon in a white background.	The Pokemon is a yellow and black slug.
Generated Image with BLIP2 Feedback			
Commentary	In this case, the OPT-generated feedback was qualitatively quite good and contained features from the original image that were not in the caption. However, this feedback alone isn't good enough for even a human, let alone a diffusion model, to generate an image that looks like the target image. The BLIP2-generated feedback also suffers from the same issue.	Here, the OPT feedback once again looks qualitatively better than the BLIP2 feedback, but neither feedback can result in an image being generated quite like the original one. The generated images from the feedbacks are significant improvements over the original. Note that both generated feedbacks exhibit signs of overfitting.	In this example, the OPT-generated feedback is better than the BLIP2-generated feedback and contains visual elements of the true image. However, the original caption is so poor that both models struggle to generate useful feedback to improve the image. Although the feedbacks are qualitatively good, the new generated image is still quite different from the true Pokemon dataset image.

Table 3: Examples comparing OPT- and BLIP-generated feedbacks and associated images with the original dataset, accompanied by commentary.

	Example 4	Example 5	Example 6
True Pokemon Image			
Pokemon Caption	a drawing of a blue and orange pokemon	a pink butterfly with black wings and a white background	a green and yellow cartoon character
Original Generated Image (caption only)			
OPT-generated Feedback	A blue Pokemon standing on four legs with a gray tail against a white background. It has a white background. A white background should not be so white as to make the image look like a gray cat with blue stripes.	The Pokemon is a butterfly with pink wings that have blue, yellow, white, and black on the edges. It has a face consisting of dots in the shape of the letter H.	A Pokemon green dragon with yellow circles and a long tail with a white background. It has a white background. It has a green tail with some yellow circles and a white background.
Generated Image with OPT Feedback			
BLIP2-generated Feedback	The image should be of a blue and orange colored pokemon that has long arms, in a white background.	The Pokemon is a butterfly with black wings and a white background.	The Pokemon is a cartoon green and yellow bug with two antennas and two eyes.
Generated Image with BLIP2 Feedback			
Commentary	Here, we can see a slight but noticeable improvement in the images generated based on the OPT and BLIP2 feedback. Although not perfect, the feedbacks and new generated images are a step closer to the intended image from the dataset than the original image generated solely using the original caption from the dataset. The OPT feedback includes some unnecessary details.	A slight but noticeable improvement can be seen in the generated images after feedback. The LLM feedback in particular is quite detailed and contains elements not in the caption or original generated image, although some of it is wrong (for example, the yellow in the wings isn't correct).	The OPT-generated feedback seems to induce an image that is a massive improvement over the one generated based on the caption from the dataset – the final image is a green/yellow creature facing the camera. The BLIP2-generated feedback is completely wrong and leads to a departure from the intended image.

Table 4: More examples comparing OPT- and BLIP-generated feedbacks and associated images with the original dataset, accompanied by commentary.

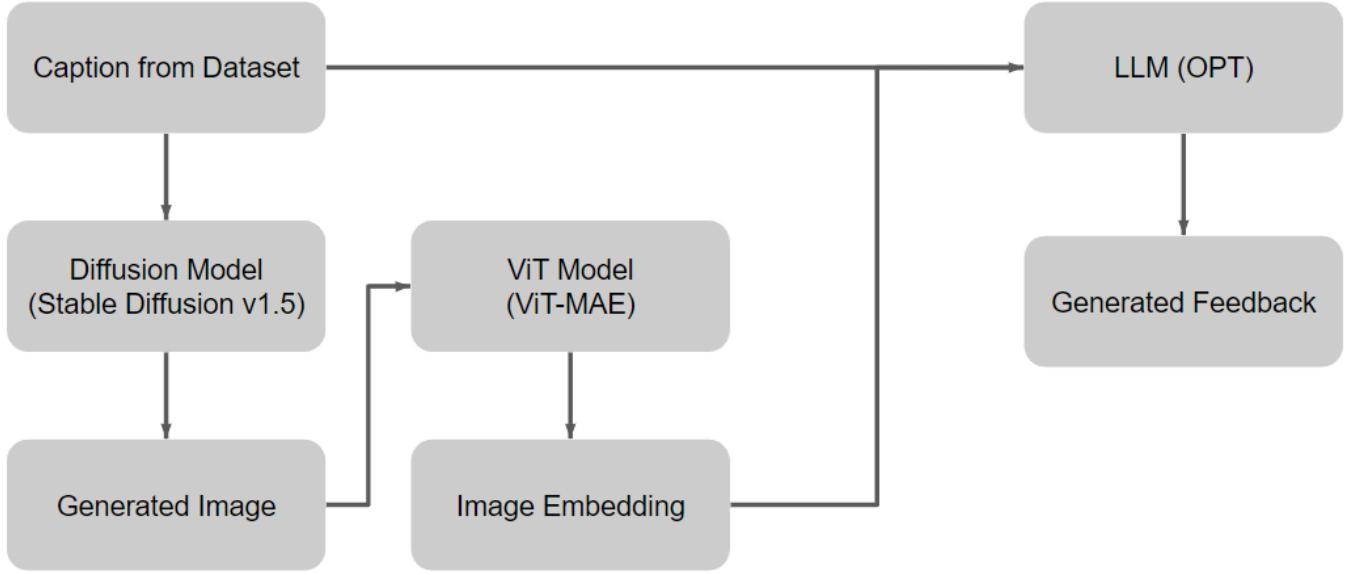


Figure 1: Diagram of our LLM-based architecture. The caption from the Pokemon dataset is passed through the Stable Diffusion model, and the output image is passed through a ViT. This embedding goes through some simple transformations before it is combined with the original caption and passed through the LLM, which generates feedback. We then use this feedback to augment the prompt to the diffusion model, and the final output image is compared to the true Pokemon image from the dataset.

- The diffusion model tends to focus on earlier parts of the prompt. Since we place the original caption before the feedback in the new combined prompt, the model tends to focus on the original prompt; the feedback has less than the desired effect.
- In some cases, the original Pokemon caption is so vague or misleading that the prompt is "uncurable": no matter what additional feedback we provide, the final image will be very different than what is desired.
- The diffusion model can misinterpret words or put undue emphasis on certain parts of the feedback. For instance, the specific diffusion model we are using strongly associates the word "Pokemon" with the specific pokemon "Pikachu" and the Pokemon logo. Thus, if feedback for a non-Pikachu Pokemon contains the word "Pokemon", the diffusion model might still output an image of Pikachu.

Many of these are faults of the diffusion model we are using, not necessarily of the feedback-generating models. Qualitatively, the feedbacks generated match reasonably well with what humans would write as feedback for the model, but the diffusion model is not able to adequately process it.

## Conclusion

### Summary

We propose the use of fine-tuning an LLM to generate feedback to improve text-to-image generation. We fine-tune OPT-2.7b using ViT-encoded images within the prompt, and compare it against a BLIP2-based VLM on a task to provide feedback on images generated by Stable Diffusion for

short Pokemon captions. We successfully fine-tuned these models to generate feedback that yielded improvements to image generation results.

### Limitations

This proof-of-concept has several limitations:

- Our training dataset was small (80 handwritten feedback samples). It was also fairly homogeneous, consisting of similar images and captions (for cartoonish beasts). These were major causes of overfitting we experienced in both the VLM and LLM models.
- The Pokemon captions dataset was of poor quality; many caption descriptions were incorrect or misleading.
- We lacked the computing power necessary for extensive training and hyperparameter tuning. We performed LoRA fine-tuning on the largest models we could, but could likely have achieved better results with larger models.

### Future Work

In the future, we hope to construct a much larger set of feedbacks on a larger and more diverse dataset. We also hope to use larger, more complex/competent LLMs and VLMs for testing, and appropriately perform hyperparameter tuning on both to avoid overfitting.

### References

- Betker, J.; Goh, G.; Jing, L.; TimBrooks, Wang, J.; Li, L.; LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Ramesh, A. 2023. Improving Image Generation with Better Captions.

Hao, Y.; Chi, Z.; Dong, L.; and Wei, F. 2022. Optimizing Prompts for Text-to-Image Generation. arXiv:2212.09611.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. H. S. 2019. Controllable Text-to-Image Generation. arXiv:1909.07083.

### **Acknowledgements**

All members of this group contributed equally.