Sketch to Reality: Enhancing AI's Understanding and Creation of Art

Alind GUPTA
Ecole Polytechnique
alind.gupta@polytechnique.edu

Vysakh RAMAKRISHNAN
Ecole Polytechnique
vysakh.ramakrishnan@polytechnique.edu

Abstract

This project aims to bridge the gap between abstract sketches and realistic images through advanced AI techniques. Students will learn to generate descriptive captions for sketches, create realistic images from these sketches, fine-tune a generative model to improve its output, and apply conditional generation to control specific aspects of the generated images.

1. Introduction

Understanding and generating visual content from abstract sketches remains a challenging task in artificial intelligence and computer vision. While realistic image generation has seen significant advancements through deep learning techniques, the ability to bridge the gap between sketches and photorealistic images is still an evolving domain. This project explores the transformation of abstract sketches into meaningful representations by leveraging multimodal AI models for caption generation, image synthesis, and conditional refinement.

Our approach involves a structured pipeline consisting of three key tasks: (1) generating descriptive captions for sketches, (2) synthesizing realistic images from these textual descriptions, and (3) enhancing the generated results using fine-tuned generative models. To evaluate performance, we employ automatic metrics such as BERTScore and CLIP-based similarity, along with human-led assessments for qualitative validation. Furthermore, we integrate ControlNet into the Stable Diffusion framework to enable fine-grained control over the image generation process, ensuring better alignment with the input sketches.

By systematically analyzing different AI-driven techniques, this work aims to enhance AI's ability to interpret, describe, and recreate artistic content with greater fidelity. Our experiments provide insights into the robustness of multimodal learning setups and the effectiveness of conditional generation strategies for sketch-to-image translation.

2. Background

In this section, we introduce all the relevant concepts that are used for various purposes throughout the project.

2.1. Image to Text Models

Vision-language models have significantly advanced the field of image captioning by leveraging multimodal learning. BLIP-2 (Bootstrapped Language-Image Pretraining) [1] introduced a two-stage pretraining framework that efficiently aligns vision and language models. It enables high-quality caption generation by first encoding visual features using a frozen vision model and then mapping them to textual representations through a lightweight query transformer.

Similarly, LLaVA (Large Language and Vision Assistant) [2] extends the capabilities of multimodal large language models by incorporating vision encoders with a conversational agent, allowing for contextual image captioning and reasoning. It fine-tunes LLaMA (Large Language Model Meta AI) to process visual information and generate descriptive textual outputs, enhancing real-world applicability.

2.2. BERTScore for Caption Evaluation

Evaluating the quality of generated text remains a crucial task in vision-language modeling. Traditional metrics like BLEU and ROUGE fail to capture semantic similarity. BERTScore [8] addresses this limitation by computing cosine similarity between contextualized word embeddings from BERT, effectively measuring the semantic alignment between reference and generated captions. It has been widely adopted for assessing text quality in NLP and vision-language tasks.

2.3. CLIP for Image Similarity

CLIP (Contrastive Language-Image Pretraining) [3] is a vision-language model trained on a vast dataset of image-text pairs. Unlike traditional image embedding methods, CLIP learns a joint representation of images and text, allowing it to compute the similarity between an image and a

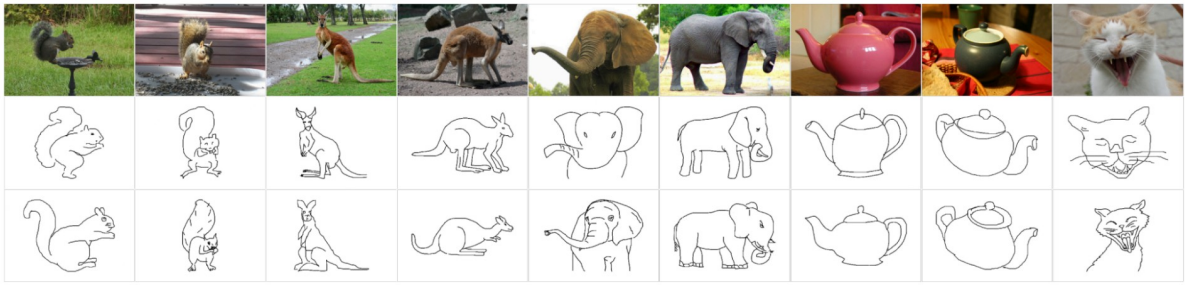

Figure 1. An example flow of the project

text description. This property makes it a powerful tool for evaluating image descriptiveness by comparing embeddings of generated captions with corresponding image representations.

2.4. Stable Diffusion for Image Generation

Generative models have demonstrated remarkable progress in synthesizing high-fidelity images from textual prompts. Stable Diffusion [4] is a state-of-the-art text-to-image model that operates in the latent space rather than pixel space, significantly reducing computational requirements while maintaining quality. It follows a diffusion-based approach, progressively refining noisy latent representations under the guidance of text embeddings to generate coherent images.

2.5. Structural Similarity Index Measure

Evaluating the quality of generated images is essential for generative modeling. The Structural Similarity Index Measure (SSIM) [7] provides a perceptual metric for comparing images based on luminance, contrast, and structural information. Unlike pixel-wise metrics such as MSE or PSNR, SSIM better reflects human visual perception, making it a preferred evaluation metric in image generation tasks.

2.6. ControlNet for Sketch-Guided Image Generation

While Stable Diffusion generates diverse images from textual prompts, ControlNet [9] enhances its controllability by incorporating additional conditioning inputs such as edge maps, depth maps, or sketches. By providing explicit structural guidance, ControlNet allows users to constrain the generation process, ensuring the synthesized image adheres to the desired shape and composition. This makes it particularly useful for transforming sketches into realistic images while preserving artistic intent.

2.7. Fine-Tuning Stable Diffusion with DreamBooth

Generating consistent subjects across different prompts remains a challenge for diffusion models. DreamBooth [5] introduces a personalized fine-tuning approach for Stable Diffusion, enabling the model to learn and reproduce specific visual concepts. By training on a small set of images with additional textual conditioning, DreamBooth allows for fine-grained control over image synthesis, making it suitable for domain-specific applications such as personalized art generation.

3. Methodology

In this section, we split the project into two parts, the image captioning task and the task of generating images from a given text input and explain the relevant methodologies for both tasks. We first start by describing the data and then in each subsequent subsection, we explain the methodology used to tackle the task.

3.1. The Sketchy Database

The Sketchy Database [6] is a large-scale dataset designed to facilitate research in sketch-based image retrieval, generation, and captioning. It contains 75,000+ hand-drawn sketches paired with real-world images, enabling multimodal learning across visual domains. Each sketch in the dataset corresponds to an object in a reference image, allowing for detailed exploration of how abstract representations can be translated into photorealistic visuals.

3.1.1 Dataset Composition

- **Sketch-Image Pairs:** Each object category contains multiple sketches linked to real images, capturing variations in human-drawn abstraction.
- **Metadata Annotations:** Each sketch is accompanied by **error level, ambiguity level, and worker annotations**, providing insights into sketch complexity and perceptual clarity.

- **Diversity & Scale:** The dataset spans multiple object categories, ensuring generalizability across different sketch-to-image transformation tasks.

3.2. Image Captioning with BLIP-2 and LLaVA

Image captioning is a fundamental vision-language task that involves generating descriptive textual representations of images. In this work, we explore two state-of-the-art models, **BLIP-2** and **LLaVA**, for generating captions from sketches. These models leverage pre-trained vision encoders and large language models to generate semantically meaningful descriptions.

3.2.1 BLIP-2: Captioning with and without Prompting

BLIP-2 employs a two-stage training process that aligns a frozen vision encoder with a lightweight transformer-based language model. It enables zero-shot and fine-tuned image captioning capabilities.

- **Zero-Shot Captioning:** BLIP-2 can generate captions directly from an input sketch without requiring additional prompts. The model extracts visual features and converts them into textual descriptions, capturing abstract object representations.
- **Prompted Captioning:** BLIP-2 also allows for input prompts that guide the captioning process. By providing context (e.g., "This sketch represents a..."), we can control and refine the generated captions, improving their specificity and accuracy.

We experiment with both prompting strategies to analyze their impact on caption quality, evaluating results based on semantic coherence and alignment with human annotations.

3.2.2 LLaVA: Multimodal Captioning with Large Language Models

LLaVA extends multimodal vision-language understanding by integrating a pre-trained vision encoder with a large language model. It enhances image captioning through interactive and context-aware processing.

- **Vision-Language Alignment:** LLaVA first encodes the image features using a frozen vision model, then feeds these features into an LLM to generate contextualized captions.
- **Conversational Captioning:** Unlike traditional captioning models, LLaVA enables iterative refinement by incorporating additional textual input, allowing for follow-up queries and adjustments to the generated descriptions.

Through comparative evaluation, we assess the strengths and limitations of BLIP-2 and LLaVA for sketch captioning. Key performance metrics include **BERTScore** for semantic similarity, CLIP-based embedding comparisons for descriptiveness, and human evaluation for qualitative assessment.

3.3. DreamBooth: Fine-tuning Sketch to Image Pipeline

Given approximately 3–5 images of a subject, the DreamBooth method fine-tune a text-to-image diffusion model using input images paired with a text prompt containing a unique identifier and the name of the class to which the subject belongs (e.g., "A [V] dog"). In parallel, the method also apply a class-specific prior preservation loss, which leverages the semantic prior that the model holds for the class and encourages it to generate diverse instances belonging to the subject's class by using the class name in a text prompt (e.g., "A dog") 2.

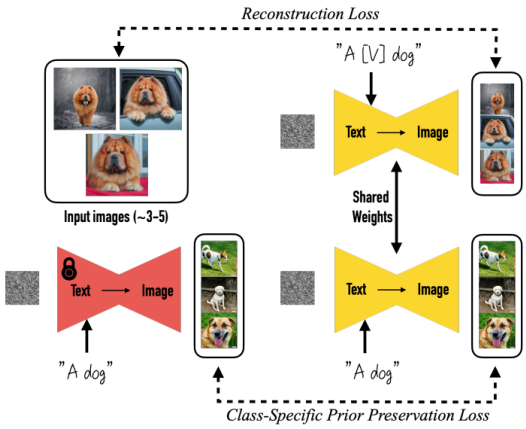

Figure 2. Finetuning pipeline of the DreamBooth method

4. Results

In this section, we describe the experiments that were done for both the tasks and we will also talk about the insights that we gained after running the experiments. We again split this section into two subsections where we talk about experiments and insights for each task.

4.1. Image Captioning Experiments

4.1.1 Experimental Setup

To evaluate the performance of BLIP-2 and LLaVA for the image captioning task, we designed a systematic experimental setup that ensures consistency and comparability across different models and configurations.

Dataset and Preprocessing We utilize the **Sketchy Database**, which contains over 75,000 sketches paired with

corresponding real-world images. Since sketches inherently lack fine details present in photographs, preprocessing is essential to standardize inputs. Any metadata associated with the sketches, such as ambiguity and error levels, is retained for analysis.

Model Configurations We experiment with both **BLIP-2** and **LLaVA** in different configurations:

- **BLIP-2 Zero-Shot:** Captions are generated without additional textual input, relying solely on the model's pre-trained vision-language alignment.
- **BLIP-2 with Prompting:** We introduce manually designed prompts (e.g., "This sketch depicts a...") to guide the captioning process.
- **LLaVACaptioning:** The model is teste in single-pass captioning mode.

Evaluation Metrics To assess caption quality, we employ both automatic and human evaluation methods:

- **BERTScore:** Measures semantic similarity between generated captions and ground-truth human annotations.
- **CLIP-Based Descriptiveness Score:** Computes embedding similarity between the sketch and the generated caption using CLIP to evaluate alignment.
- **Human Evaluation:** The authors manually also go thorugh the output to ensure the quality of the output

This experimental setup ensures a rigorous and comprehensive evaluation of the image captioning task, allowing us to analyze how well these models generalize to abstract sketch inputs.

4.1.2 Performance Comparison

Here we compare the performance of the different model configurations as previously described and give our derived insights. Refer to figures[2,3,4,5].

4.1.3 Insights

We derived the following insights from our experiments:

- **BLIP-2 is better:** In our experiments we found BLIP-2 to be performing better than LLava. This can be attributed to the presence of querying transformers in the BLIP-2 architecture.
- **Models are robust:** We do not observe significant performance difference with our without images that have been tagged erroneous or ambiguous by a human annotator. This implies that the models are robust to errors.

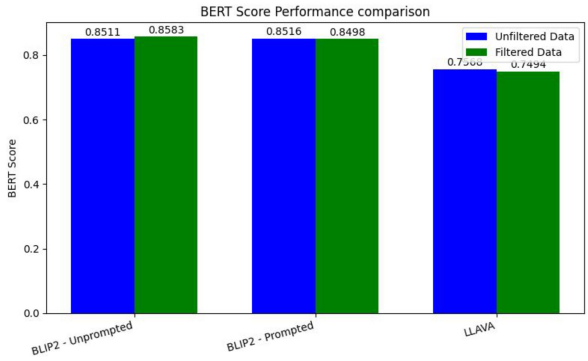

Figure 3. Comparison of Model configurations on BERTScore

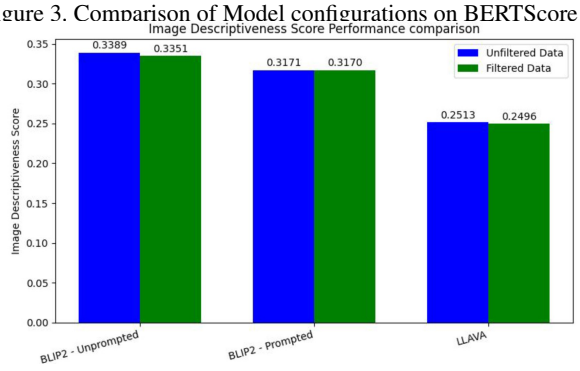

Figure 4. Comparison of Model configurations on Image descriptiveness score

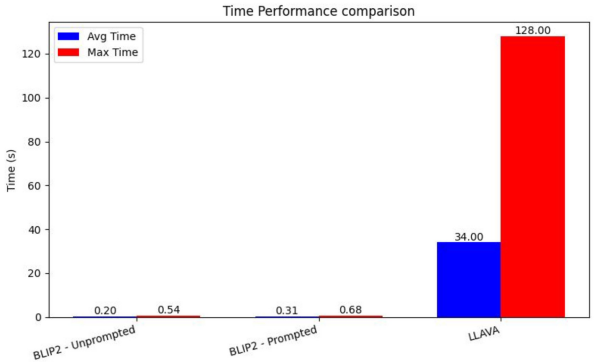

Figure 5. Comparison of Model configurations on inference times

4.2. Generating Images from Sketch Descriptions with Stable Diffusion

We introduce Stable Diffusion and explore its capabilities and limitations for the task 2. We then use the captions generated in Task 1 as input prompts for Stable Diffusion, synthesizing realistic images that align with the original sketch descriptions. Finally, we assess the quality and relevance of these generated images in relation to the initial sketches and their corresponding textual descriptions.

From this figure 7, we note the generation from sketch is not as expected and we need to incorporate the guidance to this using controlnet or finetuning Stable diffusion using Dreambooth, which is the task 3.

Figure 6. Generated caption: ”A drawing of a bird standing on a white background”

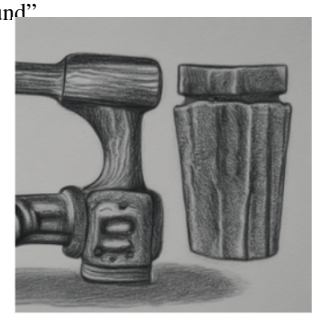

Figure 7. Prompt used for this - ”A hammer in black and white”

4.3. Fine-Tuning Stable Diffusion for Sketch-Based Image Generation

Inorder to provide better generation to the user, we used ControlNet for better guidance and finetuned the Stable diffusion using Dreambooth. Result of the ControlNet shows figure 8 us that the generation is much better and gives better fidelity to the produced image.

As a next step, we tried finetuning the model, and the results are shown in figure 9, 10. From this it's clearly evident that the betterment techniques, ControlNet and DreamBooth, both help in the process and are a way forward to bridge the good generation process between sketch and image generation.

5. Conclusions

The project ”Sketch to Reality: Enhancing AI's Understanding and Creation of Art” aimed to bridge the gap between abstract sketches and realistic images using advanced AI techniques. Our findings highlight several key

Figure 8. Result of ControlNet

Figure 9. Results of finetuning 1

Figure 10. Results of finetuning 2

achievements. In image captioning, BLIP-2 outperformed LLaVA due to its efficient vision-language alignment, and both models showed robustness to errors and ambiguities in sketches. For image synthesis, while Stable Diffusion showed promise, incorporating ControlNet for sketch-guided generation and fine-tuning with DreamBooth significantly improved image fidelity and alignment.

Evaluation through automatic metrics such as BERTScore and CLIP-based similarity, along with human assessments, validated the effectiveness of our approaches. The robustness of multimodal learning setups and the importance of conditional generation strategies were evident in our experiments. Future directions include enhancing AI models, developing user-friendly tools, and fostering interdisciplinary collaboration to drive innovative applications in AI-driven artistic creation.

References

[1] Baldridge J. Yang Y. Li, J. Blip-2: Bootstrapped vision-language pretraining with frozen image encoders and large language models, 2023. arXiv:2301.12597.

[2] Ding M. Yang Z. et al Liu, H. Llava: Large language and vision assistant, 2023. arXiv:2304.08485.

[3] Kim J. Hallacy C. et al. Radford, A. Learning transferable visual models from natural language supervision, 2021. Inter-

national Conference on Machine Learning (ICML.

[4] Blattmann A. Lorenz D. et al. Rombach, R. High-resolution image synthesis with latent diffusion models, 2022. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[5] Li Y. Jampani V. et al. Ruiz, N. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2022. arXiv preprint arXiv:2208.12242.

[6] Burnell N. Ham C. Hays J. Sangkloy, P. The sketchy database: Learning to retrieve badly drawn bunnies., 2016. ACM Transactions on Graphics (TOG).

[7] Bovik A. C. Sheikh H. R. Simoncelli E. P. Wang, Z. Image quality assessment: From error visibility to structural similarity, 2004. IEEE Transactions on Image Processing.

[8] Kishore V. Wu F. et al. Zhang, T. Bertscore: Evaluating text generation with bert, 2020. International Conference on Learning Representations (ICLR).

[9] Zhang Z. Zhang H. et al. Zhang, Z. Adding conditional control to text-to-image diffusion models, 2023. arXiv preprint arXiv:2302.05543.