# CLOTHES DESIGN UTILIZING GENERATIVE AI

AIP490_G18

# MEMBERS

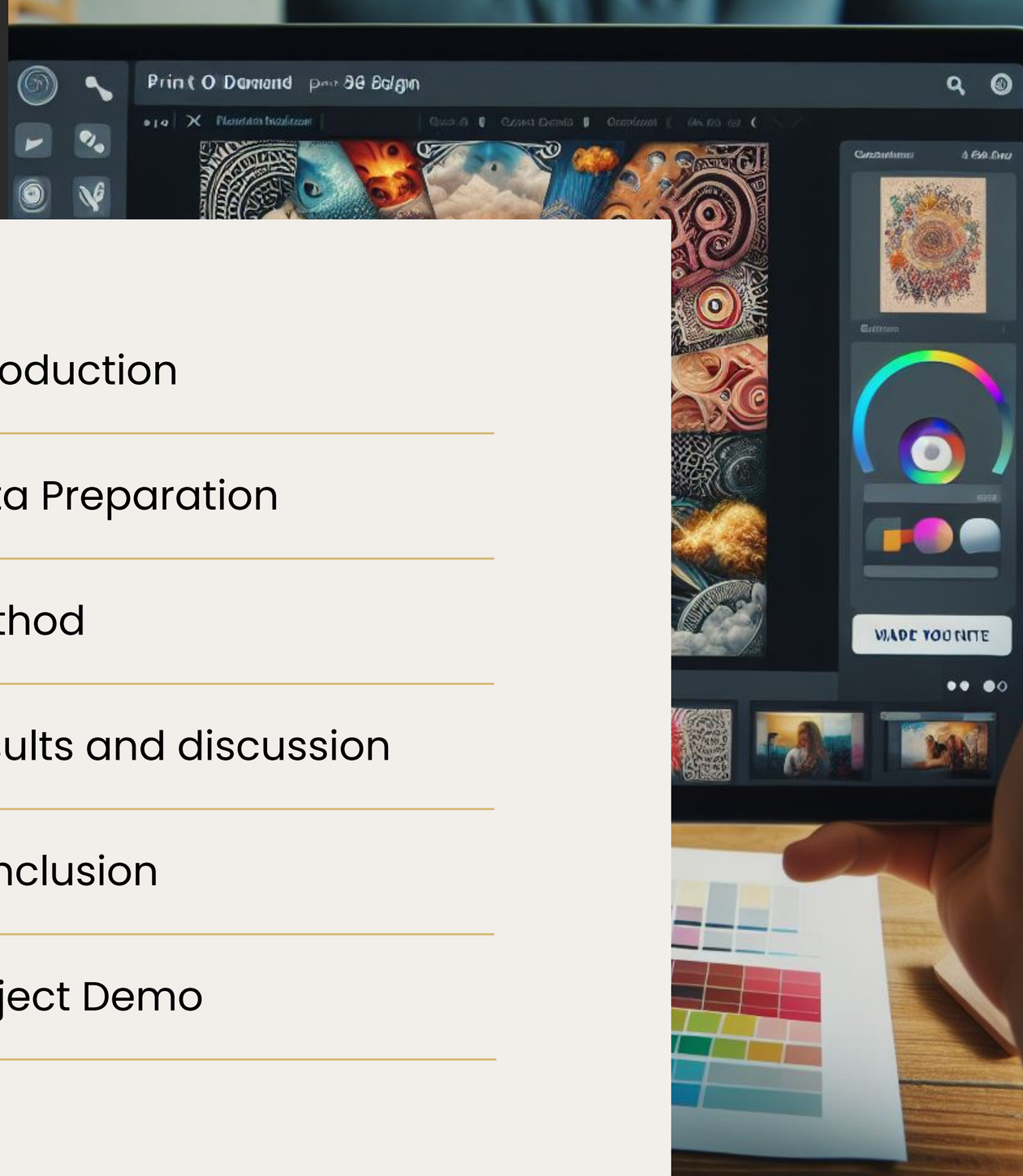
NGUYEN QUANG PHUOC

Leader

NGUYEN NGOC MINH

Member

# OUTLINE

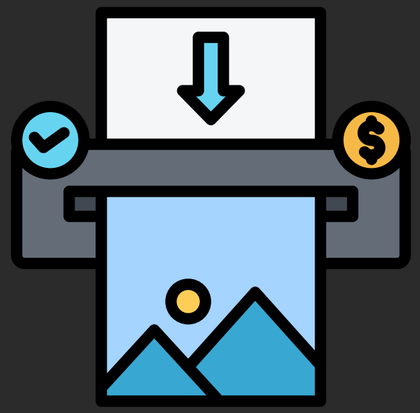


# 1. INTRODUCTION

- Text-to-image generation.
- How we came up with Text-to-image generation apply to Print on Demand.
- Our approach

# 1. INTRODUCTION

- Text-to-image generation has gained traction in recent years, with practical applications in various industries.
- Companies like Amazon see the potential of AI algorithms replacing stylists and designers.
- DeepVogue, an AI fashion design system, received recognition and awards in the fashion industry.
- AI enables individuals without design backgrounds to create their own artworks.
- Image generative models like GANs, CGAN, AttnGAN, and diffusion models like DDPMs have enabled the generation of high-quality and diverse images based on user-specified text conditions.



# 1. INTRODUCTION

- Applying image generation models to POD drawings for clothing, especially printed shirts, is an appealing concept.
- Manual labeling of image-caption pairs for this project is time-consuming.
- The Shifted Diffusion method is explored for text-to-image generation in a language-free setting.
- The referenced article shows that the proposed method achieves superior results compared to other text-to-image generation models.


# 2. DATA PREPARATION

Consists of 2 prior models:

- The Prior Model (Shifted Diffusion Model).
- Decoder (Stable Diffusion 2).

Datasets:

- MS-COCO 2017.
- Amazon clothes image dataset.

# 2. DATA PREPARATION

| Dataset | Size (pixels) | Category | #train/val/test | Total |
|---|---|---|---|---|
| MS-COCO [11] | 640x480 | Humans and objects | 118k/5k/41k | 164k |
| CUB-200-2011[10] | 500x500 | Birds | 6k/0/6k | 12k |
| CC3M [23] | >400x400 | All that pass the filters | 3.3M/16k/13k | 3.3M |
| LAION 400M [24] | 256x256 | All except NSFW | | 413M |

# 2. DATA PREPARATION

A child holding a flowered umbrella and petting a yak.

A woman wearing a net on her head cutting a cake.

A man with a red helmet on a small moped on a dirt road.

# 2. DATA PREPARATION


# 3. METHOD

- Prior Model and Decoder checkpoints are publicly available for fine-tuning.
- Decoder checkpoints were fine-tuned on various datasets, including MS-COCO, Localized Narratives, CelebA-HQ, and CUB.
- MS-COCO dataset was chosen for the Decoder checkpoint due to its diverse range of images, which is suitable for clothing design.
- Prior Model checkpoints were trained with T5-11B text embeddings, which are not directly applicable to the current system limitations.
- Retraining of the Prior Model is necessary using the Flan-T5-Large model, which has similar parameters to T5-Large but outperforms older T5 versions, including the 11 billion T5 checkpoint.

# 3. METHOD

- The model is trained on a single Q RTX 8000 GPU rented from Vast.ai with 45 GB VRAM.
- Training is conducted on 118,286 examples from the MS-COCO 2017 train split.
- Hyperparameters such as "train_batch_size," "num_train_epochs," "gradient_accumulation_steps," and "t5_model" are adjusted to optimize resource utilization.
- During inference, the number of layers is halved compared to the original model, and optional parameters such as "guidance_scale," "num_inference_steps," "height," "width," and "negative_prompts" can be adjusted to customize the generated images.

# 3. METHOD

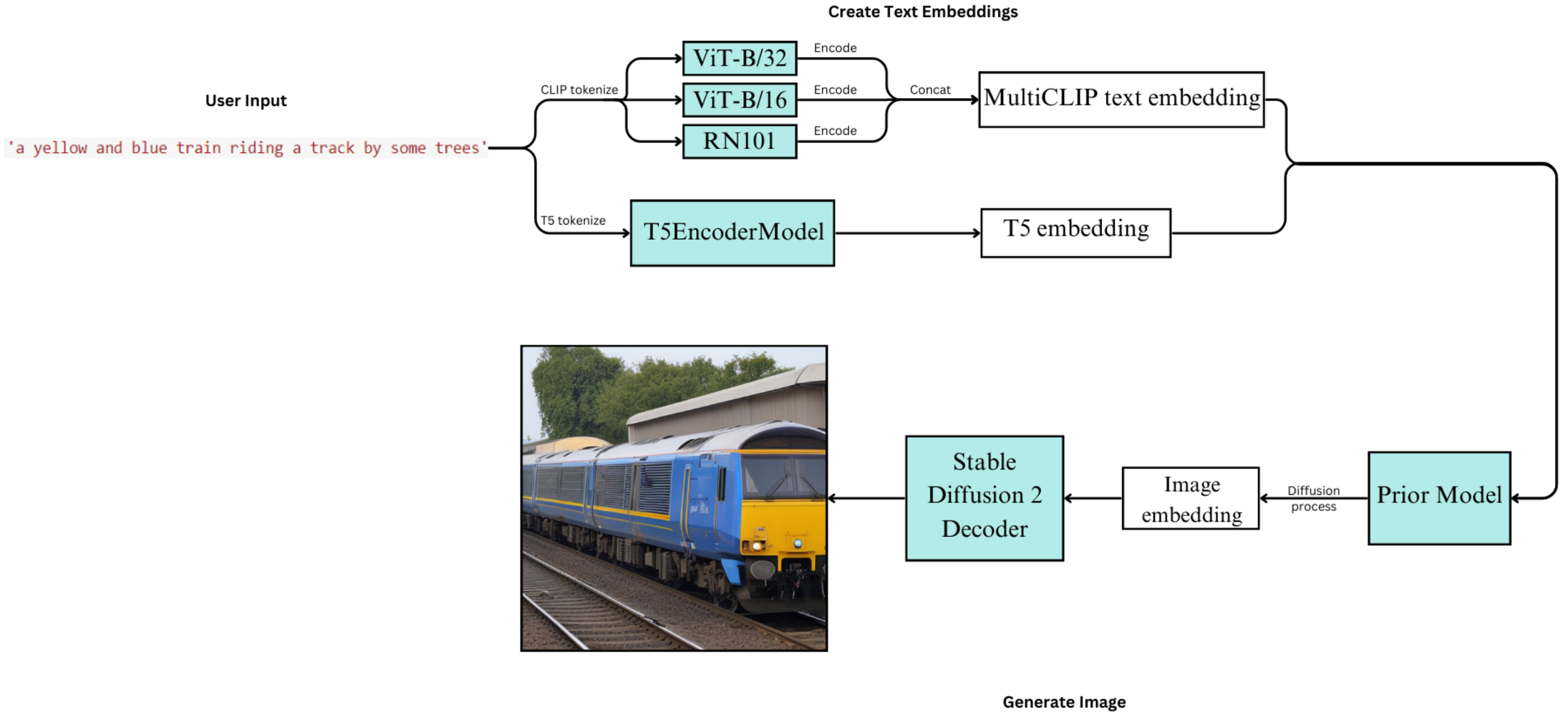
Example of CUB dataset images

Example of CelebA-HQ dataset images

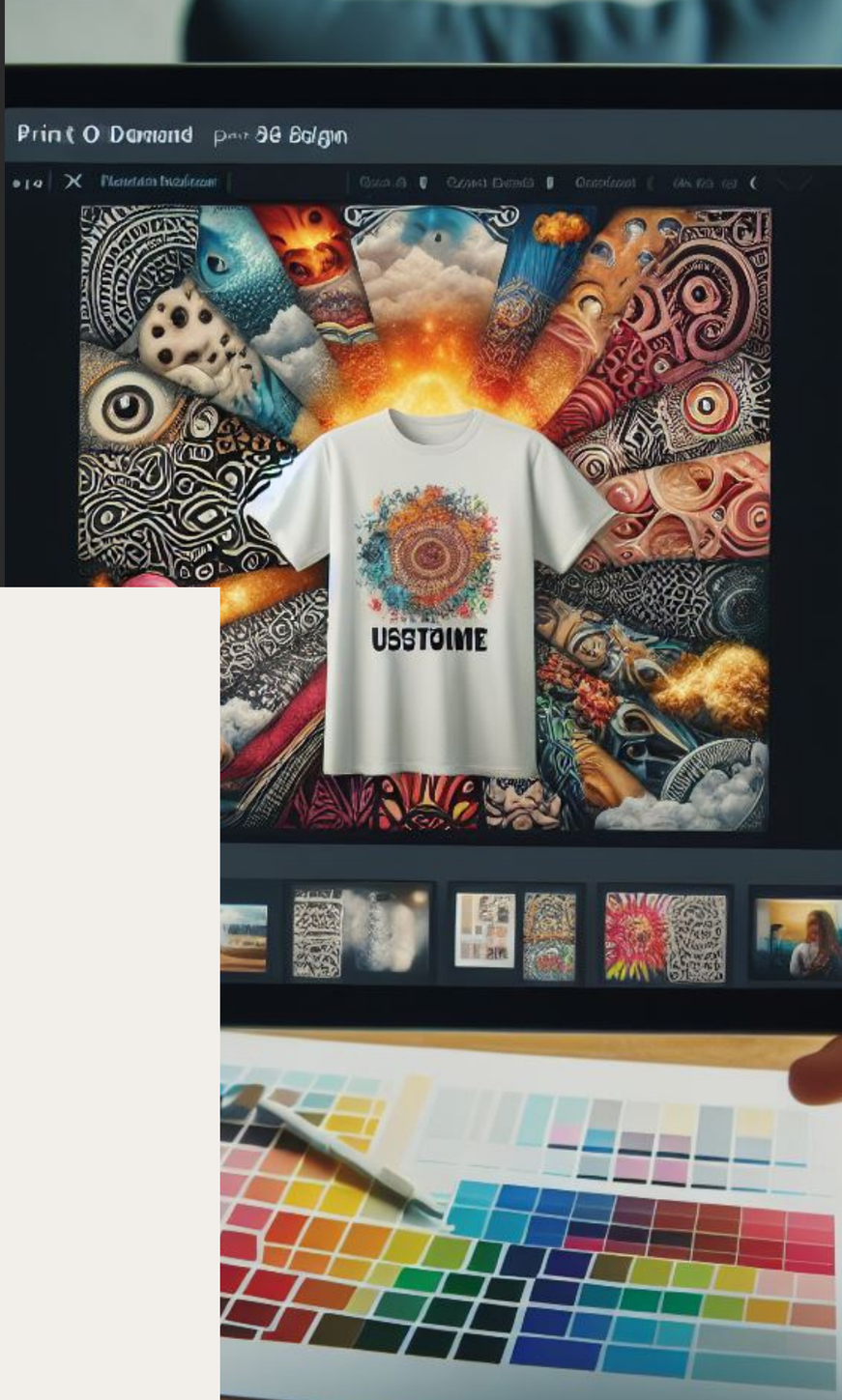
# 3. METHOD

| | | | MMLU | | BBH | | TyDiQA | MGSM |
|---|---|---|---|---|---|---|---|---|
| Params | Model | Norm. avg. | Direct | CoT | Direct | CoT | Direct | CoT |
| 80M | T5-Small | -9.2 | 26.7 | 5.6 | 27.0 | 7.2 | 0.0 | 0.4 |
| | Flan-T5-Small | -3.1 (+6.1) | 28.7 | 12.1 | 29.1 | 19.2 | 1.1 | 0.2 |
| 250M | T5-Base | -5.1 | 25.7 | 14.5 | 27.8 | 14.6 | 0.0 | 0.5 |
| | Flan-T5-Base | 6.5 (+11.6) | 35.9 | 33.7 | 31.3 | 27.9 | 4.1 | 0.4 |
| 780M | T5-Large | -5.0 | 25.1 | 15.0 | 27.7 | 16.1 | 0.0 | 0.3 |
| | Flan-T5-Large | 13.8 (+18.8) | 45.1 | 40.5 | 37.5 | 31.5 | 12.3 | 0.7 |
| 3B | T5-XL | -4.1 | 25.7 | 14.5 | 27.4 | 19.2 | 0.0 | 0.8 |
| | Flan-T5-XL | 19.1 (+23.2) | 52.4 | 45.5 | 41.0 | 35.2 | 16.6 | 1.9 |
| 11B | T5-XXL | -2.9 | 25.9 | 18.7 | 29.5 | 19.3 | 0.0 | 1.0 |
| | Flan-T5-XXL | 23.7 (+26.6) | 55.1 | 48.6 | 45.3 | 41.4 | 19.0 | 4.9 |

# 3. METHOD

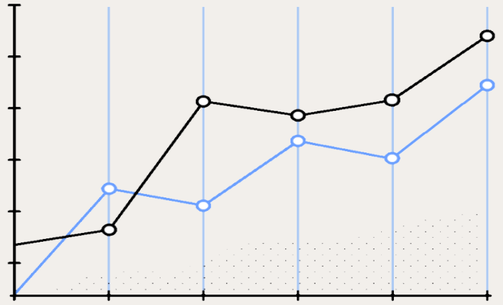

Create Text Embeddings
User Input
'a yellow and blue train riding a track by some trees'
CLIP tokenize
ViT-B/32
ViT-B/16
RN101
Encode
Encode
Encode
Concat
MultiCLIP text embedding
T5 tokenize
T5EncoderModel
T5 embedding
Prior Model
Diffusion process
Image embedding
Stable Diffusion 2 Decoder
Generate Image

# 4. RESULT AND DISCUSSION




**Results are evaluated using two methods:**

- FID score
- Visual evaluation



# 4. RESULT AND DISCUSSION

| Number of samples | FID Scores |
|---|---|
| 50 | 280.66 |
| 200 | 197.04 |
| 500 | 149.70 |
| 1000 | 107.87 |

240x1344

*Photos with dimensions of 240x1344, suitable for Shirt's logo and Long Sleeve printing, the images capture the specified details in the prompt and maintain good quality.*

384x288

*Photos with dimensions of 384x288, suitable for Adult Short Sleeve or Pocket, exhibit the highest quality.*

288x192

*Photos with dimensions of 288x192 intended for printing on Youth Short Sleeve or Pocket, there is a noticeable decline in detail and quality.*

# 4. RESULT AND DISCUSSION

*When generating images for Full Shirt Design, we have specified sizes for Adult Men's, Adult Women's, and Youth shirts as 1056x1200, 864x1056, and 864x1008, respectively. Nevertheless, with increasing resolution, the image begins to exhibit signs of deterioration. Objects start merging into one another, and the details gradually become less coherent. While the image maintains a certain level of quality, a significant portion of the details in the image loses logical consistency.*

1056x1200

864x1056

1056x1200

# 4. RESULT AND DISCUSSION

*To address this issue, our approach involves utilizing the ESRGAN model to enhance the resolution of the generated image. The optimal resolution for the ESRGAN model is determined to be 512x512 pixels and the height and width of the generated image are adjusted based on the user's original aspect ratio. If the user's resolution exceeds 512x512, the RRDB_PSNR_x4 model is used to upscale the image. By default, the model upscales the image four times compared to its original resolution. Therefore, we will first upscale the image and then resize it back to its original resolution.*

1056x1200

864x1056

1056x1200

# 4. RESULT AND DISCUSSION

*Image generation of sunset with the following parameters: prompt='romantic sunset on the beach', negative_prompt='low resolution', height=512, width=512, guidance_rescale=2.0, num_inference_steps=50.*

# 4. RESULT AND DISCUSSION

A cat on the sofa

A child playing on a sunny happy beach, her laughter as they build a simple sandcastle, emulate Nikon D6 high shutter speed action shot, soft yellow lighting

# 5. CONCLUSION

- The project showcases a pipeline for fine-tuning on an image-only dataset, reducing the need for extensive image labeling.
- The pipeline includes a user-friendly User Interface (UI) for deployment and application.
- Limitations exist in current models regarding text generation for POD images, as the models are trained to generate images based on text descriptions rather than generating the text itself.
- The Prior Model in this project was trained for approximately 10,000 steps, and further fine-tuning is expected to significantly enhance the quality of image generation.

# 5. CONCLUSION

- Future work aims to leverage Large Language Models (LLMs) for processing input prompts, including tasks like translation and pre-processing to improve image generation.
- The LLMs will be fine-tuned to predict detailed prompts and enhance image quality.
- A dataset of effective prompts will be prepared, and LLMs will be used to summarize and predict effective prompts from brief versions.

# 6. DEMO

# Thank's For Watching


Q&A