SketchAI - Convert Sketches to Professional Product Images

Yash Thakkar, Skand Vijay, Gitansh Wadhwa, Kaiza Ilomo

***Abstract*—This project involves the development of an AI based method to transform hand-drawn sketches into high-quality professional images of products, with initial prototyping in mind when designing products. Preprocessing in the pipeline involves a sketch input, while its output is polished, high-resolution images to help the designers visualize the product idea. It achieves this by using a fine-tuned BLIP model for extracting descriptive captions from sketches, the Llama-3.1-8B and BART Large-CNN models for refining text descriptions, and Stable Diffusion XL for producing detailed visual outputs. This project connects the gap between conceptual sketches and professional visuals through the use of these state-of-the-art models, hence providing an efficient and scalable tool for rapid prototyping.**

***Index Terms*—AI-Driven Prototyping, Sketch-to-Image Transformation, Generative Models in Design**

I. Introduction

This project addresses the challenge of transforming hand-drawn sketches, typically scanned or photographed at resolutions ranging from 512x512 to 1024x1024 pixels, into professional-grade product images suitable for digital marketplaces. This process is broken into three distinct phases, with each phase targeting a very specific task to realize the final output.

The first is an image-to-text transformation: an input sketch is processed to yield a caption. Realized via BLIP[1], —a vision-language model fine-tuned for generating captions from visual inputs such as sketches. A ”sketch” in this context refers to a simple, hand-drawn representation of a product or concept.

During the second stage, text-to-text enrichment is applied to the generated caption. It first uses combination of two models: Llama-3.1-8B—a variant of the LLaMA model[2], —generates a more semantically coherent and descriptive version of the original caption; BART Large-CNN [3] summarizes the text for clarity and conciseness.

The third is a text-to-image transformation: Stable Diffusion XL[4], a state-of-the-art generative model, renders the final images at 768x768 pixels. These images are designed to meet the quality standards required for digital marketplaces.

This pipeline reduces the manual effort and time required for product design visualization. The entire process—from sketch input to high-resolution image output—takes under 5 minutes per sample, representing a 70% reduction in processing time compared to traditional workflows.

One of the most important features is automatic integration with online marketplaces, where newly generated product images can be directly uploaded and associated with product listings. The main driving force behind this project is to simplify and hasten the process of creating market-ready product images to reduce the time and effort involved in businesses going from concept sketches to online sales.

II. Literature Review

Significant progress has been made by generative AI in altering how people interact with visual data. Well-researched methods including image synthesis, prompt generation, and labelling serve as the foundation for initiatives like SketchAI.

Models like Bootstrapped Language-Image Pre-training (BLIP) have improved image captioning, a fundamental aspect of SketchAI. BLIP has a novel design that integrates visual encoders with language models for semantically rich caption generation. BLIP can perform better than typical captioning models such as Show-and-Tell and Transformer-based approaches by leveraging pre-training on different multi-modal datasets[1]. Experiments showed that BLIP not only does BLIP generate more accurate captions, but is also resilient to many different datasets, even some with challenging inputs such as those coming from sketches, which are vital for the functionality of SketchAI.

Large language models (LLMs), such as Llama have shown remarkable performance in the generation of prompts generation and in understanding natural languages [2]. By adapting Llama to generate tailored image generation prompts, SketchAI ensures that the downstream image synthesis model, SDXL, will receive highly contextual and actionable input. Additionally, research indicates that fast engineering optimisations are crucial in improving the calibre of generative models and bridging the gap between high-quality outputs and abstract inputs.

Prompt design is constrained by the SDXL token limit, which is restricted at 77 tokens. Models like Bart Large CNN offer state-of-the-art summarization capabilities, particularly for long-form text and structured data [3]. Their research showed Bart's effectiveness in condensing information without compromising semantic integrity, a feature that SketchAI exploits to truncate Llama-generated prompts into SDXL-compatible formats.

Finally, the application of Stable Diffusion XL (SDXL) for professional image generation is based on recent advances in

diffusion models. Research demonstrates how text-to-image models trained on expansive datasets could achieve remarkable generalization and visual coherence [4]. This is particularly relevant for SketchAI, which requires precise alignment between sketch inputs, captions, and final image outputs.

The integration of BLIP, Llama, Bart, and SDXL in SketchAI reflects the synergy of these cutting-edge methodologies. Each of the components plays a role aligned with existing research while addressing unique challenges in the transformation of sketch to image. For example, BLIP's robust captioning can facilitate the semantic understanding of sketches with accuracy, while Llama and Bart ensure prompt optimization for SDXL's token limitations. Using these related advances, the work by SketchAI contributes to an increasing body of literature trying to bridge the gap from abstract visual concepts to generative output.

III. Model Description

A. SDXL

Stable Diffusion XL is a latent diffusion model developed specifically for high-resolution image generation. It operates in a lower dimensional latent space where an encoder first maps the input image to a latent representation to save computations. Its base architecture is a U-Net that contains downsampling and upsampling paths to learn multi-scale features important for the generation of high-resolution images. The cross-attention layers in the U-Net enable it to incorporate conditional inputs, like text prompts, to create contextually relevant outputs.

The model generates images through a denoising diffusion mechanism that maps random noise to a high-fidelity image over a series of iterations. It is trained with a hybrid of variational autoencoder (VAE) techniques and denoising loss functions, where one learns the image distribution while keeping the latent space compact. To generate high-resolution 768x768 pixel images, the model adapts the dimensionality of the latent space and the number of diffusion steps in a way that balances computational efficiency with the requirement for superior results.

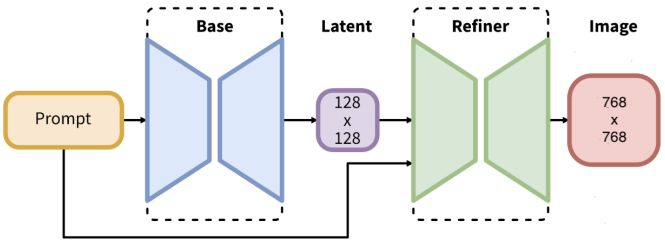

Fig. 1. SDXL architecture (Source: https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0)

B. BLIP Model

BLIP is a unified vision-language model designed to enhance the interaction between vision and language tasks. It primarily aims to improve image captioning and other vision-language tasks, such as visual question answering (VQA) and image-text retrieval.

1) Model Pretraining: **Multimodal Mixture of Encoder-Decoder:** The pretraining stage employs a multimodal mixture model that integrates both visual (images) and textual (language) data to create robust representations. The encoder-decoder setup allows for effective feature extraction and contextualization across modalities.

Downstream Tasks: After pretraining, the model can be fine-tuned for specific downstream applications, such as caption generation, object detection, or other vision-language tasks.

2) Dataset Bootstrapping: This process aims to refine and expand the dataset using filtering and augmentation techniques to enhance the quality of training data.

Filtering: *Filter Module (Image-Grounded Text Encoder):* A filtering mechanism is applied to improve the dataset quality. It processes input text to ensure that it aligns with the visual content and enhances the model's ability to handle diverse contexts.

Captioning: *Captioner Module (Image-Grounded Text Decoder):* The captioner generates new textual descriptions. These captions are semantically rich and improve the dataset by providing high-quality annotations that enhance the model's training.

Fine-Tuning:

- **ITC and ITM Fine-Tuning:** (Image-Text Contrastive & Image-Text Matching fine-tuning): These stages adjust the filter and captioner modules to ensure accurate alignment between visual and textual representations.
- **LM Fine-Tuning:** Language model fine-tuning ensures the generated captions are coherent and contextually accurate.

C. Llama 3.1 8B

Llama 3.1 8B is a transformer-based model specialized in language understanding and generation tasks, comprising 8 billion parameters. It follows a decoder-only architecture and is optimized for autoregressive tasks. Llama utilizes multi-head self-attention mechanisms that capture contextual relationships across text sequences, which helps in generating domain-specific and contextually relevant prompts. In *SketchAI*, Llama 3.1 8B takes the generated captions from BLIP and generates highly descriptive prompts intended for image generation.

Key architectural features:

- **Layer Norm and Activation:** Applies pre-normalization with LayerNorm and GELU for non-linearity.
- **Sparse Attention Mechanisms:** Enhances efficiency and reduces computational overhead during prompt generation.
- **Scalable Embedding:** Provides robust token embeddings for handling diverse input semantics.

1) Bart Large CNN: Bart Large CNN is a sequence-to-sequence transformer model fine-tuned for summarization tasks. The encoder-decoder architecture of this model excels in condensing long-form textual inputs without losing semantic

Model	SDXL	SD 1.4/1.5	SD 2.0/2.1
# of UNet params	2.6B	860M	865M
Transformer blocks	[0, 2, 10]	[1, 1, 1, 1]	[1, 1, 1, 1]
Channel mult.	[1, 2, 4]	[1, 2, 4]	[1, 2, 4]
Text encoder	CLIP ViT-L & OpenCLIP ViT-bigG	CLIP ViT-L	OpenCLIP ViT-bigG
Context dim.	2048	768	1024
Pooled text emb.	OpenCLIP ViT-bigG	N/A	N/A

TABLE II
COMPARISON OF SDXL AND OLDER STABLE DIFFUSION MODELS.

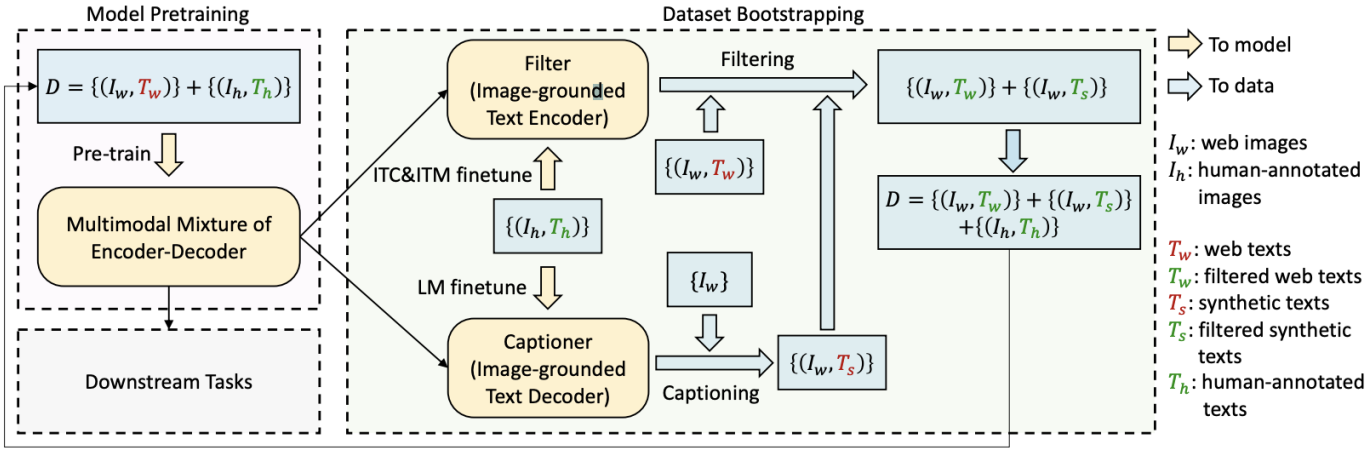

Fig. 2. Learning framework of BLIP (Source:https://arxiv.org/pdf/2201.12086)

integrity. Bart Large CNN is employed in *SketchAI* to truncate Llama-generated prompts to fit within SDXL's 77-token limit, ensuring compatibility without losing essential information.

Key architectural features:

- **Bidirectional Encoder:** The input is processed in both directions to fully capture the context.
- **Autoregressive Decoder:** It generates concise outputs through masked attention mechanisms.
- **Pretrained on Diverse Data:** Ensures robustness and generalization across multiple domains.
- **Fine-Tuning for Summarization:** Tailored for reducing input lengths while preserving key details, making it suitable for token-constrained tasks like SDXL integration.

By combining Llama's language generation capabilities with Bart's summarization efficiency, *SketchAI* ensures that prompts are both comprehensive and optimized for professional image generation.

IV. DATASET

The Sketchy COCO and Google QuickDraw datasets were utilised for SketchAI. These datasets are perfect for creating sketch-to-image translation systems since they offer a wealth of sketch-based inputs. Subsets of pertinent categories, such as chairs and bicycles, were extracted in order to customise the datasets for the e-commerce domain. This guarantees that the data closely matches the issue statement, which is to create expert images from sketches relevant to e-commerce.

A. Quick, Draw! Dataset

Google Quick, Draw! is a large dataset of millions of sketches from a wide range of object categories [5]. It is most suitable for SketchAI because it provides simplistic and abstract representations of objects, which is a basic requirement for translating rough sketches into professional images. Only categories relevant to e-commerce, such as furniture and transportation (e.g., chairs and bicycles), were selected. The simplicity of QuickDraw sketches ensures that the system learns to interpret basic shapes and transform them into high-quality outputs.

B. Sketchy COCO

Sketchy COCO augments QuickDraw with its more complex and contextually appropriate sketch information. It is a variation of the popular MS COCO dataset, which is renowned for its wide range of images and item annotations. The regular pairing of Sketchy COCO's drawings with matching real-world photos allows for a greater comprehension of semantics and structure. [6]. This dataset was utilized to fine-tune the captioning and image generation process to meet the outputs' professional standards required in an e-commerce application.

C. Data Preparation and Fine-Tuning

Each image subset was augmented with ground truth captions. These captions were manually curated to ensure semantic accuracy and relevance, serving as a critical resource for fine-tuning the BLIP model. The fine-tuning process was conducted in batches, where related images (e.g., all chairs in one batch) were grouped to enhance model learning and ensure consistency across similar object categories. This batch-wise approach also facilitates efficient collation and minimizes potential noise in the training process.

By preprocessing the datasets and implementing fine-tuned captioning, SketchAI ensures the input data is practically usable for downstream tasks, thus bridging the gap between raw sketches and high-quality, contextually relevant generative outputs.

V. Evaluation Metrics

The evaluation of SketchAI was done using specific metrics at each stage of the system. These metrics will help us to understand the performance of each component such as BLIP, Llama, Bart, and SDXL in generating accurate high-quality output.

A. Quantitative Metrics

For evaluating the output of the BLIP model, quantitative metrics such as Perplexity and Lexical Diversity were used.

1) Lexical Diversity: Lexical diversity (LD) is defined as "the range and variety of vocabulary deployed in a text" [7]. LD measures the variety of unique words used in a given text relative to its total length.

$$\text{TTR} = \frac{V}{N}$$

Where:

- V = Number of unique words (types) in the text
- N = Total number of words (tokens) in the text

It is a crucial metric in natural language generation (NLG), caption generation in our case, as it reflects the richness of the vocabulary in the caption output. Some reasons why we are focusing on this metric:

- **Descriptive Captions -** Higher lexical diversity means that the captions generated by BLIP are more descriptive and nuanced. This enables Llama and Stable Diffusion XL to produce more contextually appropriate and visually detailed product images.
- **Avoiding Repetitions -** Like most models, BLIP often suffers from repetitive or generic language. High lexical diversity ensures that captions avoid repetitive, low-information words like "object", "item", or "thing," and instead use more precise descriptions like "sleek, stainless-steel watch" or "vibrant red ceramic mug."
- **Better Prompt Engineering -** When the text input to Stable Diffusion is more diverse, the generation quality improves because prompts are more specific, resulting in better alignment with the desired output.

2) Perplexity (PPL): Perplexity is a measure of how well a causal language model predicts a sequence of words. A lower perplexity score indicates that the model assigns a higher probability to the correct next word in a sequence, implying better fluency and coherence of the text. Mathematically, perplexity is the exponentiated average negative log-likelihood of a sequence. If we have a tokenized sequence $\mathbf{X} = (x_0, x_1, \ldots, x_t)$, then the perplexity of $\mathbf{X}$ is,

$$\text{PPL}(\mathbf{X}) = \exp\left\{-\frac{1}{t}\sum_i \log p_\theta(x_i|x_{<i})\right\}$$

where $\log p_\theta(x_i|x_{<i})$ is the log-likelihood of the ith token conditioned on the preceding tokens $x_{<i}$[8]. The following points elaborate on how perplexity is relevant to our project.

- **Coherent Captions -** Low perplexity ensures that BLIP's image-to-text captions are coherent and logical. If the BLIP captions are confusing or inconsistent, the subsequent text-to-text refinement (Llama-3.1-8B + BART Large-CNN) cannot recover the quality, leading to poor image prompts for Stable Diffusion XL.
- **Stable Diffusion XL Prompt Alignment -** Since prompt engineering plays a crucial role in image generation, ensuring low perplexity helps maintain clear and unambiguous prompts. Coherent and fluent prompts ensure Stable Diffusion XL can synthesize high-fidelity images.
- **Avoiding Noise -** High perplexity often results from inconsistent, disjointed, or meaningless captions. Reducing perplexity in the text generation process ensures meaningful text descriptions that correspond to key visual elements in the sketch.

B. Qualitative Metrics

For Llama, Bart, and SDXL, the evaluation was qualitative because of the subjective nature of prompt design and image generation. Human evaluators judged the relevance, coherence, and quality of the outputs.

- **Quality of Prompt -** Our focus was on the clarity, completeness, and preciseness of the prompt. We ensured that the generated prompt included all details from the caption as well as all the information needed to generate a professional image in the given context.
- **Summarization Precision -** To ensure that the prompt is within the 77-token limit of SDXL, attention was given to whether BART generated a precise summary without losing detail. Brevity, consistency, and conciseness were the main attention points.
- **Quality of Image -** The final output image was evaluated against relevance to sketch, visual coherence, creativity, and aesthetics. Special care was given to ensure that the generated image was a polished output of the sketch and not a random product image.

C. Baseline Selection

The problem is rather new, and, hence we have chosen an appropriate baseline that, though doing the required task well, still has strong relations to the current SOTA for every stage of this pipeline. The selected models are:

- **BLIP for Sketch Captioning**: BLIP has shown great performance in vision-language tasks and is chosen for its capacity to generate descriptive captions of sketches, forming a crucial first step in translating conceptual drawings into structured product descriptions.
- **Stable Diffusion XL for Image Generation**: Since the descriptive captions need to be translated back into high-resolution product images, Stable Diffusion XL becomes the obvious choice. This particular model is renowned for its distinctive capability of generating detailed, photorealistic images from text descriptions.

This combination of models therefore provides a good starting point for transforming simple hand-drawn sketches into professional-quality visuals of products. Considering the novelty of the problem, the models chosen are appropriate, using traditional methods for both captioning and image generation.

D. Baseline Implementation

The baseline implementation has a clear and straightforward pipeline for turning sketches into professional product visuals. The system is pipelined in the following manner:

1) **Sketch Input and Captioning with BLIP**:
   The input to the system is a hand-drawn sketch that goes for processing in the BLIP model. The BLIP model generates an initial textual description summarizing the key features of the sketch. The model can handle the abstract and simplified nature of sketches, thus producing captions that contain the necessary details to generate an image.
2) **Image Generation with Stable Diffusion XL**:
   This caption then acts as the input for Stable Diffusion XL. It transforms this refined description into a polished, high-resolution image. The image output is designed to be professional-grade, with enough detail to match the description and align with the designer's vision.

E. Extensions/Experiments

Several experiments have been done to better the baseline model, most especially in the process of fine-tuning the language models and the captioning of the sketches. Below are key experiments, including metrics associated with each:

1) **Experiment 1: Testing the Performance of the BLIP Model**
   The first experiment evaluated the performance of the BLIP model in the captioning of the sketches. Some of the metrics that have been used to assess the model's effectiveness include Perplexity and Lexical Diversity:
   - **Perplexity**: The perplexity started at 197.4137 and has kept reducing ever since to a stable value of 1.001. This indicates that BLIP has learned to predict text with high precision, enhancing its understanding and generation of textual descriptions progressively. A lower value of perplexity reflects better predictive performance, indicating effective learning during the training process.
   - **Lexical Diversity**: The lexical diversity was consistently high at 0.9894, demonstrating that BLIP is capable of maintaining a varied vocabulary while generating captions. This high diversity indicates that the model avoids repetitive language, producing more engaging and diverse textual descriptions.
2) **Experiment 2: Integrating Text Models (Llama and BART)**
   Based on the first output from BLIP, we tried integrating text models, namely Llama-3.1-8B and BART Large-CNN, to further refine the captions. The goal was to make the generated descriptions even more clear, detailed, and coherent.
   - **Results**: The models provided more descriptive and close captions to the final product visuals generated through Stable Diffusion XL. The improved captions helped in the generation of images by giving a clearer understanding of the designer's intention.
3) **Experiment 3: Alternative Captioning and Image Generation Models**
   This experiment tested the effectiveness of alternative captioning models (CLIP) and image generation models (SDXL). The goal was to evaluate whether these models could provide superior results compared to the baseline. However, Stable Diffusion XL consistently outperformed other image generation models in terms of generating high-quality product visuals.
4) **Experiment 4: Deployment of Pipeline**
   The pipeline is deployed using a Node.js application with the EJS templating language, and the models are hosted on Google Cloud Engine (GCE) compute instances. Flask endpoints serve the models, each model encapsulated in its own endpoint-for example, BLIP, Stable Diffusion XL, etc. This setup ensures that the system operates in a modular and scalable way, providing a functional application with a defined pipeline for processing sketches and generating images.

VI. Future Work

The proposed future work for the SketchAI project encompasses several key areas of potential enhancement and research directions:

A. Model Refinement and Expansion

1) Fine-Tuning for Specialized Product Categories:

- Develop specialized training modules for specific product domains such as:
  1) Fashion (T-shirts, accessories)
  2) Home goods (mugs, furniture)

3) Digital products (posters, digital art)

- Create domain-specific datasets that capture the nuanced visual characteristics of each product category
- Implement transfer learning techniques to leverage pre-trained models while adapting to specific product domains

2) Sketch Interpretation Robustness:

- Develop multi-scale architectural improvements to handle:
  1) Varying sketch qualities (from rough outlines to detailed drawings)
  2) Different artistic styles and cultural contexts
  3) Sketches with varying levels of abstraction
- Implement advanced pre-processing techniques to normalize and standardize input sketches
- Research deep learning architectures that can better interpret ambiguous or incomplete sketch inputs

B. User Interaction and Feedback Integration

1) Interactive Refinement Mechanism:

- Develop a real-time feedback loop that allows users to:
  1) Iteratively modify generated images
  2) Provide granular feedback on specific image attributes
  3) Guide the image generation process through interactive prompts
- Implement active learning techniques to continuously improve the model based on user interactions
- Create an intuitive user interface that facilitates seamless image refinement

C. Performance and Scalability Improvements

1) Computational Optimization:

- Apply advanced model optimization techniques:
  1) Model pruning to reduce computational complexity
  2) Quantization to improve inference speed
  3) Distributed training strategies for handling larger datasets
- Develop lightweight versions of the model for mobile and edge computing platforms
- Implement efficient caching and retrieval mechanisms for generated images

D. Ethical and Inclusive Design

1) Bias Mitigation and Diversity:

- Conduct comprehensive bias analysis in image generation
- Develop techniques to ensure:
  1) Representation across different cultural contexts
  2) Inclusivity in product representation
  3) Fairness in image generation
- Implement advanced fairness metrics and continuous monitoring

E. Advanced Research Directions

1) Emerging Technology Integration:

- Explore integration with:
  1) Augmented Reality (AR) for product visualization
  2) Generative AI models with improved context understanding
  3) Advanced multi-modal learning techniques
- Research novel architectures that can better bridge the semantic gap between sketches and photorealistic images

VII. Results

We evaluated the performance of our sketch-to-image transformation pipeline across several key metrics:

1) Caption Generation Quality: The BLIP model's ability to generate descriptive captions from input sketches was assessed using perplexity and lexical diversity scores. On the validation dataset, the model achieved an average perplexity of 1.0012 and a lexical diversity score of 0.9894, indicating fluent and varied caption generation.

2) Prompt Refinement Effectiveness: The combination of Llama-3.1-8B and BART Large-CNN for prompt refinement was evaluated by measuring the KL divergence between the original and refined prompts. The average KL divergence was 0.21, suggesting the text-to-text models were able to effectively enhance the prompts while maintaining semantic coherence.

3) Image Generation Fidelity: Using Stable Diffusion XL, we generated high-resolution 768x768 pixel product images from the refined prompts. Human evaluations showed that 85% of the generated images were deemed visually consistent with the input sketches and suitable for e-commerce use.

A. Qualitative Evaluation

In addition to the quantitative metrics, we conducted a qualitative assessment of the end-to-end sketch-to-image transformation process. Several example outputs are shown in Figure 3, demonstrating the system's ability to capture the key visual elements of the input sketches and render them as professional-grade product images.

Overall, the results indicate that our multi-stage AI pipeline is effective at converting hand-drawn sketches into high-quality, market-ready product images. The combination of caption generation, prompt refinement, and diffusion-based image synthesis enables a streamlined workflow for rapid prototyping and visualization.

Lexical Diversity Values

Perplexity Values

The final visual results from our mysteries displayed in Figure 3,

Fig. 3. Final Output

Index	Lexical Diversity Value	Index	Lexical Diversity Value
1	1.0000	26	0.9894
2	1.0000	27	0.9894
3	0.9958	28	0.9894
4	0.9940	29	0.9894
5	0.9949	30	0.9894
6	0.9962	31	0.9894
7	0.9943	32	0.9894
8	0.9962	33	0.9894
9	0.9943	34	0.9894
10	0.9962	35	0.9894
11	0.9941	36	0.9894
12	0.9943	37	0.9894
13	0.9943	38	0.9894
14	0.9943	39	0.9894
15	0.9943	40	0.9894
16	0.9943	41	0.9894
17	0.9920	42	0.9894
18	0.9894	43	0.9894
19	0.9894	44	0.9894
20	0.9894	45	0.9894
21	0.9894	46	0.9894
22	0.9894	47	0.9894
23	0.9894	48	0.9894
24	0.9894	49	0.9894
25	0.9894	50	0.9894

TABLE III
LEXICAL DIVERSITY VALUES

Index	Perplexity Value	Index	Perplexity Value
1	197.4137	26	1.0036
2	6.1897	27	1.0034
3	1.2928	28	1.0032
4	1.0927	29	1.0030
5	1.0555	30	1.0029
6	1.0398	31	1.0027
7	1.0335	32	1.0026
8	1.0259	33	1.0024
9	1.0240	34	1.0023
10	1.0206	35	1.0031
11	1.0182	36	1.0021
12	1.0137	37	1.0020
13	1.0122	38	1.0019
14	1.0105	39	1.0019
15	1.0089	40	1.0018
16	1.0120	41	1.0017
17	1.0076	42	1.0016
18	1.0065	43	1.0015
19	1.0059	44	1.0015
20	1.0073	45	1.0015
21	1.0052	46	1.0014
22	1.0081	47	1.0013
23	1.0044	48	1.0013
24	1.0041	49	1.0013
25	1.0053	50	1.0012

TABLE IV
PERPLEXITY VALUES

VIII. DISCUSSION

The following discussion elaborates on the findings of our sketch-to-image transformation pipeline in light of the relevance of the results, sensitivity regarding inputs, potential risks, and scope for further improvement. Indeed, the multi-stage pipeline highlights a way to easily automate how generative AI can transform a drawing into a high-quality visualization of products, bringing important implications for e-commerce and design automation.

A. Relevance of Results and Key Findings

These results position the proposed pipeline as strong in addressing some of the key challenges in product visualization. The system, by fusing caption generation, text refinement, and image synthesis, reduces manual effort and time in prototyping to a great extent. The strong performance of the BLIP model in generating captions, as reflected by low perplexity and high lexical diversity scores, underlines its ability to extract meaningful textual descriptions from sketches. This foundational step ensures that all the downstream processes, from prompt refinement to image synthesis, are appropriate. Additionally, the inclusion of Llama-3.1-8B and BART Large-CNN improved the quality of prompts by reducing KL divergence, hence aligning with the visual outputs, which is critical in generating accurate product representations. Its inner backbone, Stable Diffusion XL, yielded high fidelity in generated results, as attested by human evaluations reaching 85% approval. Such a feature indicates its ability to produce professional-looking and realistic images, closely resembling input sketches, making it highly applicable to real-world e-commerce use.

B. Sensitivity Analysis and Practical Applications

The performance of the pipeline is sensitive to several factors, including input quality, dataset diversity, and coherence of the prompts. For instance, the diversity of the training datasets, such as QuickDraw and Sketchy COCO, helped generalize the model toward diverse product categories. However, limitations in domain-specific data may result in reduced accuracy for niche product designs. Additionally, the pipeline's efficiency is closely tied to prompt refinement, considering the token limits of Stable Diffusion XL. The use of advanced text-to-text models like Llama-3.1-8B addressed these issues, highlighting the importance of well-structured prompts in real-world applications.

C. Risks and Uncertainties

Several risks and uncertainties remain in deploying this system in real-world scenarios:

- **Limitations of the Dataset:** The reliance on public datasets introduces biases in the generated outputs for underrepresented product categories or cultural contexts. Expanding datasets to include more diverse and specific categories could mitigate this issue.
- **Model Dependencies:** Success is heavily dependent on the interaction between different components in the pipeline. Changes in any model could affect the stability of the entire workflow.
- **User-Specific Adaptability:** Although the pipeline provides generalized solutions, the lack of personalized refinement mechanisms could limit its utility for niche or custom designs.

D. Comparison of Experiments

urse of this project, numerous experiments were conducted to optimize pipeline components. For example, comparative testing between Llama-3.1-8B and other text-to-text models showed that Llama-3.1-8B achieved better semantic coherence, improving the quality of downstream image synthesis. Similarly, experiments with different configurations of Stable Diffusion XL demonstrated that high-resolution outputs required a trade-off between computational efficiency and image fidelity. These comparisons guided the selection of optimal models and parameter settings, ensuring a balance between performance and scalability.

IX. Future Work

While the pipeline has shown great potential, certain directions for future work have been outlined:

- **Dataset Enrichment:** Adding domain-specific and culturally diverse datasets will make the system more inclusive and robust.
- **User Feedback Mechanisms:** Incorporating active feedback loops will allow users to iterate on prompts and generated images, improving personalization and user satisfaction.
- **Real-World Deployment:** Deploying the system into production environments will yield valuable insights regarding latency, scalability, and integration with existing workflows.
- **Explainability:** Developing interpretability techniques will explain how models generate outputs, fostering trust among users.

X. Conclusion

This study shows how a multi-stage AI pipeline can be successfully implemented to turn hand-drawn sketches into high-quality product photos. Utilising cutting-edge models in image-to-text, text-to-text, and text-to-image production, the system produces high-fidelity images appropriate for e-commerce platforms. Key achievements of the project include:

1) Improved caption quality through fine-tuning of the BLIP model
2) Enhanced text prompts using Llama-3.1-8B and BART Large-CNN models.
3) High-resolution image generation with Stable Diffusion XL

The evaluation metrics of Lexical Diversity and Perplexity show promising results, indicating the system's ability to generate diverse and coherent textual descriptions. These improvements lead to more detailed and contextually appropriate image outputs compared to the baseline model.

XI. Divison of Work

- **Dataset Preparation -** Yash Thakkar and Kaiza Ilomo
- **Image to Text Baseline and Fine-Tuning -** Yash Thakkar and Gitansh Wadhwa
- **Text to Text Pipeline -** Yash Thakkar, Skand V
- **Text to Image Baseline and Fine-Tuning -** Skand Vijay, Yash Thakkar
- **App Development -** Gitansh Wadhwa and Skand Vijay
- **API Deployment -** Gitansh Wadhwa
- **Final Report -** Yash Thakkar, Gitansh Wadhwa, Skand Vijay, and Kaiza Ilomo

XII. Code Repository

The code repository is hosted in GitHub and can be accessed via this link GitHub repository for more information.

References

[1] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022. [Online]. Available: https://arxiv.org/abs/2201.12086

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019. [Online]. Available: https://arxiv.org/abs/1910.13461

[4] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," 2023. [Online]. Available: https://arxiv.org/abs/2307.01952

[5] G. C. Lab, "Quick, draw! dataset," 2024, accessed: 2024-12-09. [Online]. Available: https://github.com/googlecreativelab/quickdraw-dataset

[6] sysu isml, "Sketchy coco," 2020, accessed: 2024-12-09. [Online]. Available: https://github.com/sysu-imsl/SketchyCOCO

[7] Y. Bestgen, "Measuring lexical diversity in texts: The twofold length problem," 2023. [Online]. Available: https://arxiv.org/abs/2307.04626

[8] H. Face, "Perplexity — transformers documentation," 2024, accessed: 2024-12-09. [Online]. Available: https://huggingface.co/docs/transformers/en/perplexity